A thesis submitted in part requirement for the

degree of Doctor of Philosophy

Investigating the Role of Structural Variation in Male Infertility

Oguzhan Kalyon

Biosciences Institute

Faculty of Medical Sciences, Newcastle University

Supervisors: Prof Joris A Veltman, Dr Giles S Holt, Dr Aneta Mikulasova

March 2025

Abstract

Infertility affects approximately one in six couples globally, with male factors contributing to roughly half of all cases. Although chromosomal abnormalities such as Klinefelter's syndrome and Y chromosome microdeletions are well-established causes of male infertility (MI), nearly 40% of cases remain idiopathic. The role of structural variants (SVs) and dominant inheritance pattern has been understudied, primarily due to technical limitations in identifying SVs and the lack of patient-parent trio analyses required to investigate dominant *de novo* variants.

In this thesis, we performed whole-genome sequencing (WGS) on 216 patients with idiopathic azoospermia and their parents to investigate SVs across different inheritance models, with a particular focus on the dominant model. Additionally, whole-exome sequencing (WES) was conducted on 234 additional patients with azoospermia. In these cohorts, we identified several SVs that clearly explained the patients' phenotypes, as well as numerous potentially causative SVs that revealed novel candidate male infertility genes and loci.

This study demonstrates that WGS is an effective tool for studying SVs and significantly advances our understanding of the genetic basis of male infertility, particularly regarding the contribution of SVs.

Acknowledgements

First and foremost, I would like to offer my deepest gratitude to God for bestowing upon me all that I have. I also wish to express my heartfelt thanks to the late Ahmet Yagbasan, who sadly passed away and whom I miss so much. He is one of the main reasons I came to Newcastle, and there are no words to describe his greatness or how much he helped me. His memory continues to inspire me every day.

I would like to extend my sincere thanks to my supervisors, Prof Joris Veltman, Dr Aneta Mikulasova, and Dr Giles Holt, for their continuous support and guidance throughout my doctoral journey. In particular, I am immensely grateful to Giles for his daily support, to Aneta for making me an expert in CNV analysis, and to Joris, one of the finest geneticists of our time, for his invaluable insights and encouragement.

I am also grateful to all the members of the Genomics of Male Infertility Research Group, including research technicians Dr Bilal, Lois, Inas, our research nurse Joan, and postdoctoral researcher Dr Miguel. It was a pleasure to work with all of you, your camaraderie and expertise enriched my research experience.

I extend my thanks to all our collaborators, especially our colleagues at Radboud University, with special mention to Dr Godfried van der Heijden. My gratitude also goes to Frank Tüttelmann for his significant contribution to the replication studies, and to Dr Kevin Mceleny from Newcastle Fertility Centre for his kind support in patient recruitment.

I would also like to thank the members of my PhD progression panel, Dr Mauro Santibanez-Koref and Prof Gavin Hudson, for their constructive feedback and valuable suggestions. Furthermore, I am appreciative of all the staff at Newcastle University, particularly those at NUBI and the Centre for Life, who made administrative and logistical tasks seamless and well-organised.

I gratefully acknowledge my country, Türkiye, and the Republic of Türkiye Ministry of National Education for awarding me the scholarship that enabled me to study here. This opportunity has been instrumental in my academic progress and development.

Finally, I want to acknowledge everyone who made my time in this city memorable and enjoyable. In particular, my dear friends Hüseyin Ayan, Cafer Deniz, Fatih Cure, Yusuf Ugurluoglu, Muhammed Karayagli, Usame Altuntas, Usame Gurbuz, Ismail Kan, Muhammed

Cavus, Furkan Celiktas, Emre Ozturk, Nurettin Ayvali, and Umit Demirbaga, thank you for making life here lively and fun. I also owe much to all my former teachers, especially my former supervisors Assoc Prof Alper Yilmaz and Prof Hamza Muslumanoglu, who shaped my academic journey.

I am forever indebted to my family for their unconditional support, especially my mother Zeynep, my father Mehmet, and my sister Arzu. Their endless love and belief in me has been a pillar of strength throughout this process. Last but not least, I offer my heartfelt gratitude to my beloved wife, Feyza Gul, who has been by my side through every step of this journey, offering unwavering support, patience, and love. This thesis would not have been possible without her.

Declaration

I hereby declare that 90% of the work presented in this thesis is my independent contribution. Collaborative works have been appropriately acknowledged.

Contents

List of Figures

List of Tables

List of Abbreviations

aCGH: Array Comparative Genomic Hybridisation

ACMG: American College of Medical Genetics and Genomics

AD: Autosomal Dominant

AIS: Androgen Insensitivity Syndrome

AMP: Association for Molecular Pathology

AR: Androgen Receptor

AR: Autosomal Recessive

ART: Assisted Reproductive Technologies

BMI: Body Mass Index

BOM: Bionano Optical Mapping

BSU: Bioinformatics Support Unit

BTB: Blood-Testis Barrier

CBAVD: Congenital Bilateral Absence of the Vas Deferens

ClinGen: The Clinical Genome Resource

CN: Copy Number

CNV: Copy Number Variation

cnnLOH: Copy-Neutral Loss of Heterozygosity

CS: Complex SVs

DEL: Deletion

DUP: Duplication

FDR: False Discovery Rate

FISH: Fluorescence In Situ Hybridisation

FoSTeS: Fork-Stalling and Template Switching

FSH: Follicle-Stimulating Hormone

gnomAD: Genome Aggregation Database

GO: Gene Ontology

GWAS: Genome-Wide Association Studies

HMW: High Molecular Weight

ICSI: Intracytoplasmic Sperm Injection

IGV: Integrative Genomics Viewer

IMIGC: International Male Infertility Genomics Consortium

INS: Insertion

INV: Inversion

IVF: In Vitro Fertilization

KCMIG: Known and Candidate Male Infertility Genes

LH: Luteinizing Hormone

LMM: Linear Mixed-effects Model

LoF: Loss-of-Function

MAF: Minor Allele Frequency

MAR: Medically Assisted Reproduction

MERGE: Male Reproductive Genomics, Münster, Germany

MMAF: Multiple Morphological Abnormalities of the Sperm Flagella

MI: Maternally Inherited

NAHR: Non-Allelic Homologous Recombination

NHEJ: Non-Homologous End-Joining

NGS: Next-Generation Sequencing

NOA: Non-Obstructive Azoospermia

OAT: Oligoasthenoteratozoospermia

OGM: Optical Genome Mapping

ONT: Oxford Nanopore Technologies

PARs: Pseudoautosomal regions

PCOS: Polycystic Ovary Syndrome

PESA: Percutaneous Epididymal Sperm Aspiration

PGCs: Primordial Germ Cells

PI: Paternally Inherited

PON: Panel of Normals

RD: Read Depth

REs: Repeat Expansions

RP: Read Pair

SR: Split Read

SMRT: Single-Molecule Real-Time

SNVs: Single Nucleotide Variations

SPGF: Primary Spermatogenic Failure

SV: Structural Variation

TESE: Testicular Sperm Extraction

TRA: Translocation

TRs: Tandem Repeats

TS: Targeted Sequencing

UPD: Uniparental Disomy

VEP: Variant Effect Predictor

WES: Whole-Exome Sequencing

WGS: Whole Genome Sequencing

WHO: World Health Organization

srWGS: Short Read Whole Genome Sequencing

Chapter 1. Introduction

1.1 Overview of Male Infertility

1.1.1 Definition, Prevalence and Classification

Infertility affects approximately 10-15% of all couples globally, with approximately half of the cases arising due to male factors (Neto et al., 2016). Defined as the inability of a couple to conceive after one year of unprotected intercourse (Zegers-Hochschild et al., 2009) infertility affects approximately 7% of the male population worldwide (Krausz & Riera-Escamilla, 2018).

Male fertility is commonly assessed through semen analysis, evaluating parameters such as ejaculate volume, sperm concentration, motility, and morphology (World Health Organization, 2021). According to the latest World Health Organisation (WHO) reference values, normozoospermic men should present an ejaculate volume of at least 1.4 mL, a total sperm number of 33 million or more per ejaculate, a concentration of 12 million sperm per mL or higher, at least 42% motile sperm (of which 30% should show progressive motility), and at least 4% morphologically normal forms (World Health Organization, 2021). Deviations from these thresholds may indicate underlying male factors contributing to infertility. Additionally, the 6th edition of WHO manual (2021) emphasises the importance of consistent laboratory techniques, standardization of assessments, and precise calculations to minimize variability and ensure reliable results.

From an aetiological standpoint, male infertility is often categorised into four broad groups. The largest category comprises quantitative defects in spermatogenesis (Krausz & Riera-Escamilla, 2018). Other categories include ductal obstruction or dysfunction (obstructive pathologies), disorders of the hypothalamic-pituitary axis (secondary testicular failure or secondary hypogonadism), and qualitative defects in sperm structure or function (Tournaye et al., 2017).

Azoospermia is defined as the complete absence of sperm in the ejaculate (Figure 1.1a). It represents one of the most severe quantitative abnormalities and affects approximately 10-15% of infertile men. (Tüttelmann, Werny, et al., 2011). This phenotype is highly impactful since, without sperm in the ejaculate or testicular tissue, assisted reproductive techniques become challenging and in many cases unfeasible. Azoospermia can arise from a physical

blockage in the excurrent ductal system (obstructive azoospermia, accounts for roughly 40% of cases) or from intrinsic spermatogenic failure (non-obstructive azoospermia, accounts for roughly 60% of cases) (Hubbard et al., 2025). The latter is more severe, as it often results in very limited or even absent sperm retrieval during testicular sperm extraction (TESE). Other well-characterised male infertility phenotypes include oligozoospermia (reduced sperm count) (Figure 1.1b), asthenozoospermia (reduced motility) (Figure 1.1c), teratozoospermia (abnormal morphology) (Figure 1.1d), and combined defects (oligoasthenoteratozoospermia (OAT) syndrome). Less common forms include globozoospermia (<0.1 % in infertile males), characterised by round-headed sperm lacking an acrosome (Figure 1.1e), and flagellar anomalies classified under conditions such as Multiple Morphological Abnormalities of the Sperm Flagella (MMAF) (<0.1 % in infertile males) (Esteves et al., 2018; Krausz et al., 2015).

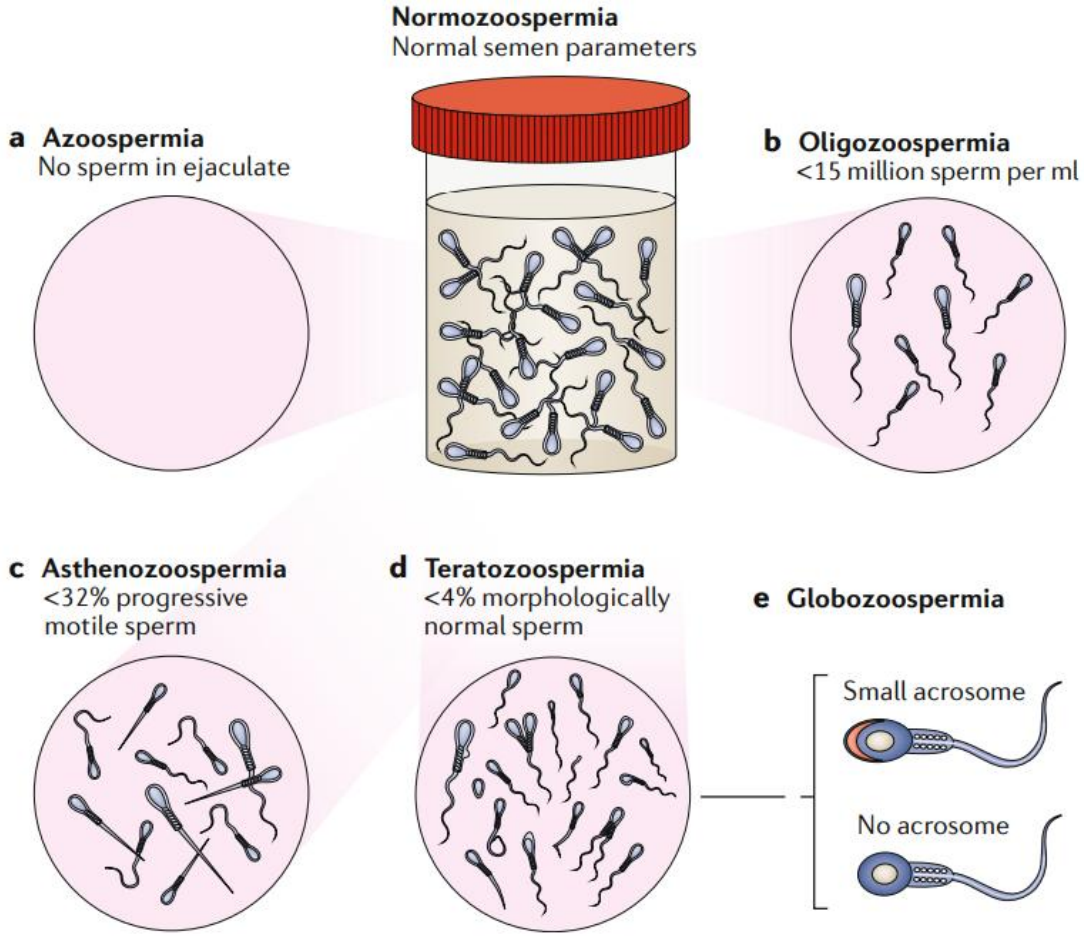


Figure 1.1 Seminal alterations associated with male infertility. **a**) Azoospermia. **b**) Oligozoospermia. **c**) Asthenozoospermia. **d**) Teratozoospermia. **e**) Globozoospermia. Figure courtesy of Esteves et al., (2018)

1.1.2 Understanding Spermatogenesis

Spermatogenesis is a highly intricate and tightly regulated process that is essential for male fertility (Njogu et al., 2010; Zheng et al., 2013), accurate interpretation of male infertility requires an understanding of this complex process. The hormonal control of spermatogenesis is governed by the hypothalamic-pituitary-testicular axis, which coordinates the release of key reproductive hormones (Sharma Rakeshand Agarwal, 2011; Wang et al., 2017; Ding et al., 2021). Gonadotropin-releasing hormone (GnRH) from the hypothalamus stimulates the anterior pituitary to secrete follicle-stimulating hormone (FSH) and luteinizing hormone (LH) (Sharma Rakeshand Agarwal, 2011). FSH binds to receptors on Sertoli cells, which provide critical support and nurturing for the developing germ cells. LH acts on Leydig cells to stimulate the production of testosterone, a crucial hormone for the maintenance and progression of spermatogenesis (Figure 1.2) (Sharma Rakeshand Agarwal, 2011). In addition to this hormonal regulation, spermatogenesis is also modulated by various other factors, including growth factors, cytokines and epigenetic modifications such as DNA methylation and N6-methyladenosine (m6A) RNA modification (Toragall et al., 2018; Guo, 2023). These epigenetic mechanisms play important roles in the coordinated expression of genes involved in germ cell development and differentiation (Guo, 2023; Toragall et al., 2018).

Primordial germ cells (PGCs) are the precursors of spermatogonial stem cells, which give rise to the entire population of germ cells during spermatogenesis (G. Wang et al., 2019). PGCs originate from the epiblast during embryonic development and migrate to the genital ridge, where they undergo a series of developmental changes to become spermatogonial stem cells (Sharma Rakeshand Agarwal, 2011; Wang et al., 2019). This process of PGC specification and migration takes place during the early stages of embryonic development.

Spermatogonia are the stem cells located along the basement membrane of the seminiferous tubules. They serve as the foundation for continuous sperm production by either self-renewing or differentiating into primary spermatocytes (Sharma Rakeshand Agarwal, 2011). There are two main types: Type A spermatogonia, which can self-renew or begin differentiation, and Type B, which are committed to differentiation (Figure 1.2) (Rato et al., 2012). The transition from spermatogonia to primary spermatocytes takes about 16 days in humans.

During meiosis, primary spermatocytes undergo two cell divisions to produce haploid spermatids (Figure 1.2). This process involves the reduction of the diploid chromosome number to the haploid state, ensuring that each mature spermatozoon contains a single set of chromosomes (Sharma Rakeshand Agarwal, 2011). The meiotic phase of spermatogenesis takes approximately 24 days in humans, during which the genetic material is recombined and segregated to produce the haploid germ cells.

Following meiosis, the haploid cells produced are known as spermatids, which must undergo spermiogenesis to attain a mature spermatozoon form. During this process, the spermatids condense their chromatin, form an acrosome, and develop a functional flagellum, ultimately resulting in highly specialised, elongated spermatozoa (Sharma Rakeshand Agarwal, 2011). Spermatids originating from a single spermatogonium remain interconnected through intercellular bridges, enabling coordinated development and the transfer of essential biochemical signals necessary for their maturation (Dym & Fawcett, 1971; Sharma Rakesh and Agarwal, 2011). After approximately 21 days of spermiogenesis, the mature spermatids disengage from the Sertoli cells and are released into the lumen of the seminiferous tubules, marking the final step of sperm cell formation known as spermiation. (Sharma Rakeshand Agarwal, 2011). Although these spermatozoa are morphologically complete, their fertilizing capacity is only fully realised after “capacitation,” a series of biochemical and physiological changes occurring in the epididymis and female reproductive tract Oud et al., 2017).

Spermatogenesis is a complex, gene-intensive process, with at least 2,000 genes contributing to its regulation, progression, and successful completion (Krausz & Riera-Escamilla, 2018). The precise regulation of these genes is essential for the successful completion of spermatogenesis and the production of functional spermatozoa. Disruptions in the expression or regulation of these genes can lead to various forms of male infertility, highlighting the importance of this intricate biological process.

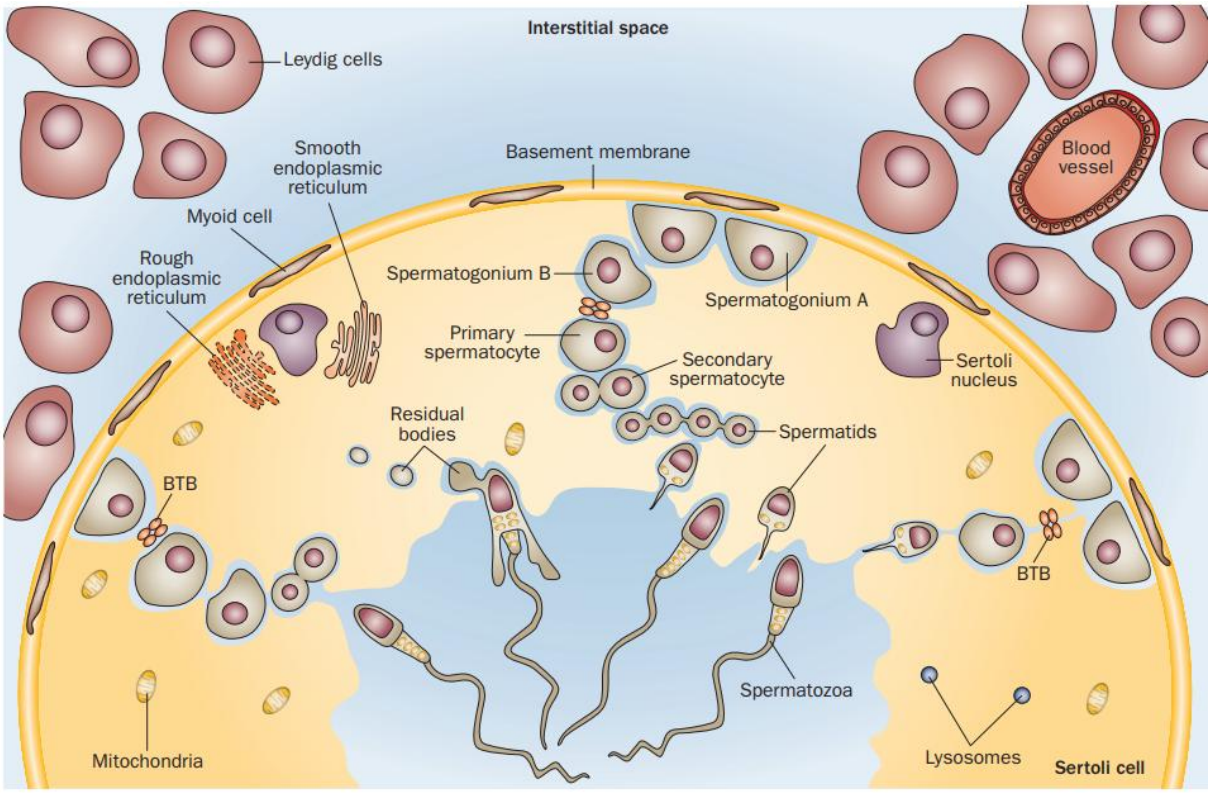


Figure 1.2. Overview of spermatogenesis. Spermatogenesis occurs in the seminiferous epithelium, which consists of Sertoli cells and germ cells at various stages of development. Surrounding the epithelium are Leydig cells and blood vessels in the interstitium. This process transforms diploid spermatogonial stem cells into haploid sperm cells through cellular division. Continuous sperm production relies on both intrinsic factors (Sertoli and germ cells) and extrinsic factors (hormonal regulation). Spermatogonia type A develop into type B, which then enter meiosis. Meiosis I produces haploid secondary spermatocytes, while Meiosis II results in spermatids. These spermatids migrate to the lumen, where they mature into spermatozoa. (BTB= blood-testis barrier) Figure courtesy of Rato et al., (2012)

1.1.3 Risk Factors and Non-genetic Causes of Male Infertility

Male infertility arises from a complex interplay of genetic, biological, environmental, and lifestyle factors (Okonofua et al., 2022). Several conditions directly impair male fertility without a genetic basis. Sexually and non-sexually transmitted diseases are common contributors, with uro-genital infections, such as those caused by Chlamydia trachomatis or other sexually transmitted pathogens, frequently implicated (Schuppe et al., 2017; Okonofua et al., 2022). These infections can lead to conditions like epididymo-orchitis or obstructive azoospermia, which severely impact male fertility (Masarani et al., 2006). Varicocele, a condition characterised by the enlargement of scrotal veins, is also prevalent among infertile men (approximately 40% in infertile males whereas around 15–20% in male population) and is often considered a non-genetic cause of infertility. (Okonofua et al., 2022). Studies have

highlighted its association with reduced sperm quality and abnormal parameters, including DNA fragmentation (Cozzolino, 2001; Jarow, 2001). Furthermore, testicular trauma and surgical procedures, such as hernia repairs, can adversely affect testicular function (Okonofua et al., 2022). Hormonal imbalances, including low testosterone levels or hyperprolactinemia, also play a critical role in disrupting spermatogenesis and libido (Jarow, 2003). Additionally, research has shown that malignancies (Nagirnaja et al., 2018) and immunological factors (Brugh & Lipshultz, 2004) affect male fertility.

Environmental factors pose significant risks to male fertility (Okonofua et al., 2022). Exposure to industrial chemicals, pesticides such as DDT, and heavy metals like lead has been strongly linked to testicular dysfunction and low sperm counts (Cherry, 2001). Elevated scrotal temperatures caused by prolonged sitting, tight clothing, or occupational hazards, such as working in high-temperature environments, can impair spermatogenesis (Okonofua et al., 2022). Similarly, radiation and chemotherapy, while medically necessary, can irreversibly damage testicular tissue and spermatogenic cells, further contributing to infertility. Given the significant risk of fertility loss in prepubertal and adolescent males undergoing such treatments, fertility preservation strategies have become increasingly important. Sperm cryopreservation is the standard approach for adolescent males capable of producing a sample, while testicular tissue cryopreservation offers a promising alternative for prepubertal boys who have not yet initiated spermatogenesis (Picton et al., 2015).

Advanced age (Brugh & Lipshultz, 2004) and poor lifestyle choices are also considered as risk factors of male infertility. Also, body mass index (BMI) is closely associated with sperm quality, as both underweight and obesity can negatively affect reproductive outcomes. Obesity, in particular, is associated with hormonal imbalances such as hypotestosteronaemia, which can contribute to fertility problems, while lifestyle modifications may reverse these (Kort, 2006). Smoking and excessive alcohol consumption are detrimental as well, increasing oxidative stress and causing DNA damage to sperm cells (Emanuele MA & Emanuele NV, 1998). Chronic use of anabolic steroids is known to disrupt the hypothalamic-pituitary-gonadal axis, resulting in a significant reduction in sperm production. Additionally, both physical and psychological stress, though still debated as contributors to infertility, have been shown to suppress testosterone production through elevated cortisol levels (Okonofua et al., 2022).

1.1.4 Clinical Management of Male Infertility

Clinical management of male infertility have advanced considerably since the introduction of assisted reproductive technologies (ART), notably in vitro fertilization (IVF) and intracytoplasmic sperm injection (ICSI). IVF, first performed in 1978 (Steptoe & Edwards, 1978), involves fertilising oocytes by incubating them with sperm in a Petri dish. Building upon the IVF technique, ICSI, introduced in 1992 (Palermo, 1992), refines the process by injecting a single sperm directly into the oocyte cytoplasm using a glass micropipette (Figure 1.3). For both IVF and ICSI, sperm is ideally obtained from the ejaculate, but in cases of severe male infertility such as azoospermia where no sperm is identified in the ejaculate, surgical retrieval of sperm by approaches such as TESE or percutaneous epididymal sperm aspiration (PESA) becomes necessary (Esteves et al., 2018). Globally, since 1978, an estimated 60 million ART cycles have been carried out, resulting in the birth of around 10 million infants according to ICMART Preliminary World Report (Adamson et al., 2024). The success of ART varies across different forms of infertility, with higher live birth rates for unexplained (35-40%) and male factor infertility (30-35%), but lower rates for conditions like endometriosis and polycystic ovary syndrome (PCOS) at 25-30%, often necessitating multiple cycles, typically 2-3, to achieve a successful pregnancy (Centers for Disease Control and Prevention, 2022). This process can be emotionally and physically demanding, particularly for the female partner. Genetic studies help identify couples for whom ART may be ineffective, emphasising the need for personalised approaches to improve success rates.

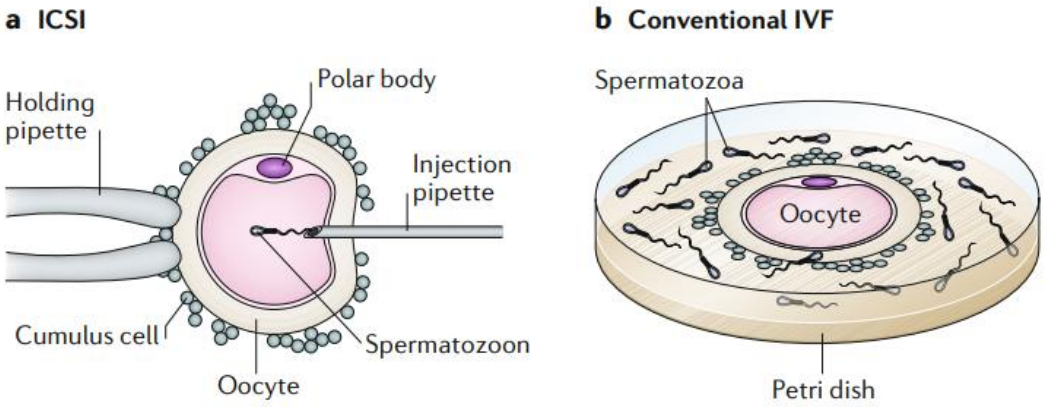


Figure 1.3 Assisted fertilization methods. a) Intracytoplasmic sperm injection (ICSI) is a procedure in which a single sperm cell is directly introduced into the cytoplasm of an oocyte using a fine glass micropipette. **b**) Conventional in vitro fertilization (IVF) involves placing an oocyte in a Petri dish alongside sperm, enabling the male gametes to naturally fertilize the oocyte.(Esteves et al., 2018)

Despite the widespread use and success of ART, there have been concerns about the health of ART-conceived offspring. Studies suggest that children born through IVF or ICSI may have a slightly increased risk of congenital malformations (Relative Risk ~1.33) (Qin et al., 2015).. When it comes to the effect of ART on the reproductive health of offspring, it has been understudied. After decades of use, the first ICSI-conceived children have reached adulthood, allowing researchers to begin investigating their reproductive fitness. An initial study on 54 ICSI-conceived young adult males reported more pronounced reductions in semen quality, such as lower sperm concentration and total sperm count, when compared with spontaneously conceived peers (Belva et al., 2016). In contrast, more recent and relatively larger-scale research that included men conceived with IVF/ICSI found no significant differences in sperm output or concentration compared to men conceived without ART, although subtle variations in sperm motility and morphology were observed (Catford et al., 2022). These studies underscore the importance of long-term monitoring and large-scale studies to fully elucidate the genetic and reproductive health implications of ART. Additionally, since IVF, particularly ICSI, may result in the transmission of pathogenic genetic variants from infertile fathers to their offspring, understanding the genetics of male infertility has become increasingly important. For instance, Patsalis et al., 2002, showed that the inheritance of Y chromosome deletions through ART can lead to infertile male progeny with azoospermia or severe oligozoospermia and consequent severe clinical implications. Thus, a comprehensive understanding of the genetic factors underlying male infertility, as well as the effects of IVF and ICSI on the reproductive health of offspring, is crucial for providing appropriate counselling and treatment.

1.2 Genetics of Male Infertility

Known genetic abnormalities account for the genetic aetiology of male infertility in approximately 15-20% of infertile men with still 40% of the total classified as idiopathic (Krausz & Riera-Escamilla, 2018). Genetic factors are involved in four of the major aetiological categories of male infertility: spermatogenic quantitative defects; ductal obstruction or dysfunction; hypothalamic-pituitary axis disturbances; and spermatogenic qualitative defects (Tournaye et al., 2017). Men with azoospermia face the highest likelihood of carrying genetic abnormalities (25%), with this risk gradually declining as sperm production increases (Figure 1.4) (Krausz & Riera-Escamilla, 2018).

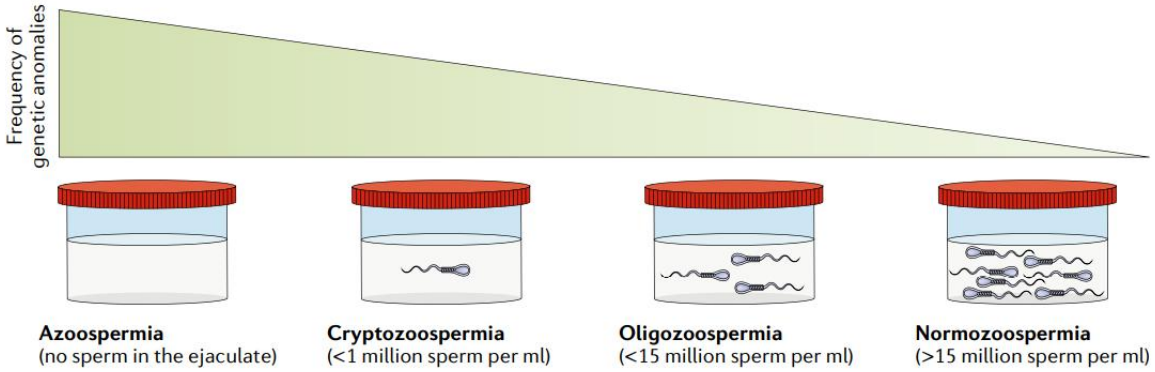


Figure 1.4 The different types of quantitative disturbances of spermatogenesis and the frequency of genetic factors in each category (Krausz & Riera-Escamilla, 2018).

1.2.1 Chromosome Anomalies

Standard karyotyping is the first-line genetic test for men with nonobstructive azoospermia and severe oligozoospermia, as around 15% of men with nonobstructive azoospermia and up to 4% of those with moderate oligozoospermia have detectable chromosomal anomalies (Minhas et al., 2021).

The most common chromosomal abnormality resulting in NOA is Klinefelter syndrome (47, XXY and mosaic variants), with an estimated prevalence of 1 in 600 in the general population and approximately 1 in 7 among men with NOA (Punab et al., 2016). Klinefelter syndrome is characterised by testicular failure, androgen deficiency, and an increased risk of various co-morbidities, including metabolic, autoimmune, and neuropsychiatric disorders (Belling et al., 2017). Although azoospermia is typical, some men (especially younger or mosaic cases) may have rare spermatozoa in the ejaculate, enabling possible fertility treatment through TESE and ICSI. Early diagnosis and fertility preservation, along with lifelong health monitoring, are

advised for patients with Klinefelter syndrome as spermatogenesis progressively declines with age due to germ cell loss and testicular fibrosis, reducing the likelihood of successful sperm retrieval later in life. (Krausz & Riera-Escamilla, 2018). Another related condition is the 46,XX male (de la Chapelle) syndrome, which results from translocation of the SRY gene onto an X chromosome in most cases (Zenteno-Ruiz et al., 2001). Men with this syndrome are invariably azoospermic, often have smaller stature, and can present with gynaecomastia or ambiguous genitalia. TESE is not recommended in these patients due to the absence of Y-linked AZF regions (Skaletsky et al., 2003). Other structural chromosomal anomalies, such as Robertsonian translocations, inversions, and reciprocal translocations, are more common in men with oligozoospermia than in those with normal sperm counts (Vincent et al., 2002). These rearrangements can disrupt meiosis or gene expression, impairing spermatogenesis. Preimplantation genetic diagnosis (PGD) is advised when using assisted reproductive techniques in these cases to reduce the risk of transmitting chromosomal imbalances to offspring (Krausz & Riera-Escamilla, 2018).

Sub-microscopic deletions of the Y chromosome, including those affecting AZF regions (AZFa, AZFb, AZFc), are another important genetic cause of severe male infertility (Vogt, 1996). The proximal region of AZF harbours repetitive homologous sequences which elevate the likelihood of non-allelic homologous recombination (NAHR) events and, as a result, increase the overall risk of SVs occurring in these regions. AZF deletions occur in about 10% of men with NOA and 5% with severe oligozoospermia (Lo Giacco et al., 2014). They remove essential spermatogenesis genes, leading to varying degrees of sperm production impairment (Krausz et al., 2024). The phenotypic outcomes associated with deletions in these regions differ depending on the specific area affected, ranging from azoospermia to oligozoospermia. Typically, azoospermia is seen in individuals with extensive AZFa deletions, whlist those with AZFc deletions generally exhibit only a reduced sperm count (Krausz & Riera-Escamilla, 2018). Partial deletions like the gr/gr deletion in the AZFc region increase the risk of oligozoospermia (Rozen et al., 2012) and may also be transmitted to offspring, potentially escalating to a complete AZFc deletion in future generations (Krausz & Riera-Escamilla, 2018).

1.2.2 Monogenic Causes

Monogenic factors are associated with all three aetiologic categories of decreased sperm count, motility, and morphology. Currently, only a small number of genes have been clearly linked to male infertility (IMIGC has classified 48 genes as definitively linked to male infertility),

and even fewer are regularly assessed in the clinical evaluation of male infertility (e.g. *CFTR* and *AR* genes) (Wyrwoll et al., 2024). One example is the cystic fibrosis transmembrane regulator (*CFTR*) gene, which, when both alleles are compromised, leads to cystic fibrosis. More than 95% of men with cystic fibrosis present infertility due to congenital bilateral absence of the vas deferens (CBAVD). The vas deferens are ducts that carry sperm from the testes to the urethra. Blockage of these ducts results in obstructive azoospermia. (Bieth et al., 2021). In 75-80% of cases, men with bi-allelic *CFTR* mutations are infertile due to isolated CBAVD but do not exhibit other cystic fibrosis symptoms. In these cases, milder *CFTR* mutations likely cause a variant of cystic fibrosis limited to the reproductive tract, with isolated CBAVD considered a mild form of the disease (Bieth et al., 2021). The other well-known disease gene is the Androgen Receptor (*AR*) gene. Pathogenic variants in both copies of the *AR* gene cause Androgen Insensitivity Syndrome (AIS), which ranges from complete AIS (CAIS) to mild AIS (MAIS) (Ferlin et al., 2006). While CAIS patients present with a female external phenotype and lack typical masculinization, individuals with milder forms, MAIS and partial AIS (PAIS), usually have a male external phenotype, more typical masculinisation, and testicular development, but still experience infertility (O'Hara & Smith, 2015). Interestingly, an analysis of andrological parameters in infertile versus control men found that a longer *AR* CAG repeat length was significantly associated with lower sperm count, reduced motility, and higher serum FSH levels ($P<0.05$) (Mosaad et al., 2012). However, a later meta-analysis conducted by Xiao et al., 2016, concluded that there were contradictory results among different studies regarding CAG repeats between azoospermic cases and controls. As WGS allows for the analysis of repeat expansions throughout the genome (see Section 1.4.2 on SV Detection and Next-Generation Sequencing), these expansions can be analysed in a targeted or a whole-genome manner. In our cohort, known regions linked to the male infertility, including the *AR* gene, are currently undergoing targeted analysis by colleagues within our research group. Both *AR* and *CFTR* genes are considered known male infertility genes, and testing for these genes is often incorporated into routine diagnostics (Houston et al., 2021). Nevertheless, the mainstay of genetic diagnostics in male infertility has not evolved much since the discovery of the AZF regions two decades ago, primarily focusing on karyotype analysis, Y chromosome deletion testing, and *CFTR* and *AR* mutation screening (Houston et al., 2021; Xavier et al., 2021).

From 2008 onward, next-generation sequencing (NGS) technologies have been widely adopted by labs worldwide, including those studying male infertility. However, despite the surge in research resulting in many candidate gene discoveries, practical improvements, reflected in standard diagnostic procedures have been limited. A systematic review published in 2021, which examined studies on potential male infertility genes, showed a steady increase in the amount of genetic research since the 1990s (Oud et al., 2019).

Houston et al., 2021, as of 2020 July, identified a total of 120 genes that were moderately, strongly, or definitively linked to 104 infertility phenotypes. For non-syndromic forms of isolated infertility, only 28 autosomal recessive genes, 11 autosomal dominant genes, and 7 X-linked genes were definitively implicated in infertility phenotypes (Kasak & Laan, 2021) . Many of the recessive genes were identified in patients from consanguineous backgrounds and were associated with specific qualitative sperm defects, suggesting these variants likely do not account for a significant proportion of the more common quantitative sperm impairments in the general population. It still remains unclear for the limited autosomal and X-linked genes whether the harmful variants were inherited or originated as *de novo* (Houston et al., 2021). Even though genetic factors are believed to play a significant role in a substantial proportion of male infertility cases (Krausz et al., 2015), they have not significantly contributed to diagnosis, with at least 40% of all cases still classified as idiopathic (Kasak & Laan, 2021).

1.2.3. Genomics Studies in Idiopathic Male Infertility

The development of high-throughput genomic technologies and the completion of major international genomic initiatives have driven large-scale genomic research across various medical fields, significantly advancing the understanding of complex diseases (Auton et al., 2015). Over the past 25 years, genomic tools have evolved and can be broadly categorised into microarray-based methods (such as SNP arrays, and comparative genomic hybridisation (CGH) arrays) and NGS approaches. For instance, SNP arrays have been instrumental in identifying candidate genes such as *SPATA16* and *DPY19L2*, which are associated with globozoospermia (Dam et al., 2007; Harbuz et al., 2011). While genome-wide association studies (GWAS) using SNP arrays have highlighted a few common SNPs linked to male infertility, their overall contribution appears minimal, with validated SNPs showing small effects when co-occurring in individuals (Krausz & Riera-Escamilla, 2018) This limited contribution is expected, as variants with strong effects on male fertility are subject to negative selection and are gradually eliminated from the population through evolutionary

processes. Similarly, aCGH arrays have proven valuable in detecting copy number variations (CNVs), such as Y chromosome-linked deletions (e.g., AZF and gr/gr deletions), which directly affect spermatogenesis (Krausz & Riera-Escamilla, 2018). This technology also enabled the discovery of a *TEX11* intragenic deletion linked to azoospermia, underscoring its utility in identifying specific genetic alterations with clinical significance (Yatsenko et al., 2015).

With the advancements of NGS, the list of genes implicated in male infertility has grown steadily, revealing a marked genetic heterogeneity behind male infertility (Houston et al., 2021). Initially, the use of NGS in infertility assessments focused on targeted gene panels that allowed parallel sequencing of up to hundreds of genes suspected or known to be disease-related (Neto et al., 2016). As NGS costs declined, these targeted approaches gave way to exome sequencing, encompassing nearly all protein-coding regions and their exon-intron boundaries. This transition made it possible to identify the underlying homozygous variants in consanguineous families and uncover potentially significant *de novo* mutations through trio-based exome sequencing of an infertile man and his parents (Gershoni et al., 2017; Oud et al., 2021). Moreover, exome sequencing supports the analysis of large patient cohorts simultaneously, offering a powerful strategy for pinpointing extremely rare genetic defects responsible for the condition (Stallmeyer et al., 2024).

Stallmeyer et al., 2024, conducted a comprehensive review of exome sequencing studies on genetic male infertility and identified 70 genes with at least moderate evidence of contributing to the condition. Their analysis highlighted that different phenotypic subgroups are associated with distinct sets of disease genes. For instance, several genes implicated in NOA, such as *C14ORF39, HFM1, KASH5, M1AP, MEI1, MSH4, MSH5, MEIOB, SHOC1, STAG3,* and *TEX11*, encode proteins essential for meiosis, a critical process in spermatogenesis. Notably, genes involved in the piRNA pathway, including *PNLDC1, PIWIL2, FKBP6, TEX15*, and *ADAD2*, have also been recognised as key contributors to NOA and impaired spermatogenesis (see Table 1.1 for examples with more details).

Building upon this work, a recent study, published after aforementioned review, screened likely pathogenic and pathogenic variants across 638 candidate genes for male infertility in the ESTAND cohort (Lillepea et al., 2024). This analysis included 521 individuals with idiopathic primary spermatogenic failure (SPGF) and 323 normozoospermic controls. The study validated the association of SPGF with several newly proposed candidate genes, including *ACTRT1, ASZ1,*

GLUD2, GREB1L, LEO1, RBM5, ROS1, and *TGIF2LY*, further expanding the genetic landscape of male infertility (see Table 1.1 for examples with more details).

Table 1.1 Example of genes implicated in male infertility.

Gene Symbol	Function / Pathway	Associated Phenotype(s)
TEX11	Encodes a protein critical for homologous chromosome synapsis and crossover during meiosis; its disruption causes a failure of this process, leading to meiotic arrest.	NOA
M1AP	Encodes "Meiosis 1 Associated Protein," which is primarily expressed in male germ cells and is crucial for the progression of meiosis.	NOA, severe oligozoospermia
STAG3	Encodes a meiosis-specific component of the cohesin complex, which is necessary for chromosome pairing and sister chromatid cohesion. Its disruption causes meiosis to fail, leading to meiotic arrest.	NOA
FKBP6	Acts as a piRNA-pathway factor essential for pachytene piRNA biogenesis. Its loss disrupts translational regulation during spermiogenesis, causing an arrest at the round spermatid stage.	NOA, severe oligozoospermia
PNLDC1	Acts as an exonuclease in the piRNA pathway that trims the 3’ ends of pre-piRNAs to create mature piRNAs. Its disruption leads to faulty piRNA processing and a collapse of the piRNA machinery.	NOA
FANCM	Acts as a DNA translocase that maintains genomic stability by assisting in DNA repair and replication. Disruption of this function during spermatogenesis compromises the ability to repair DNA damage, leading to germ cell apoptosis and subsequent germ cell loss.	NOA, Oligoasthenozoospermia
RBM5	Acts as an essential regulator of pre-mRNA splicing in haploid male germ cells. Its disruption leads to incorrect splicing of transcripts required for spermatid differentiation, causing developmental arrest.	NOA
ASZ1	Encodes a protein that co-localizes with the piRNA-pathway protein *PIWIL2*, suggesting it is involved in germline cell development and maintenance. Its disruption leads to a failure in spermatogenesis.	NOA, severe oligozoospermia
ROS1	Encodes a signalling protein essential for the epithelial differentiation of the epididymis. Its disruption leads to arrested sperm maturation and sterility in the mouse model.	NOA, severe oligozoospermia

1.2.3.1 De novo Paradigm

Due to the involvement of roughly 2300 genes in testis function and spermatogenesis (Skaletsky et al., 2003), identifying all genetic factors in male infertility requires large-scale unbiased studies; yet, most research has focused on recessive or X-linked variants, leaving the role of dominant genes largely unexplored and poorly understood (Houston et al., 2021; Stallmeyer et al., 2024). Dominant mutations typically fail to transmit through the paternal line in male infertility, but insights from other conditions affecting reproductive fitness suggest that *de novo* germline mutations and SVs can still lead to dominant forms of the disease. For example, WES and WGS studies have shown that damaging *de novo* germline mutations explain the majority of severe intellectual disability cases (Gilissen et al., 2014). The frequency of a genetic disorder caused by *de novo* mutations correlate with the number of genes (mutational target) that can produce the condition when mutated, meaning more genetically heterogeneous disorders are relatively more common (Veltman & Brunner, 2012). This principle is illustrated by the fact that, by 2016, more than 700 genes had been identified as causes of intellectual disability and related disorders when mutated, most of which act in a dominant manner (Vissers et al., 2016). A similar scenario may apply to male infertility, which affects about 7% of men, as its likely high genetic heterogeneity and a *de novo* mutation paradigm could explain how it persists despite severe reproductive challenges. This idea aligns with known *de novo* genetic events causing NOA, such as Klinefelter syndrome and Y chromosome deletions. However, comprehensive large-scale studies to investigate *de novo* SNVs and CNVs in male infertility remain lacking.

To thoroughly explore potential dominant causes of male infertility, it is essential to carry out extensive studies that include large cohorts of patients and their parents in order to pinpoint harmful *de novo* and maternally transmitted mutations affecting male fertility. However, only a small number of researchers are involved in male infertility genetics, where they struggle with limited funding and patient participation (Barratt et al., 2021).

1.2.3.2 Challenges in Conducting Research in Male Infertility

Funding agencies and policymakers have only recently started to recognize human reproduction and infertility as issues of broad societal importance, and those affected by male infertility often continue to face stigma, even within their own families (Veltman & Tüttelmann, 2024). To address these challenges, research groups are increasingly collaborating in both formal and informal ways, for instance, through the International Male

Infertility Genomics Consortium (IMIGC). The IMIGC has driven various large-scale collaborative studies, identifying diverse autosomal recessive and X-chromosomal monogenic causes and underscoring the role of *de novo* mutations in male infertility (Oud et al., 2021; Nagirnaja et al., 2022; Riera-Escamilla et al., 2022). Furthermore, improvements in standardised clinical phenotyping help to reveal new candidate genes and diagnosis. Historically, the classification of male infertility phenotypes has lacked uniformity, often complicating cross-study comparisons and the establishment of genotype-phenotype correlations. To address this issue, Wyrwoll et al., 2024, integrated existing clinical guidelines, evidence-based thresholds, and expert consensus into the Human Phenotype Ontology (HPO) classification system. Their framework ensures that each phenotypic description is well-defined and logically positioned within a branching hierarchy of related terms. In this updated HPO system, a broad range of non-syndromic male infertility phenotypes is encompassed under one principal term, "Decreased male fertility," from which all further sub-phenotypes derive. Moreover, collaborative research efforts have led to recommendations for integrating genomic approaches into routine diagnostics (Wyrwoll et al., 2023; Stallmeyer et al., 2024).

1.2.3.3 The Importance of Genetic Diagnosis in Male Infertility

Establishing a molecular genetic diagnosis is critically important for affected men and couples, as it offers clarity and resolution regarding their condition (Veltman & Tüttelmann, 2024). Genetic diagnoses play an important role in clinical decision-making; for instance, deletions in the AZFa, AZFb, or AZFc regions on the long arm of the Y chromosome can accurately predict sperm recovery outcomes during TESE. Success rates can be as high as 50% in men with complete AZFc deletions, whereas they approach zero in cases involving complete AZFa, AZFb, or AZFbc deletions (Krausz & Riera-Escamilla, 2018). Recent studies have also identified monogenic causes that may predict TESE outcomes (Houston et al., 2021). However, small sample sizes currently limit the accuracy of these predictions, much like the early stages of AZF testing in the 1990s. Additionally, specific mutations, such as those in the *AURKC* gene, suggest that sperm are unsuitable for achieving viable pregnancies even through medically assisted reproduction (MAR) (Wyrwoll et al., 2023; Stallmeyer et al., 2024). Similarly, homozygous deletions in the *CATSPER2* gene highlight the need for ICSI (Stallmeyer et al., 2024). Pathogenic variants in genes like *DNAH1* are associated with MMAF, leading to infertility and requiring oocyte activation during MAR (Stallmeyer et al., 2024). Understanding the genetic basis of male infertility not only helps clinicians predict treatment success but also

enables them to counsel patients and their families on potential health risks beyond infertility and the effectiveness of MAR techniques (Veltman & Tüttelmann, 2024).

1.3 The Role of Structural Variations in Male Infertility

Known causes of male infertility include SVs such as CNVs (Carvalho et al., 2011; Tüttelmann, Simoni, et al., 2011; Dong et al., 2015; Krausz & Riera-Escamilla, 2018; Luo et al., 2019). Infertile men have been found to exhibit a higher burden of CNVs, particularly deletions, compared to fertile men. (Krausz & Riera-Escamilla, 2018). Additionally, it has been reported that oligozoospermic males carry ~4.6% of chromosomal aberrations, whereas chromosomal abnormalities have detected in up to 10-15% of azoospermic males (Carvalho et al., 2011).

aCGH and high throughput genome-wide sequencing has improved our understanding of CNVs role in male infertility in recent years (Tüttelmann, Simoni, et al., 2011; Dong et al., 2015; Krausz & Riera-Escamilla, 2018). However, little is known about the role of CNVs outside the AZF regions, as most studies have been small-scale and relied on low-resolution microarray approaches. Additionally, CNV analysis is not yet part of routine diagnostics for male infertility, unlike its established role in many other genetic diseases Nevertheless, researchers have identified CNVs causing male infertility, mainly using microarrays. These CNVs affect genes such as *EDDM3A, EDDM3B, HLA-DRB1, HLA-DQA1, POTEB, GOLGA8C, DNMT3L, ALF, NPHP1, NRG1, RID2, ADAMTS20, TWF1, COX10, MAK, DNEL1, MAST2,* and *TSPY1* (Shen et al., 2013; Dong et al., 2015; Huang et al., 2015). A more recent study by Luo et al. ,2019, identified a CNV that significantly decreased the expression of *CATSPER2* proteins in an infertile man with normal semen parameters. Another recent study by Wyrwoll et al. ,2022, identified the genetic cause of infertility in three patients, two with homozygous *SYCE1* deletions and one with a heterozygous *SYCE1* deletion combined with a pathogenic SNV on the other allele, using aCGH and exome sequencing. In addition, they identified *MLH3, EIF2B2, SLX4, CLPP,* and *TEKT5* as candidate genes, emphasising the advantages of exome-based CNV analysis over aCGH for future research (Wyrwoll et al., 2022). The same year, by utilising a combined genome-wide aCGH and WES approach, Hardy et al., 2022, detected pathogenic and likely pathogenic SNVs and CNVs in 15 patients (15%) with unexplained SPGF. A year after Xin et al., 2023, identified high-frequency CNVs-loci, including loci at Xp22.31 and 2p24.3, as well as candidate genes such as *VCX and NACAP9*, by performing karyotyping and CNV-seq[1] in 1157 azoospermia and

[1] CNV-seq is a low-coverage, massively parallel next-generation sequencing method used to detect chromosomal CNVs, offering a cost-effective and accessible alternative to traditional array-based methods.

oligospermia patients. It should be noted that the results are based on samples from a single-centre and have not been replicated. Additionally, aCGH or any other assay were not conducted to validate the identified CNV loci. Zhou et al., 2024, identified SNVs in 17 known causative genes and 12 candidate genes potentially involved in spermatogenesis and male infertility by performing WES and gene expression analysis in 167 infertile patients and 210 fertile controls. They also examined CNVs but found no pathogenic CNV in the cohort that could explain patients' infertility.

Moreover, researchers have identified repeat expansions (REs) in genes associated with male infertility, including *AR, HO-1 POLG, ATXN1, DMPK, ATXN3*, and *SHBG* (Dowsing et al., 1999; Mosaad et al., 2012; Siasi et al., 2011; Wagner et al., 2023). However, association studies often yield conflicting results, and replication studies generally failed to validate initial findings (Wagner et al., 2023). Therefore, further investigations are needed to clarify the relationship between REs and male infertility.

In the literature, researchers have investigated SVs that occur more frequently in infertile men compared to controls or have focused on specific regions associated with male infertility in both patients and unaffected individuals. The main challenge is that identifying one or two SVs in a patient cohort, with none detected in controls, does not provide strong evidence of pathogenicity. Large cohorts are necessary to leverage frequency data in both cases and controls to establish the clinical relevance of these variants. Also, for many reported CNVs, the lack of inheritance information makes it difficult to determine whether these variants warrant further investigation. Parental data could provide insight into the origin of a SV by determining whether it is inherited from a fertile father (reducing its likelihood as a causal factor in the proband's infertility), maternally inherited, or a *de novo* mutation arising in the germline of the affected individual. The first CNV analysis in patient-parent trios, conducted by my former colleague Kumara in our group using WES, identified two rare *de novo* CNVs involving a total of seven genes (Oud et al., 2021).

1.4 Structural Variations and Detection Methods

1.4.1 Structural Variations and Identification with Conventional Methods

Though there is no distinct characterisation of SV, they are usually considered as genomic rearrangement events over the size of 50 bp, divided into sub-categories as deletions (DEL), insertions (INS), duplications (DUP), inversions (INV), translocations (TRA), and tandem repeats (TRs) (Stankiewicz & Lupski, 2010; Fazal et al., 2020). Of these, DEL, INS, DUP and TRs are grouped under the broader classification of CNV (Figure 1.5).

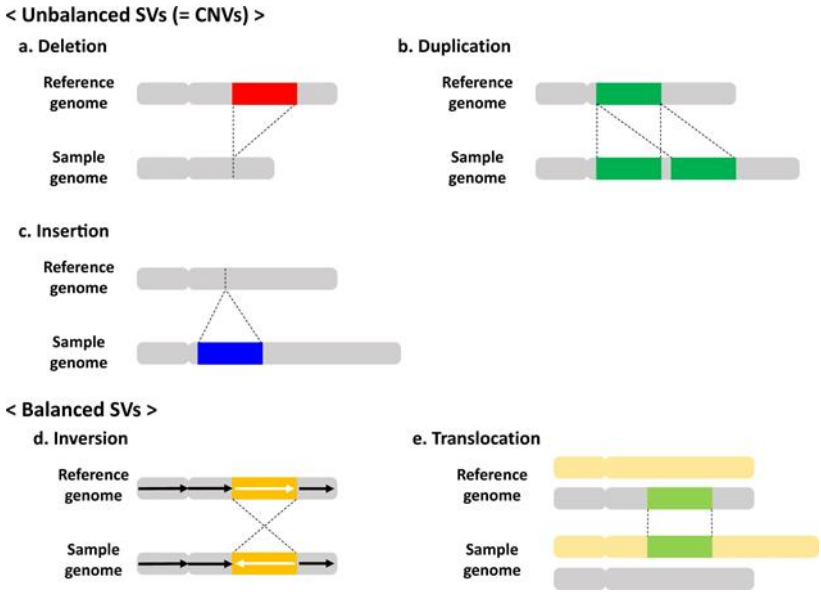


Figure 1.5. Illustration of SV types. i) Unbalanced SVs (CNVs); **a.** Deletion: Loss of a segment of DNA. **b.** Duplication: Copying of a DNA segment, resulting in multiple copies. **c.** Insertion: Addition of extra DNA sequences into the genome. ii) Balanced SVs; **d.** Inversion: Reversal of a DNA segment within the genome. **e.** Translocation: Rearrangement of DNA segments between non-homologous chromosomes without loss or gain of genetic material. (Nakatochi et al., 2021)

SVs arise through three main mechanisms: errors during DNA replication (polymerase slippage events), the movement of transposable elements within the genome, and improper repair of DNA breaks (either double-strand or single-ended) by processes non-homologous end-joining (NHEJ) and homologous recombination (NAHR) (Figure 1.6) (Hastings et al., 2009; Weckselblatt & Rudd, 2015; Scully et al., 2019).

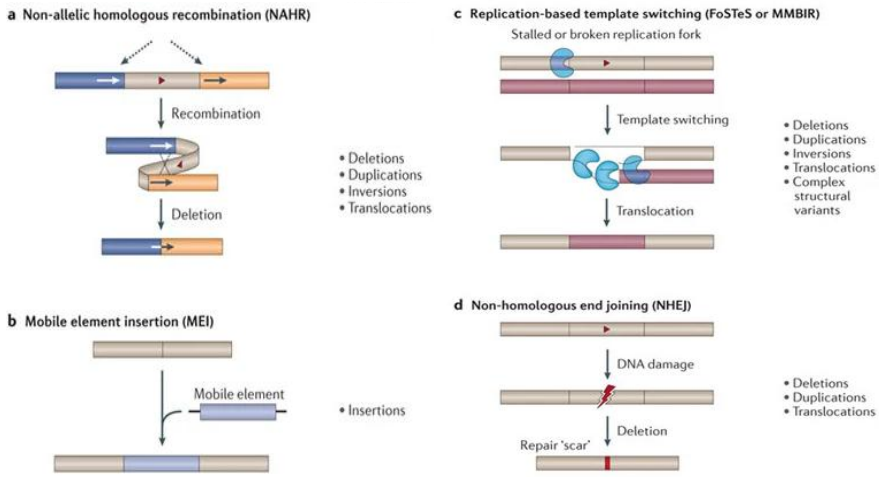


Figure 1.6. Mechanism of SVs (Weischenfeldt et al., 2013) (**A.** Recurrent SVs generally due to non-allelic homologous recombination (NAHR) which involves recombination between long highly homologous low-copy-number repeats (blue and orange segments). **B.** Novel genomic insertions might be caused by mobile element insertion of transposable elements by retrotransposition. **C.** DNA-replication-associated template-switching events, including the fork-stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR) mechanisms, can cause simple or complex structural variants, frequently resulting in duplicative events. **D.** Many SVs in humans are related to non-homologous end joining (NHEJ), which is a mechanism that repairs DNA double-strand breaks.)

SVs have been linked to several diseases such as Crohn's disease (McCarroll et al., 2008), schizophrenia (Walsh et al., 2008; Kirov et al., 2012), autism spectrum and other neurodevelopmental disorders (Pinto et al., 2010; Girirajan & Eichler, 2011; Elia et al., 2012). Though SVs have been linked to disease, they also play an important role in the diversity and evolution of the human genome (Stankiewicz & Lupski, 2010), with genomic differences between individuals being approximately 3-10 fold greater in SVs than single nucleotide variations (SNVs) (Pang et al., 2010; Alkan et al., 2011; Sudmant et al., 2015).

Traditionally, karyotyping and hybridisation-based approaches are used to ascertain SVs in patients/individuals. Though its low-resolution, karyotyping is still being used in clinics and is effective at detecting large (>5-10 Mb) chromosomal aberrations, in particular balanced translocations (Zhang et al., 2025). One of the hybridisation-based techniques is Fluorescence *in situ* hybridisation (FISH) using fluorescent probes that hybridize to complementary chromosomal DNA. FISH has the same capabilities as karyotyping in terms of detectable SV types but has higher resolution and is efficient at identifying sub-telomeric rearrangements (Linardopoulou et al., 2005). The other hybridisation-based technique is microarray comparative genome hybridisation (aCGH), which performs large-scale CGH at higher

resolution (Conrad et al., 2010). Though commonly used, aCGH can only detect certain types of SVs, has a lower sensitivity for small SVs, and has less capability to determine breakpoints than the sequencing-based methods (Kosugi et al., 2019).

1.4.2 SV Detection and Next-Generation Sequencing

Next-generation technologies and computational algorithms (Korbel et al., 2007; Bentley et al., 2008; McKernan et al., 2009; Alkan et al., 2011) have made profound improvements in SV detection and have superseded most other genomics approaches for SV detection. WGS enables us to detect various types of SV at base-pair resolution and allow of the integration of SV studies with studies of other genetic variants like SNVs and REs. WGS methodologies centre around short-read and long-read sequencing platforms, both of which have strengths and weaknesses.

1.4.2.1 Short Read Sequencing

Having broken DNA down into fragments, each fragment is sequenced from both ends (paired-end) using the most commonly sequencing-by-synthesis technique after labelling fragments. The size of fragments is generally in the range of 75-150 bp. This platform approach is commonly used in whole-exome sequencing (WES) and WGS. WES is solely focused on protein-coding regions of genomic DNA (~ 1.5% of the known genome) and requires capturing and enriching the coding sequence of the genome with PCR. By contrast, each base of the genome are sequenced in WGS (Slatko et al., 2018). Though WES is more economical than WGS, it excludes the intronic regions of the genome. WGS can make use of a PCR-free library preparation, providing the ability to determine exact breakpoints of large SVs along with the decreasing of PCR amplification biases and errors (Belkadi et al., 2015). Though these technologies enable comprehensive analysis of the genome, further techniques have been developed such as linked-read sequencing, Hi-C sequencing, long-read sequencing (LRS) and optical mapping to overcome challenges in some loci such as repetitive genomic regions or lengthy segmental duplications.

1.4.2.2 Long Read Sequencing

Single-molecule real-time (SMRT) sequencing developed by Pacific Biosciences (PacBio) and Nanopore sequencing developed by Oxford Nanopore Technologies (ONT) are the two most prominent long-read sequencing (LRS) platforms (Balachandran & Beck, 2020). LRS can generate reads >10kb and therefore requires high-molecular-weight DNA for best results

(Pollard et al., 2018). LRS has been shown as an effective technique to detect SVs, identifying SVs missed by the short-read sequencing in genomic studies. Despite higher error rates in SNVs and greater cost, LRS is a promising approach and is being used to unveil the entire genomic sequence and detect SVs with high confidence (Amarasinghe et al., 2020), particularly SVs occurring within long repetitive sequences (Chaisson et al., 2019). Crucially, this capability extends to accurately detecting and sizing pathogenic REs, which are notoriously difficult to resolve with short-read data. It is also important to highlight that LRS provides the added advantage of detecting DNA methylation directly from the native DNA strand. This allows for a more comprehensive analysis that integrates genetic and epigenetic information (Fu et al., 2025).

1.4.2.3 Bionano Optical Mapping

Although LRS offers kilobase to occasionally megabase read lengths, they average at approximately 15kb which is still insufficient to cover some large complex regions (Staňková et al., 2016). Recently developed next-generation mapping also uses very long double-stranded fragments of DNA (~300kb in length) labelled with fluorescent markers at specific sites. During the generation of optical maps, images of the fluorescent signal patterns are used (Yuan et al., 2020). This new approach offers greater detection rate of SV because of the absence of the pre-fragmentation steps required, allowing the analyses of ultra-high molecular weight (HMW) DNA molecules. Though this emerging technology still needs to prove its value within the clinical genetic diagnostic practice, it has been shown to contribute to the identification of translocation and inversion breakpoints, large insertions and deletions in the genome, and more complex SV (Ho et al., 2020; Yuan et al., 2020). One study showed that pathogenic structural variants missed by PCR-based techniques or chromosomal microarrays have been detected through next-generation mapping technology developed by Bionano (Barseghyan et al., 2017). Despite promising results further research is needed to affirm bionano optical mapping (BOM) efficiency and reliability.

1.4.2.4 Integrated Approaches

Each platform has its own unique advantages and disadvantages, integrating several approaches improves the analysis of inaccessible regions of the genome and the ability to generate a fully phased genome. For example, researchers have applied integrated approaches to elucidate genome sequencing of humans as telomere-to-telomere without gaps (Miga et al., 2020). These approaches have made profound changes in the understanding

of patterns of human genome (Levy-Sakin et al., 2019; Sethi et al., 2020; Ebert et al., 2021), and could help reveal underlying genetic causes of human diseases such as in cancer by enabling the detection of complex and large rearrangements in the genome (Dixon et al., 2018).

1.4.3 SV Variant Calling

1.4.3.1 Variant Calling

Recent advances in bioinformatic tools have been equally important to developments in sequencing technologies for improving the detection of SVs. In sequencing-based methods of SV detection, there are four sequencing signatures used to ascertain the SVs, these are read depth (RD), read pair (RP), split read (SR), and beta-allele frequency (BAF, also known as MAF: Minor Allele Frequency) (Figure 1.7) (Medvedev et al., 2009). Each method has its advantages and limitations in the different SV types, for example, the read depth method is known to accurately estimate copy numbers, but breakpoint resolution is generally poor (Alkan et al., 2011b). To overcome limitations and increase the accuracy of SV variant calling, combinatorial approaches are used by recently developed tools. These integrated approaches offer to increase the performance and reduce false-positive discoveries (Kosugi et al., 2019). Due to the technical and computational limitations, a vast amount of benchmarking and comparison studies have been done by researchers to determine the most efficient SV variant detection tools (Guan & Sung, 2016; Zare et al., 2017; Kosugi et al., 2019; Liu et al., 2020; Sarwal et al., 2020). Integrating multiple SV tools into a pipeline has been shown to increase specificity and sensitivity in identifying SVs, reducing the false positive rate (Lin et al., 2015).

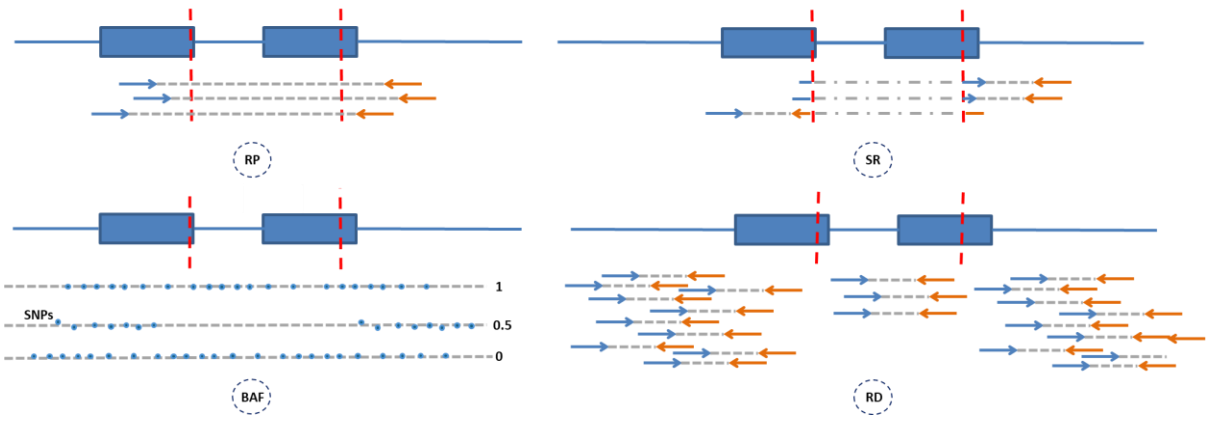


Figure 1.7. Signatures for SV detection in Next Generation Sequencing data. This figure shows a deletion as an example (indicated by two red vertical lines) in the reference genome (blue horizontal line, with rectangles representing exons) and the corresponding signatures observed in short-read sequencing data. RP (read-pair) analysis shows an increased distance

between mapped read pairs, SR (split-read) mapping identifies reads spanning the breakpoint, BAF (B-allele frequency) reveals the absence of heterozygous SNPs (BAF = 0.5) due to the presence of only one allele, and RD (read-depth) analysis indicates a reduction in sequencing coverage.

1.4.4. Challenges in Structural Variation Detection and Interpretation

There is an array of technological and computational tools, but no established standardised workflow to comprehensively analyse genome-wide SVs. The varied success of several methods is the result of different sequencing technologies and signals used (Medvedev et al., 2009). Repetitive and GC rich regions in the genome cause major challenges during sequencing and analysis of SVs. These difficult-to-sequence regions often lead to inconsistent results between different methods, causing significant variation in the number of SVs identified in the same genome. This variability highlights the need for integrated approaches that combine multiple methods to accurately identify SVs (Sarwal et al., 2020). Comparison and interpretation of SVs is an additionally challenging step, due to the lack of information present in the available databases, and WGS data storage and processing. Systematic annotation and prioritisation methods must be developed to overcome challenges.

1.5 Project Aims and Outline of the Chapters

This thesis investigates the contribution of SVs to the pathogenesis of severe male infertility phenotypes, specifically focusing on idiopathic NOA and severe oligozoospermia. Both of these phenotypes represent the most severe and rare forms of male infertility, making them the most likely to have an underlying genetic aetiology. Through WGS analysis, the study aims to characterise the spectrum of SVs, encompassing both inherited rare variants and *de novo* mutational events, and elucidate their functional implications in spermatogenic failure. The findings improve our understanding of the molecular aetiology of male infertility, enhance current diagnostic strategies, and provide evidence-based frameworks for genetic counselling of affected individuals.

This chapter provides an overview of male infertility, including definitions, prevalence, and classification; the fundamental biology of spermatogenesis; common risk factors and diagnostic approaches; and current treatments. It then delves into the genetics of male infertility, covering chromosomal anomalies, monogenic causes, and methods for uncovering the genetic basis of unexplained cases, before introducing structural variations and detection technologies.

In Chapter 2, I describe the cohorts of patients with idiopathic NOA and severe oligozoospermia. I then detail the sequencing methodologies, data analysis pipelines, the approaches used for validating the SVs and replicating the results.

In Chapter 3, I focus on optimising SV detection methods. I present results from *in silico* validation studies, compare SV calls from real WGS data, and contrast these findings with those from optical genome mapping. The discussion addresses the strengths and limitations of various tools and pipelines, culminating in a refined approach for SV detection in subsequent chapters.

In Chapter 4, I summarise the baseline structural variation landscape observed in the recruited cohort. I discuss how SVs are distributed across patients, compare probands to their parents where available, and interpret patterns that may hint at a potential role of SVs in male infertility.

In Chapter 5, the focus shifts specifically to *de novo* SVs. I detail the discovery of *de novo* deletions and duplications, describe the phasing of these variants, and interpret how such events might lead to spermatogenic failure.

In Chapter 6, I investigate the significance of rare inherited SVs, particularly those transmitted through the maternal line. I present analyses of autosomal and sex chromosome SVs, assess their potential pathogenicity, and describe a replication study conducted to validate promising findings.

In Chapter 7, I explore biallelic and compound heterozygous variations and include a systematic search for copy-neutral loss of heterozygosity (cnnLOH). I evaluate whether such recessive mechanisms contribute to idiopathic severe male infertility and discuss the clinical implications of these findings.

In Chapter 8, I investigate CNVs in WES data from a cohort of 234 patients with idiopathic quantitative male infertility, where parental DNA samples were not available.

In Chapter 9, Finally, I synthesise the results from all experimental chapters, contextualize the findings in the broader field of male infertility research, and discuss potential avenues for refining SV-focused diagnostics. I also propose directions for future investigations aimed at uncovering novel genetic contributors and improving patient outcomes.

Chapter 2. Material and Methods

2.1 Recruitment of Patients with Azoospermia and Severe Oligozoospermia

A total of 216 patient-parent trios and 234 singletons with unexplained non-obstructive azoospermia or severe oligozoospermia (with or without asthenozoospermia) were recruited from various centres. The 216 trios included 186 from Nijmegen (the Netherlands), 7 from India (FRIGE Institute of Human Genetics), and 23 from Newcastle (UK), while the 234 singletons comprised 11 from Sheffield, 145 from Nijmegen, 52 from Newcastle, and 26 from Manchester. Samples from Nijmegen were collected between 2015 and 2023. The remaining samples were collected after 2018 following the establishment of the Male Infertility Genomics group in Newcastle. An additional cohort of 29 patient-parent trios with idiopathic non-obstructive azoospermia was recruited as part of a collaborative project between Newcastle University and the University of Münster under the German Male Reproductive Genomics (MERGE) study.

To ensure consistency across centres, standardised protocols for sample collection, sequencing, and data processing were used. All patients were men under 50 years old, diagnosed with idiopathic azoospermia (no sperm in the ejaculate) or severe oligospermia (fewer than 5 million sperm/mL in at least two semen analyses). The reference values and semen nomenclature were used according to the WHO guidelines (World Health Organization, 2021). Blood samples were collected at the respective fertility centres, and saliva samples were obtained from the parents of trio participants. Saliva samples were used for parents since it provides a non-invasive and easy method for sample collection, thereby improving participant compliance. Saliva samples from all probands' parents were collected using the Oragene OG-500 kit (DNA Genotek, Ottawa, Canada). Y microdeletions and known chromosomal anomalies were screened as part of routine diagnostics in most laboratories using PCR and cytogenetic karyotyping assays, respectively, and were used as exclusion criteria. It should be noted that the samples from Sheffield were not screened for chromosome abnormalities and Y microdeletions. Additionally, patients conceived through ART were excluded from this cohort, so paternal transmission of pathogenic variants via ART is not expected. DNA was extracted from blood and saliva, and all participants gave written consent for genetic testing and use of their clinical data. The study was approved by the relevant Ethics Committees and/or Institutional Review Boards (Nijmegen: NL50495.091.14

version 4; Newcastle, Manchester, Sheffield: REC Ref: 18/NE/0089; India: FRIGE/IEC/25/2022; MERGE: nr. 2010-578-f-S).

2.2 Sequencing

2.2.1 Whole Genome Sequencing and Data Pre-Processing

Whole Genome Sequencing (WGS) was performed for all recruited 216 patient-parent trios, except for the German cohort. Having retrieved blood samples from probands and saliva samples from parents, DNA was extracted using the QIAGEN® Gentra® Puregene® DNA extraction kit using the vendor's protocol (QIAGEN®, Venlo, NL). Using Qubit and Nanodrop isolated DNA quality has been tested to determine whether these are suitable for the next process. Samples were prepared and enriched following the vendor's protocols for Illumina TruSeq DNA PCR-free® library preparation kit, then sequenced on the NovaSeq 6000 Sequencing System (Illumina) at the Genomic Core Facility (Newcastle University, UK) in the International Centre for Life. For blood samples, sequencing was performed with 24 samples per S4 flow cell (2 × 150 bp), and for saliva samples, 16 samples were run per S4 flow cell (2 × 150 bp). In cases where the minimal required depth of 30x coverage for robust variant calling was not achieved, samples were re-sequenced, and the data were then merged.

After BCL to FASTQ conversion and quality control using MultiQC (Ewels et al., 2016) and FastQC (Andrews, 2010), reads were aligned to the Genome Reference Consortium human assembly 38 (GRCh38) through BWA-MEM (version 0.7.17) (Li & Durbin, 2009). BWA-MEM generated SAM files, and then SAM files were converted into BAM files which were sorted by SAMtools (version 0.1.19) (Li et al., 2009). Duplicate reads were marked and removed by Picard tool (https://broadinstitute.github.io/picard/) after which BAM files were converted to the Compressed Columnar File (CRAM) file format via SAMtools (version 0.1.19) to optimise storage capacity. Regions were masked using the ENCODE blacklist (version 2), which identifies genomic regions exhibiting anomalous, unstructured, or excessively high signals in next-generation sequencing experiments, regardless of cell line or experimental conditions (Amemiya et al., 2019). The absence of reads in certain regions of CNV plots for both patients and parents may indicate that these areas were excluded from analysis. For quality control, the genomic data from each sample was examined in terms of sex, ancestry and relatedness using peddy (Pedersen & Quinlan, 2017). If mistakes were identified, i.e. either an incorrect sex or a lack of relatedness to the other patient or parents, we examined for sample swaps

within a run and corrected if identified. Flagged samples that could not be resolved were excluded. These data pre-processing steps were performed by Bioinformatics Support Unit (BSU) at Newcastle University and by Dr Miguel Xavier.

2.2.2 Whole Exome Sequencing and Data Pre-Processing

Whole Exome Sequencing (WES) was performed by the Genomic Core Facility (Newcastle University, UK) at the International Centre for Life for all singletons and German patient-parent trios. WES samples were prepared and enriched following the manufacturer's protocols of either Illumina's Nextera DNA Exome Capture kit or Twist Bioscience's Twist Human Core Exome Kit. All sequencing was performed on the NovaSeq 6000 Sequencing System. Sequencing was performed with 96 samples per S2 flow cell (2 × 100bp). Sequence reads were then aligned to the Genome Reference Consortium human assembly 38 (GRCh38) through BWA-MEM (version 0.7.17) (Li & Durbin, 2009). The sex, ancestry and relatedness of each sample was calculated using peddy (Pedersen & Quinlan, 2017). These data pre-processing steps were performed by Dr Miguel Xavier.

2.2.3 Bionano Optical Genome Mapping with Initial Data Processing

Nine samples were selected for this approach based on high-quality short-read sequencing data, as high-quality DNA was required for analysis on the Bionano Optical Mapping platform. Bionano Prep SP Frozen Human Blood DNA Isolation Protocol v2 (Bionano Genomics) was applied using Bionano Prep SP DNA Isolation Kit (Bionano Genomics) to extract HMW DNA which was then labelled using Bionano Prep DLS Kit (Bionano Genomics). Afterwards labelled DNA was scanned through Saphyr (Bionano Genomics). HMW DNA is required to sequence long stretches of DNA.

SV analysis in OGM was performed using the Rare Variant Analysis pipeline in Bionano Access 1.6.1. The last output file format is a SMAP file that contains a list of annotated SVs. Detailed information about each SV call was stored in a tab-delimited, text-based format in an SMAP file. Also, the CNV pipeline based solely on the depth of coverage of reads was run on Bionano Access and results were retrieved in TXT format. Using just depth of coverage signature in this pipeline caused a high-rate of false positive calls due to technological limits. These SMAP and txt files were used for further analysis.

2.3 Data Analysis

2.3.1 Generating *In Silico* Genome Sequencing Data for Optimisation of CNV Calling

In silico data sets were constructed to optimise CNV calling with the help of Dr Aneta Mikulasova. Two separate WGS datasets were generated for deletions and duplications at around 30X coverage. The datasets were constructed with CNVs of different types and sizes (665 losses and 176 gains, ranging from 2 kb to 1 Mb in size). Firstly, the ClinVar CNV database was downloaded and relevant information (CNV type, start and end positions) was extracted using a custom-made R script. CNVs classified as pathogenic or likely pathogenic, with sizes between 2 kb and 1 Mb, and that were located at least 100 kb apart from each other were then filtered. Using this filtered CNV list, FASTA files were generated by simuG (1.0.0) which incorporated the CNVs into a simulated genome (Yue & Liti, 2019). ART (version MountRainier-2016-06-05) was then used to simulate raw FASTQ files from the simulated genome (Huang et al., 2012). The same protocol was applied as described in the sequencing section above for aligning FASTQ reads to the reference genome. GRCh38 reference genome was used in all steps.

2.3.2 SV Calling in Method Optimisation Study

Initially, for optimisation, CNV analysis in *in-silico* WGS data was performed by LUMPY (Layer et al., 2014), DELLY (Rausch et al., 2012), Manta (Chen et al., 2016), dysgu-sv (version 1.1.8) (Cleal & Baird, 2021) and GATK-based CNVRobot (version 3.2) (https://github.com/AnetaMikulasova/CNVRobot) with default parameters. sv-callers (version 1.1.0) (Kuzniar et al., 2020) was used, which integrates LUMPY, DELLY and Manta. Docker (Merkel, 2014) images were built and pushed onto docker hub (https://hub.docker.com/u/ozyyk61) for sv-callers, CNVRobot and dysgu-sv. The created docker images compile all dependencies, making the tools ready-to-run, while supporting automated deployment and reproducibility.

2.3.3 SV Calling in WES and WGS Data

SV calling in WGS trios was conducted using a combination of CNVRobot (version 3.2) and dysgu-sv (version 1.1.8) tools as per the method optimisation study results (see chapter 3). CNV calling in WES data was performed with CNVRobot (version 3.2). CNVRobot was selected for WES analysis due to its methodological strengths, which are particularly well-suited for targeted sequencing data. The tool integrates read-depth information with SNP zygosity,

providing a robust and reliable method for identifying CNVs where other signatures, such as discordant and split reads, are not available. Furthermore, the use of CNVRobot for both WES and WGS ensured data compatibility. From a practical standpoint, the tool's detailed plots greatly facilitate the visual inspection and validation of all variant calls, and it is a well-maintained tool with direct developer support. All data processing was performed on a server with 24 CPU and ~1.5 terabytes of memory.

2.3.3.1 CNVRobot

CNVRobot (https://github.com/AnetaMikulasova/CNVRobot) is a comprehensive tool for detecting CNVs and loss of heterozygosity (LOH) in human sequencing data. It integrates read-depth information and SNP zygosity to identify deletions, duplications, and regions of copy-neutral LOH with high sensitivity, making it well suited for rare germline CNV discovery as well as somatic alterations in tumours. Crucially, it works on short-read data from any NGS platform, including WES, targeted sequencing (TS), or WGS. To reduce systematic noise caused by PCR amplification, GC content, and variable capture efficiency, CNVRobot relies on a Panel of Normals (PON). A minimum of 40 controls are recommended, though more controls significantly enhance the normalisation step and improves CNV detection accuracy. Herein 100 controls were used (data of 50 mothers and 50 fathers), which was the maximum number storable on the server at one time due to the large size of the BAM files. While we used a PON of 100 controls, it should be acknowledged that using saliva-derived DNA from parental samples introduces a potential technical consideration for CNV analysis, as CNV detection is highly sensitive to the DNA source and quality. This is particularly relevant for *de novo* calling, where subtle differences in technical noise could lead to false positives. To mitigate this, all candidate *de novo* CNVs were individually examined using CNVRobot plots and the IGV.

CNVRobot uses GATK4 to build a coverage model and correct for technical biases in WES, TS and WGS data, then applies an R-based segmentation algorithm to pinpoint CNVs. This segmentation step incorporates SNP allele frequencies to detect subtle LOH. For visualisation, CNVRobot generates detailed genome-wide and chromosome-wide plots, plus additional “detail plots” for each abnormal region. The visualisation was highly adaptable for both singleton and trio analyses, facilitating trio-based visualisation and case-control comparisons. It also generates data outputs in BigWig and BED formats for interactive browsing in software such as Integrative Genomics Viewer (IGV).

CNVRobot run details are specified across a set of “master files” that provide paths to your BAM files, reference FASTA, and capture BED (if working with TS or WES). CNVRobot is started using its run.sh script, which orchestrates coverage collection, denoising, segmentation, and the generation of all report files and plots. For reproducibility, it should be noted that the bin size used for coverage collection was set to the default, automatically using 1000bp. Similarly, default segmentation conditions were applied for CNV calling.

2.3.3.1.1 CNV Calling in Sex Chromosomes

Initially, CNVRobot version 3.2 was used, which was the latest available version at the time I began data processing. However, when the latest version of CNVRobot was released in October 2023 (version 4.2), featuring improvements in gonosomal analysis by doubling the coverage of chromosome X and Y in male samples and controls, we decided to use the updated version specifically for analysing sex chromosomes. Reprocessing all the data was not feasible within the given timeframe. Fortunately, the updated version allowed me to reuse the existing read counts and re-analyse the sex chromosomes without the need to reprocess the entire dataset.

2.3.3.1.2 CNVRobot Visualisation and Plot Description

The CNVRobot pipeline generates three different plots: whole genome profile plots (Figure 2.1), plots for individual chromosomes (Figure 2.2), and CNV plots (Figure 2.3). Each plot presents data for three individuals. The lower panel (e.g., proband NIJ_MI_0584 in Figure 2.1) shows the individual who is examined, while the middle (mother NIJ_MI_0584) and upper panels (father NIJ_MI_0584) can represent either the parents or male and female controls. Whole genome profile plots depict all chromosomes at once, allowing the detection of whole-chromosome gains or losses and confirmation of each individual’s sex (Figure 2.1). CV plots display a single chromosome, providing a closer look at the regions surrounding a CNV (Figure 2.2). They are also useful for identifying noisy samples, which show highly variable log2ratio signals across targets and tend to produce many CNV calls.

Figure 2.1. A whole genome profile plot for the NIJ_MI_0584 trio. From top to bottom (Father, Mother, and Proband), each track features a log2ratio illustrating the copy number status of the sequenced segments on every chromosome, alongside a minor allele frequency track indicating SNP zygosity through all chromosomes.

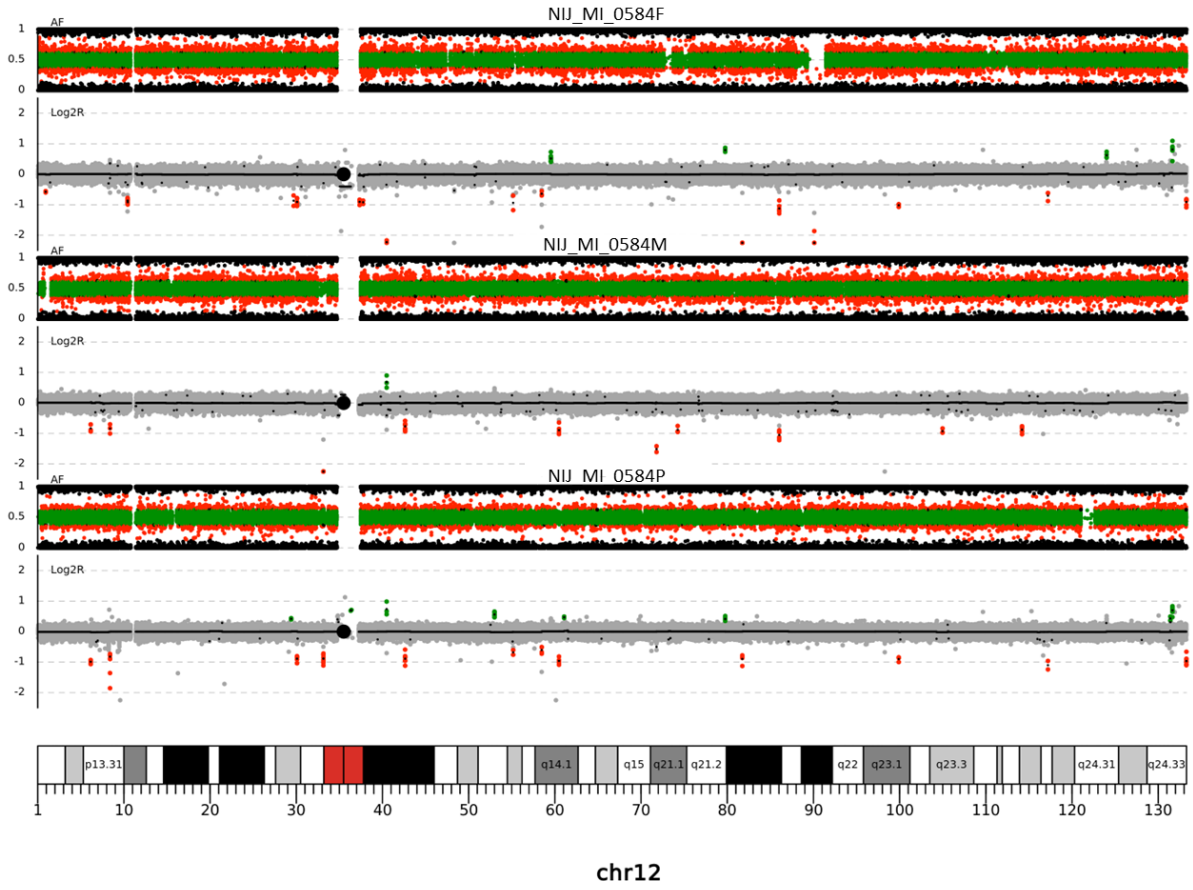


Figure 2.2. A chromosome view plot for chromosome 12 of the NIJ_MI_0584 trio. Displayed from top to bottom (Father, Mother, and Proband), each individual has a log2ratio plot showing the copy number status of all sequenced segments on chromosome 12 and a minor allele frequency plot depicting the allele frequency at each SNPs along the chromosome 12.

CNV plots illustrate specific deletions or duplications in detail (Figure 2.3). Each sequencing target appears individually, enabling fine-scale examination of the region of interest. In all three plot types, each sample has a log2ratio track and a MAF track. The log2ratio is calculated as:

log2ratio = log2 (observed normalised coverage / expected normalised coverage)

A log2ratio of 0 indicates normalised coverage matching the reference (i.e., the same copy number). In autosomes, a value of -1 implies the loss of one copy, while -2 implies the complete absence of the locus. When analysing the sex chromosomes, a single copy in the male reference also appears as a log2ratio of 0, and loss of that copy is shown as -2. In female samples, having two copies of the X chromosome corresponds to a log2ratio of 1 for chromosome X, while the Y chromosome is absent (log2ratio = -2). It is important to note that the interpretation of normalised coverage data as indicative of copy number (CN) relies on the

assumption of a diploid genome, where the most frequent coverage level is standardised to CN=2. However, in polyploid organisms, this assumption may not hold, as median-centred normalisation could align the most common coverage to CN=2, while the actual copy number may be higher (e.g., CN=3, CN=4, or greater), depending on the ploidy level. This inherent limitation applies to all coverage-based CNV detection methods, including array-CGH. Nonetheless, in the context of our analysis, which focuses on germline samples from healthy diploid individuals, this assumption is unlikely to introduce significant bias.

The MAF track is defined as:

MAF = B / (A+B)

where A and B are the read counts of the two alleles (B represents the read count of the minor allele in the population and A represents the count of the major allele). MAF values of 1 or 0 indicate homozygosity, whereas 0.5 denotes heterozygosity. This helps confirm losses of heterozygosity (e.g., in deletions, heterozygous SNPs disappear) and identify regions of copy number neutral loss of homozygosity.

CNV plots also include additional annotations below the log2ratio profiles:

- Cytoband information for the genomic region.
- A control track (yellow bars) indicating how frequently each target was lost or gained in control samples. Taller bars signify more control data for that target; red shading within a bar indicates multiple control samples with losses, and green shading indicates gains.
- DGV variants (Database of Genomic Variants), depicted as horizontal bars in green (duplications), red (deletions), or purple (complex variants). The more intense the color, the more frequently that variant has been reported. This provides a quick indication of how variable the locus is in healthy populations.
- A genome mappability track, where black shading represents high mappability and a transition toward white indicates lower mappability.
- RefSeq genes, shown at the bottom of the plot to help interpret whether any genes of interest fall within the CNV.

Through these visualisations, CNVRobot makes it possible to pinpoint and characterize copy number changes, assess their frequency in control samples, and distinguish frequently variable regions (possible polymorphisms) from rare or potentially pathogenic variants.

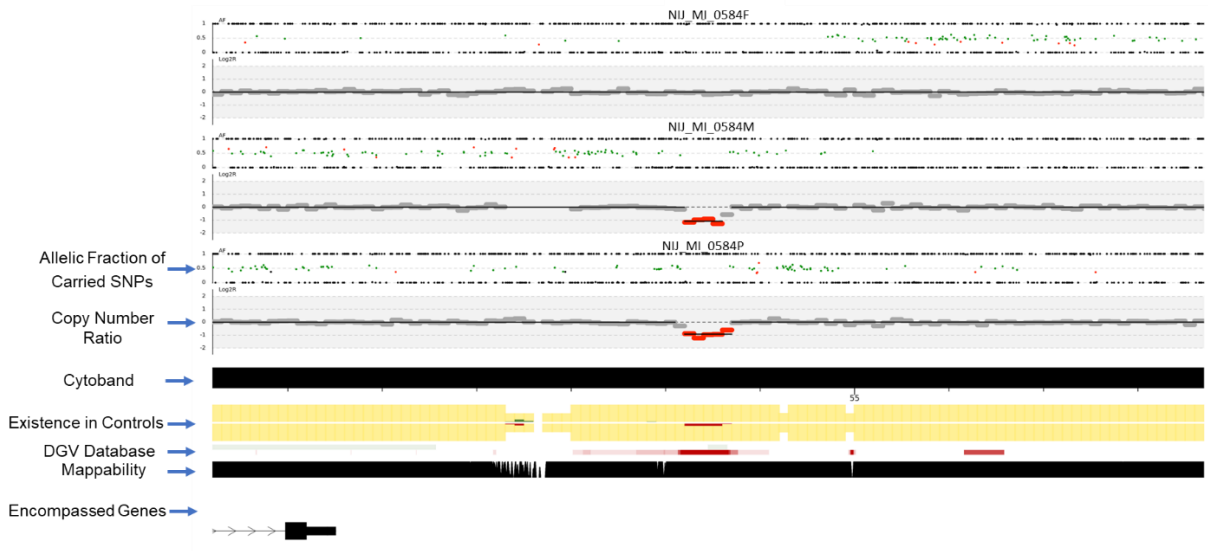


Figure 2.3. CNVRobot plot for a maternally inherited heterozygous deletion detected in proband NIJ_MI_0584. The plot includes the Log2 Ratio and Minor Allele Frequency (MAF) tracks, along with additional informative tracks. The Cytoband Information indicates the specific genomic region affected. The Control Track (yellow bars) shows how often each target was lost or gained in control samples, with red shading indicating losses and green shading indicating gains. DGV variants (green for duplications, red for deletions, and purple for complex variants) reflect how frequently a variant has been reported, with darker colors showing higher frequencies. The Genome Mappability Track uses black shading for high mappability and white for lower mappability (full black is 100% mappabiity and full white 0% mappability). RefSeq genes are displayed at the bottom to highlight any genes of interest within the CNV region. The heterozygous deletion is present in both the mother and the proband, confirmed by the absence of heterozygous SNPs (MAF = 0.5) and a Log2ratio value of -1 Additionally, it can be seen that the deletion is observed in many control samples, as shown by presence of red bars in the control track, and is highly frequent in the DGV database, indicated by the dark red region.

2.3.3.2 Dysgu-sv

Dysgu-sv was used for SV detection (version 1.1.8) in tandem with CNVRobot to broaden our coverage of SVs (Cleal & Baird, 2021). Dysgu-sv combines multiple signals, including split reads and discordant read pairs, thereby enabling detection of both unbalanced events (e.g., deletions, duplications) and balanced events (e.g., inversions, translocations). Split reads also enable us to identify breakpoints at single base pair resolution. Dysgu-sv flags of *-lengh* 1000 and *-support-read 10*, were applied to reduce false positives by focusing on variants ≥1000 bp supported by at least 10 reads. Following dysgu-sv's machine-learning-based scoring, calls with probability values below 0.5 were filtered out. Unlike CNVRobot, which can natively assess inheritance, dysgu-sv does not provide family-based prediction. Therefore, after running dysgu-sv, a custom R script was used to merge and compare dysgu-sv calls across

proband-parent trios, assigning inheritance where at least 70% of an SV's length overlaps among family members. For visualisation, a custom-made IGV batch script was used to generate IGV screenshots for interpreting individual SVs (Robinson et al., 2011). This complementary dysgu-sv pipeline, especially its capacity to detect balanced structural changes, enhances our overall variant discovery beyond the copy-number-focused results of CNVRobot, ultimately yielding a more comprehensive view of SVs across the genome.

2.3.3.3 Combining CNVRobot and dysgu-sv Calls

To combine SV calls from CNVRobot and dysgu-sv, a custom R script was developed to integrate results from both tools. The datasets were cleaned, after which the script used bed_intersect from the valr package to identify overlapping variants between the two tools (Riemondy et al., 2017). Variants were considered overlapping if they shared the same type (e.g., LOSS vs. DEL or GAIN vs. DUP) and exhibited at least 70% reciprocal overlap based on their genomic coordinates. For overlapping events, the start and end positions from dysgu-sv were retained, as it provides better breakpoint resolution due to its use of split-read and discordant-read signals. The output was divided into three categories: (1) concordant calls supported by both tools, (2) unique calls detected only by CNVRobot, and (3) unique calls detected only by dysgu-sv. Overlapping events were prioritised, while non-overlapping events were retained to capture potentially unique findings.

2.3.4 Variant Annotation and Interpretation

All SVs identified in this study were annotated using AnnotSV (version 3.1.1) with default parameters (Geoffroy et al., 2018). AnnotSV consolidates information from several databases and resources, including GnomAD (v2.1.1), ClinVar, ClinGen, the DGV, the Deciphering Developmental Disorders (DDD) project, and the 1000 Genomes Project (1000g), to provide comprehensive annotations for each SV. These annotations cover the affected genomic regions, overlapping or nearby genes, predicted functional consequences, variant frequency in reference populations, and any known clinical associations. Two primary settings were applied during annotation: (1) a minimum overlap threshold of 100% between user features and annotated SVs to be reported, and (2) an allele frequency threshold of 1% for identifying common variants in the respective databases (GnomAD (v2.1.1), ClinVar, ClinGen, DGV, the DDD project, and the 1000g). Consequently, SVs that overlapped 100% and were present at frequencies greater than 1% in respective databases were classified as common variants, while the remainder were considered rare and were examined more closely for potential

pathogenicity. It should also be noted that comparisons against databases were restricted to SVs matching type (e.g., deletions with deletions).

2.3.4.1 Additional Databases Employed during SV Interpretation

In addition to the population databases mentioned earlier, gnomAD-SV version 4 (released November 2023) was used extensively for SV interpretation. This resource integrates data from 807,162 individuals, nearly five times larger than the combined v2/v3 releases, and is split into two callsets: exome sequencing data from 730,947 individuals (including 416,555 from the UK Biobank) and genome sequencing data from 76,215 individuals. Given its unprecedented scale, this database was particularly valuable for characterizing individual SVs.

The DECIPHER database (version 11.29) (https://www.deciphergenomics.org) was also consulted for CNV data, as it contains entries from over 36,000 patients exhibiting various pathogenic phenotypes (Firth et al., 2009). Although DECIPHER offers insight into microdeletion and microduplication syndromes frequently spanning multiple megabases and genes, its usage in this study was approached cautiously. Pinpointing the pathogenicity of CNVs associated with infertility among the numerous genes listed can be complex, given that many of these structural variations are linked primarily to developmental disorders and might not directly inform fertility outcomes.

To help interpret SVs, RNA and protein expression levels of selected genes in human tissues were obtained from the Human Protein Atlas (www.proteinatlas.org), a comprehensive project aimed at mapping human protein distribution across cells, tissues, and organs (Uhlén et al., 2015).

STRING (string-db.org) was used to retrieve known and predicted protein-protein interactions, covering 24,584,628 proteins from 5,090 organisms (1st January 2025)(Szklarczyk et al., 2023).

Additionally, DOMINO (domino.iob.ch) was used in predicting inheritance patterns for genes (Quinodoz et al., 2017). The DOMINO scores were used while filtering dominant genes (>0.5 scores indicate possibility of dominant inheritance).

The UCSC Genome Browser (http://genome.ucsc.edu), which interactively visualizes genomic data collected from a wide range of resources (Perez et al., 2024), was used primarily to investigate regulatory elements within genomic regions. In addition, the ENCODE database

was utilised to examine regulatory regions across the genome (ENCODE Project Consortium, 2012).

Literature searches were conducted on PubMed (https://pubmed.ncbi.nlm.nih.gov).

2.3.4.2 SV Prioritisation and Interpretation

While prioritising SVs with potential clinical relevance to male infertility, I employed different strategies and thresholds depending on the inheritance models under investigation. All prioritisation strategies are detailed in the relevant chapters or sections. However, two complementary approaches were primarily followed. First, Known and Candidate Male Infertility Genes (KCMIG), described below, were investigated. Second, efforts were made to identify novel candidate genes by analysing rare SVs (population allele frequency <1%), prioritising those affecting constrained genes based on pLI score (Collins et al., 2020), which indicate tolerance to loss-of-function (LoF) mutations, and considering gene inheritance models predicted by DOMINO (Quinodoz et al., 2017).

All prioritised SVs were visually inspected using IGV. Additionally, the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen) have established a scoring framework for classifying SVs based on an evidence-based method (Riggs et al., 2020). While acknowledging the key role of these standards in clinical interpretation, this study primarily aims to reveal new candidate genes rather than providing definitive genetic diagnoses for patients.

Creating Known and Candidate Male Infertility Genes (KCMIG) List

The gene prioritisation framework was developed through systematic integration of multiple evidence sources. The foundation was established using the comprehensive male infertility gene catalogue by Houston et al., (2021), comprising 657 genes with evidence-based classifications. These classifications were quantified using a 5-point scoring system: Definitive (5), Strong (4), Moderate (3), Limited (2), and Unable to classify/No evidence (1). Genes scoring above 2 were considered as known disease genes.

This core dataset was supplemented with additional candidate genes from multiple sources. A preliminary gene list developed by Dr. Giles Holt incorporated data from Protein deep profiling and model predictions for male infertility (Xu et al., 2021), catalogues of human genes associated with pathozoospermia (lowered semen quality) (Ignatieva et al., 2021), and the Male Infertility Knowledgebase (Shaini Joseph & Smita D Mahale, 2021). Further candidate

genes were identified through large-scale cohort studies (Oud et al., 2021; Nagirnaja et al., 2022), adding 195 genes. Literature review of 60 post-2020 publications yielded 13 additional candidates. The dataset was further expanded through analysis of male infertility-related subphenotypes (MONDO:0005372) in OMIM, including spermatogenic failure, androgen insensitivity, adrenal hyperplasia, gonadal dysgenesis, hypogonadotropic hypogonadism, hyperprolactinemia, leprosy, pituitary hormone deficiency, and trichothiodystrophy, identifying 2 additional genes.

The frequency of gene citations across these resources was quantified using a tiered scoring system: single mention (1C), 2-5 mentions (1B), and >5 mentions (1A). While not as systematically curated as the Houston et al. (2021) review (as a comprehensive analysis is beyond the scope of this thesis), this scoring system provided an evidence-based framework for variant prioritisation. The final compiled dataset encompassed 3,437 genes with evidence scores ranging from 1C to 5 (Table 2.1).

Table 2.1 The number of genes in each assigned score in Known and Candidate Male Infertility Genes list

Score	5	4	3	2	1	1A	1B	1C
No. of gene	48	28	44	138	399	67	223	2490

2.3.5 Data Exploration/Manipulations, Statistical Tests and Plots

All data exploration, manipulation, statistical testing, and visualisation were performed using basic terminal commands, custom-made Bash scripts, and R scripts. I used R version 4.1.0 (2021-05-18) for data analysis. For general data analysis, I used the valr (version 0.8.3) (Riemondy et al., 2017), tidyverse (version: 2.0.0) (Wickham et al., 2019), and dplyr (version 1.0.6) (Wickham et al., 2021) packages. To generate plots, I employed the GenomicRanges (version 1.58.0) (Lawrence et al., 2013), ggplot2 (version 4.1.2) (Wickham, 2016), GenVisR (version 1.38.0) (Skidmore et al., 2016), ggpubr (version 0.4.0) (Kassambara, 2020), and karyoploteR (version 1.32.0) (Gel & Serra, 2017) packages. For outlier detection, I used the outliers package (version 0.15) (Komsta, 2022), and for statistical tests, including the chi-square test, ANOVA, and Tukey's test, I used the janitor (version 2.1.0) (Firke, 2021), lmerTest (version 3.1.3) (Kuznetsova et al., 2017) and stats (version 3.6.2) (R Core Team, 2021) packages.

2.4 SV Validation

All validation experiments were conducted with the help of senior lab technician Dr. Bilal Alobaidi. The ten *de novo* SVs, including eight deletions and two duplications (see Chapter 5), as well as a maternally inherited inversion (see Chapter 6) were validated. All of these variants were in the heterozygous state except for the *de novo* deletion on chromosome X.

We validated the eight *de novo* deletions and the maternally inherited inversion using a PCR assay followed by Sanger sequencing. I designed primers for both the intact and the altered alleles using Primer3 (version 0.4.0) (Untergasser et al., 2012). Then, PCR reactions were performed using AmpliTaq 360 DNA Polymerase (ThermoFisher, MA, USA). After performing PCR with both sets of primers with controls, we analysed the products on 1% agarose gels, which confirmed the presence of both the intact and altered alleles. Next, we performed Sanger sequencing on an Applied Biosystems SeqStudio Genetic Analyzer (ThermoFisher, MA, USA) to validate the DNA breakpoints.

To validate the two *de novo* duplications, we employed a quantitative PCR (qPCR) assay. I designed qPCR primers for both the duplicated regions and the control regions, then performed the assays using the Applied Biosystems QuantStudio 7 Flex Real-Time PCR System (ThermoFisher, MA, USA) to detect copy number changes. By comparing the relative copy number in each sample to a reference region, we confirmed the presence of additional copies in the duplicated regions.

2.5 Replication Study

The most promising candidate genes identified through the SV analyses in this thesis were further evaluated in an independent cohort of infertile men, as well as a fertile control cohort. We also investigated potentially pathogenic SNVs in these candidate genes in our Genomics of Male Infertility Group cohort, which is primarily comprised of samples from the Nijmegen/Newcastle cohort.

2.5.1 SNVs in our Genomics of Male Infertility Group Cohort

Dr. Miguel Xavier extracted variants from our patient-parent trios, in which patients underwent WGS, and from singleton cases, in which WES was performed. The total number of patients was 794 (processed by the end of 2024.).

Analyses were conducted following previously introduced methods (Oud et al., 2021), with improvements implemented as our group transitioned to WGS. Briefly, variants with an allele frequency greater than 1% in the gnomAD database were excluded to prioritise rare variants. Variants with fewer than ten reads and/or fewer than 15% of reads containing the mutation were removed. Synonymous variants and those with no impact on protein function, including non-protein-altering splice site variants, were also removed. Pathogenicity was evaluated using SIFT, MutationTaster, and PolyPhen pathogenicity predictors. Variants were then classified based on the ACMG and the Association for Molecular Pathology (AMP) 2015 guidelines.

2.5.2 German Male Reproductive Genomics (MERGE) Cohort

We investigated exome data from the MERGE study, comprising 887 men with azoo-, crypto-, or severe oligozoospermia. Known genetic causes of male infertility, including chromosomal aberrations and microdeletions of the AZF region, were excluded beforehand. WES was carried out as described previously (Wyrwoll et al., 2020). Genetic variants within the candidate genes were extracted from this dataset, and variants with fewer than ten reads and/or fewer than 15% of reads containing the mutation were removed. To ensure consistency, genomic coordinates were converted to GRCh38 where necessary and reannotated with Variant Effect Predictor (VEP). The same variant prioritisation and interpretation mentioned above were applied by Dr. Miguel Xavier.

2.5.3 Fertile Control Cohort: Dutch Parents

We utilised an anonymised exome dataset from 5784 Dutch men and 5803 Dutch women, all confirmed to have at least one child, as a control group. These individuals, sequenced at the Radboud Diagnostics Center, were healthy parents of children with severe illnesses and underwent routine exome sequencing. Despite their children's intellectual disabilities, the fertility of these men and women is presumed to align with that of the general Dutch population. Genetic variants within the candidate genes were retrieved from this dataset and analysed using a variant prioritisation and interpretation strategy akin to that employed by Dr. Miguel Xavier, with adaptations performed by myself. One difference in my approach was the inability to assess variant read counts due to their absence in the dataset. Additionally, as I focused on LoF variants I incorporated SpliceAI and CADD scores for prediction of pathogenicity. Specifically, for frameshift, stop-gain, stop-loss, and start-loss variants, pathogenicity was determined using a "best-of-three" predicted pathogenic approach with a

CADD score threshold of ≥30, while splicing variants were assessed using the same "best-of-three" method with a SpliceAI score threshold of >0.8.

Chapter 3. Method Optimisation for Structural Variation Detection in the Human Genome

3.1 Introduction

SVs are implicated in a wide range of disorders, including mendelian diseases such as intellectual disabilities and autism spectrum disorders (Pinto et al., 2010; Girirajan & Eichler, 2011; Elia et al., 2012)., as well as genetic syndromes like DiGeorge syndrome (McDonald-McGinn & Sullivan, 2011), Prader-Willi/Angelman syndromes (Paparella et al., 2023), and various other microdeletion/microduplication syndromes. Beyond monogenic conditions, SVs also contribute to complex polygenic disorders, such as schizophrenia (Kirov et al., 2012), and serve as critical drivers in cancer pathogenesis (Cosenza et al., 2022).

Traditionally, karyotyping and hybridisation-based approaches are used to ascertain SVs in patients/individuals. Emerging state-of-the-art next-generation technologies and computational algorithms are making profound changes in SV detection (Korbel et al., 2007; McKernan et al., 2009; Bentley et al., 2008; Alkan et al., 2011). In sequencing-based methods of SV detection, there are four sequencing signatures used to ascertain the SVs, these are read depth (RD), read pair (RP), split read (SR), and B-allele frequency (BAF) (also known as MAF: Minor Allele Frequency) (Figure 3.1.) (Medvedev et al., 2009).

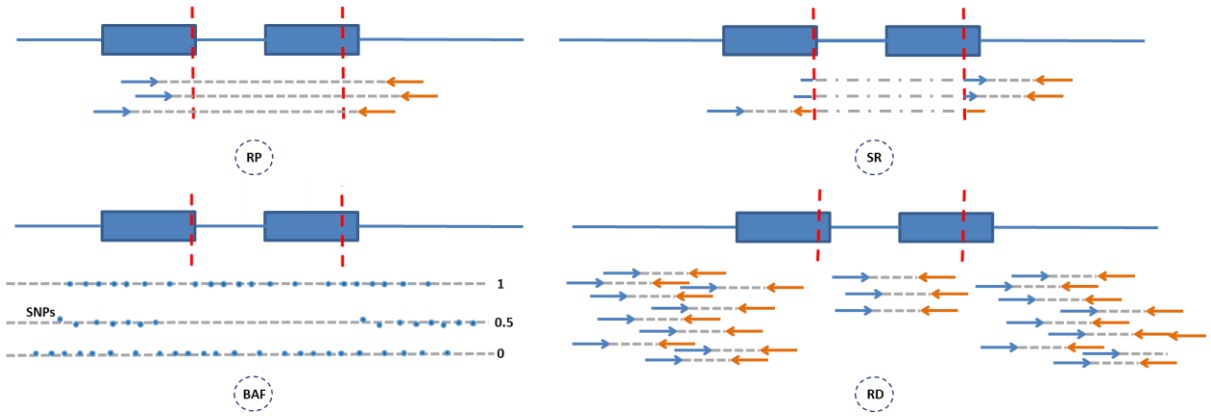


Figure 3.1. Signatures for SV detection in Next Generation Sequencing data. This figure shows a deletion as an example (indicated by two red vertical lines) in the reference genome (blue horizontal line, with rectangles representing exons) and the corresponding signatures observed in short-read sequencing data. RP (read-pair) analysis shows an increased distance between mapped read pairs, SR (split-read) mapping identifies reads spanning the breakpoint, BAF (B-allele frequency) reveals the absence of heterozygous SNPs (BAF = 0.5) due to the presence of only one allele, and RD (read-depth) analysis indicates a reduction in sequencing coverage.

Each method has its advantages and limitations for the detection of the different SV types. For example, the read depth method is known to accurately estimate copy numbers, but resolution is limited due to its interval-based coverage collection approach (Alkan et al., 2011). To overcome limitations and increase the accuracy of SV variant calling, combinatorial approaches are used by recently developed tools. Integrating multiple SV tools into a pipeline has been shown to increase specificity and sensitivity in identifying SVs, reducing the false positive rate (Lin et al., 2015). Repetitive and GC rich regions in the genome cause major challenges during sequencing and analysis of SVs. These challenges partly explain the wide variation in the number of observed SVs per genome and drive the need for integrative approaches to accurately identify SVs.

Beside short and long-read sequencing technologies, recently developed next-generation OGM uses very long double-stranded fragments of DNA (~ 300Kbp in length) labelled with fluorescent markers at specific sites. During the generation of optical maps, images of the fluorescent signal patterns are generated (Yuan et al., 2020). This new approach offers greater detection rate of SVs because of the absence of the pre-fragmentation steps required, allowing the analyses of ultra-high molecular weight DNA molecules. Though this emerging technology still needs to prove its value within the clinical genetic diagnostic practice, it has been shown to contribute to the identification of translocation and inversion breakpoints, large insertions and deletions in the genome, and more complex SV (Yuan et al., 2020; Ho et al., 2020).

In this chapter, I optimised SV detection by analysing the genome of 9 infertile patients with both short read WGS as well as OGM. To further aid optimisation two *in-silico* WGS datasets were created, which included gains and losses generated using simuG (Yue & Liti, 2019) and ART (Huang et al., 2012) tools. SV analysis in *in-silico* WGS data was performed using LUMPY (Layer et al., 2014), DELLY (Rausch et al., 2012), Manta (Chen et al., 2016), dysgu-sv (Cleal & Baird, 2021) and GATK-based CNVRobot (https://github.com/AnetaMikulasova/CNVRobot). The 9 samples run on WGS and OGM platforms were analysed using Bionano Access software, CNVRobot and dysgu-sv.

3.2 Aims

This chapter aims to

- Establish a comprehensive pipeline using integrational approaches to reliably identify SVs in short read WGS data
- Understand the importance of optical mapping in SV analysis and set a standard from which further study herein can be carried out

3.3 Results

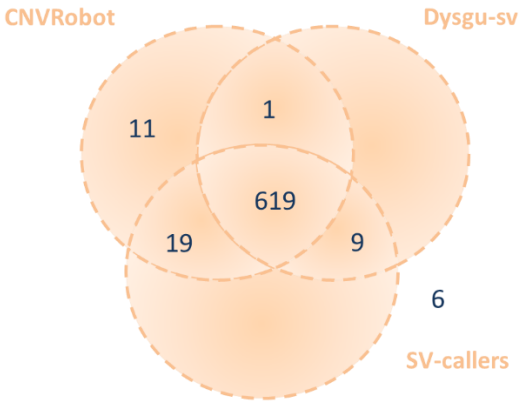
3.3.1 In silico Validation Study of CNV Calling in Genome Sequencing Data

To interrogate the strengths and weaknesses of CNV caller tools to be used in developing a combined approach, we simulated 2 separate WGS datasets for deletions and duplications at around 30X and a set of CNVs (665 deletions and 176 duplications) having different types and sizes in it. Furthermore, the simulation was limited to deletions and duplications because not all of the tools included in this analysis were designed to detect balanced structural variants, such as translocations and inversions. While creating CNVs, just autosomal chromosomes were chosen and 2kb has been taken as a minimum threshold of size, to compare tools accurately. The simulated CNVs were generated with perfect breakpoints to test the core detection capabilities of each tool under ideal conditions, which was crucial for an accurate method comparison. All simulated CNVs were generated at a consistent depth to eliminate sequencing coverage as a confounding variable. Furthermore, the CNVs were not limited to coding regions nor were they evenly distributed across the chromosomes. Instead, their genomic locations were based on a filtered list of pathogenic CNVs from the ClinVar database, which provides a realistic distribution. The simulation was designed to be comprehensive and to avoid a bias toward easy-to-detect variants, as a small number of deletions were intentionally placed in challenging regions, such as telomeres and centromeres, to test the tools' performance in these difficult-to-analyse areas.

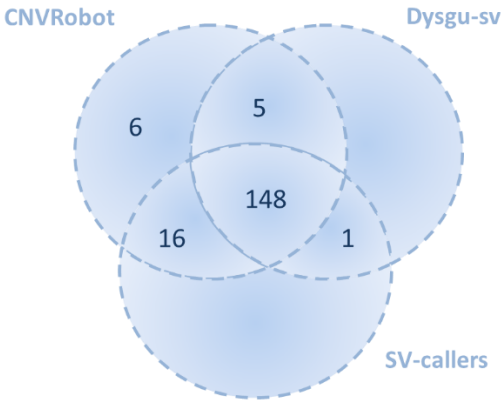
We selected highly cited and widely used LUMPY (Layer et al., 2014), DELLY (Rausch et al., 2012), Manta (Chen et al., 2016), dysgu-sv (Cleal & Baird, 2021) and CNVRobot (https://github.com/AnetaMikulasova/CNVRobot) integrated GATK4 package's read counts normalisation with custom R-based segmentation and visualisation. SV-callers tool was used to efficiently combine Lumpy, DELLY and Manta tools (Kuzniar et al., 2020). Additionally, these tools have already been compared in many studies, and we just aimed to measure how

effective integration of these state-of-the-art tools was compared to others used in this study. Once the data processed, CNVs annotated as being in low mappability region and “high noise in controls” were filtered out in CNVRobot. For dysgu-sv, SVs were included if they were both flagged with PASS in the quality column and had more than 0.5 probability, which is calculated by the tool itself corresponding to the probability of being true positive. Default filtration parameters were applied in the SV-callers. While assessing CNV overlap, a threshold of more than 50% reciprocal overlap was used.

Of the simulated CNVs, 619 were identified by at least one tool. The remaining six deletions were not detected, as they were intentionally created in the telomeric and centromeric regions, which are notoriously difficult for variant calling. This may be due to repetitive sequences leading to low mappability and their partial representation in the reference genome. 619 out of the total of 665 simulated deletions (93%) were ascertained by every tool, 29 (4 %) were detected by two tools and 11 (2 %) by just CNVRobot. CNVRobot and SV-callers tool (combination of Lumpy, DELLY and Manta) have the highest sensitivity (97 %) and lowest false discovery rate (FDR) (2%) in deletion identification. 148 out of simulated 176 duplications (84 %) were ascertained by every tool, 22 (13 %) by two tools and 6 (3 %) by just CNVRobot. CNVRobot has the highest sensitivity (99 %) and lowest FDR (2 %) in duplication identification. Whereas CNVRobot has the highest sensitivity and lowest FDR for copy number detection, it cannot be used to assess precise DNA breakpoint positions as it is based on a read depth strategy (Figure 3.2.). We also measured the processing time for each tool per 2 samples and the fastest tool was dysgu-sv with 1 hour, it took approximately 2 hours with CNVRobot and 9 hours with SV-callers (Table 3.1.). The same infrastructure was used for each tool (see chapter 2.3.3).



Total No. of Deletions = 665



Total No. of Duplications = 176

Deletion	CNVRobot	dysgu-sv	SV-callers
No. of Detected SV	650	629	647
Sensitivity	97 %	94 %	97 %
FDR	2 %	6 %	2 %
Breakpoint Detection Sensitivity	95 %	99 %	99 %

Duplication	CNVRobot	dysgu-sv	SV-callers
No. of Detected SV	175	154	165
Sensitivity	99 %	87 %	93 %
FDR	2 %	8 %	2 %
Breakpoint Detection Sensitivity	98 %	99 %	99 %

Figure 3.2. **Comparison of CNV analysis tools using *in silico* WGS data**. CNVRobot, dysgu-sv and SV-callers (combination of Lumpy, DELLY and Manta tools) were compared in terms of sensitivity, FDR (False discovery Rate) and breakpoint detection sensitivity[2].

Table 3.1. Analysis duration time for 2 samples across tools. CNVRobot, dysgu-sv and SV-callers (combination of Lumpy, DELLY and Manta tools) were run on the same server for created 2 *in silico* WGS data.

Tool	CNVRobot	dysgu-sv	SV-callers
Time (for two sample)	~2 h	1 h	9 h

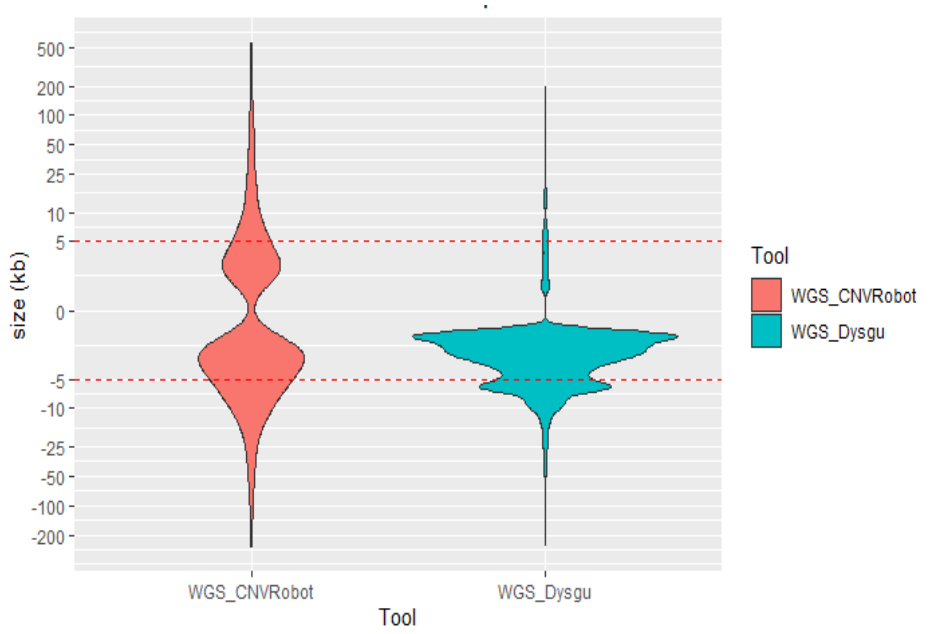
3.3.2 SV Calling in Real Genome Sequencing Datasets

We performed short read WGS and OGM on 9 samples to investigate the effectiveness of our approach in real data. All samples passed the quality controls. Short read Whole Genomw Sequencing (srWGS) data were processed by CNVRobot and dysgu-sv and the same filtration strategy was applied. The number and range of detected calls varied a lot for each tool which we did not observe during the *in-silico* analysis (Table 3.3.) (Figure 3.3. A). This could be because real data contains noise and artifacts, such as base-calling errors, PCR amplification biases, and signal degradation, introduced by the limitations of sequencing technologies. A

[2] Breakpoint detection sensitivity was quantified as the average percentage of overlap between detected and actual CNVs.

total of 1,620 CNVs were identified by CNVRobot across 9 samples, with an average of 180 CNVs per patient, while 6,712 CNVs were detected by dysgu-sv, averaging 745 CNVs per patient. Additionally, the dysgu-sv tool, which uses split and discordant reads (paired reads with unexpected distance or orientation), identified 823 inversions and translocations. Most of the CNVRobot calls were concentrated in the range of 1 kb to 5 kb, whereas dysgu-sv calls were predominantly stacked in the range of 1 kb to 2.5 kb (Figure 3.3.A). The calls showed limited overlap between tools, with the majority of overlapping calls being deletions within the 0-10kb range. We presume that these overlapping calls are the most reliable. (Figure 3.3. B).

A



B

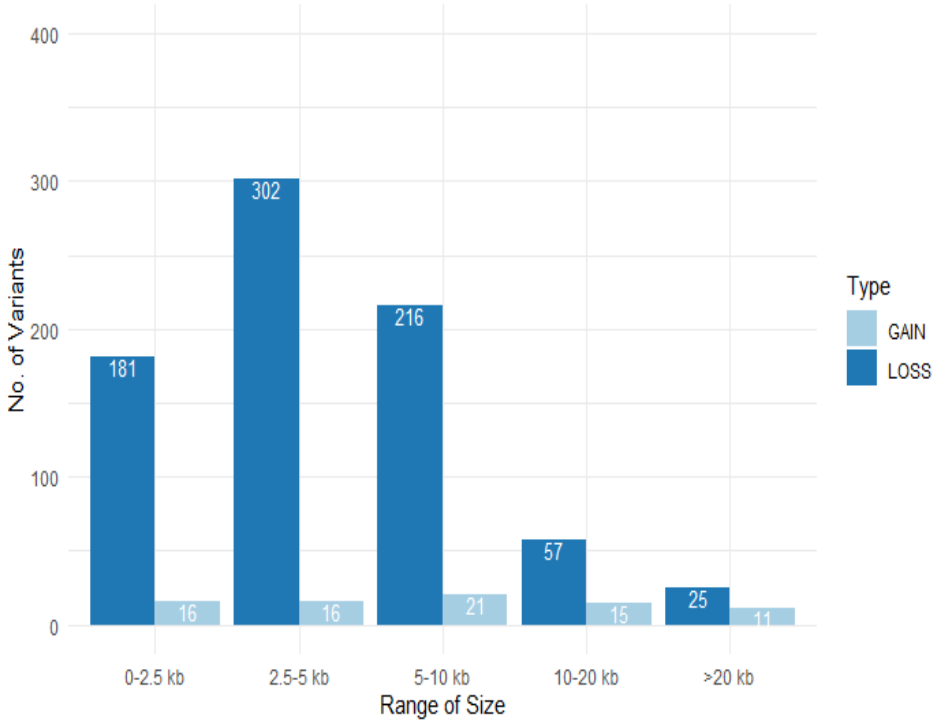


Figure 3.3. Overview of SV detection results across the nine samples. A. Size distribution of CNVs identified with CNVRobot and dysgu-sv tools in 9 samples **B.** No. of overlapped CNVs detected by CNVRobot and dysgu-sv in different ranges

3.3.3 SV Calling in Optical Genome Mapping

Around 11 thousand CNVs were detected by OGM technology without applying any filters in 9 samples (Table 3.2.). Both the Bionano SV and CNV pipelines were used for analysis (see chapter 2.2.3). While the SV pipeline relies on analysing technology-specific signatures to detect SVs, the CNV pipeline, which is based on depth of coverage, was prone to high false-positive rates due to technological limitations.

Table 3.2. No. of total variants detected by OGM in 9 samples before filtration, using Bionano SV and CNV pipelines

	Bionano SV Pipeline	Bionano CNV Pipeline
Deletions	3,941	254
Duplications	800	172
Insertions	5,637	
Inversions	306	
Total	10,684	456

Most of the SVs were larger than 5kb (Figure 3.4.A). When examining individual SV types by size, the observed ranges aligned with platform expectations: insertions ranged from 5kb to 50kb, deletions were larger than 5kb, and duplications larger than 100kb. (Figure 3.4.A). After filtration based on the manufacturer`s recommendations and visual inspection, all CNVs detected by Bionano CNV Pipeline were filtered out due to probably usage of read depth signature in the pipeline which is not effective for this technology. Calls detected by Bionano SV Pipeline were also filtered out based on recommended filters. Interestingly, all duplications and most of the inversions were below quality thresholds and filtered out. It might be that the efficiency of the platform is variable with different SV types. 132 SVs were detected per patient on average by Bionano SV Pipeline varying between 85 and 155 (Table 3.3.).

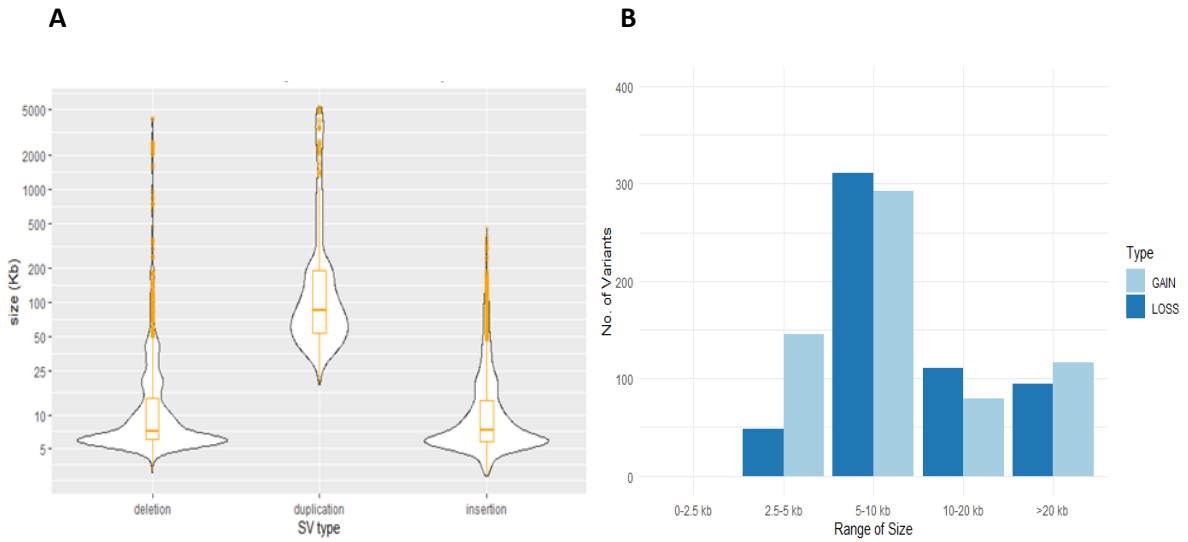


Figure 3.4. Overview of SVs identified by OGM in 9 samples A. The size distribution of different type of SVs detected by OGM in 9 samples before filtration **B.** The no. of variants after filtration in different size ranges

Table 3.3. No. of copy number losses and gains per patient identified by WGS tools (CNVRobot and dysgu-sv) and OGM.

■: LOSSES ■: GAINS

	CNVRobot	dysgu-sv	OGM	CNVRobot	dysgu-sv	OGM
0001P	105	706	70	52	26	36
0071P	101	670	52	45	18	130
0088P	97	707	50	58	29	61
0090P	127	753	63	49	28	76
0103P	286	877	155	78	34	100
0105P	74	715	35	65	26	42
0106P	96	659	54	74	30	61
0108P	88	654	40	67	31	49
0112P	85	684	45	85	35	77

3.3.4 Comparison of SV calling in WGS and OGM

I observed clear differences in the size distributions of CNVs detected on both platforms. While CNVs detected by srWGS were predominantly below 10kb, OGM identified a broader size distribution with most CNVs above 10kb (Figure 3.5, Table 3.4). This difference is unlikely to be an artifact of segmental calling (i.e., detecting the same large CNV in multiple smaller segments), as the CNVRobot tool used in our srWGS pipeline is designed to handle this by merging adjacent segments.

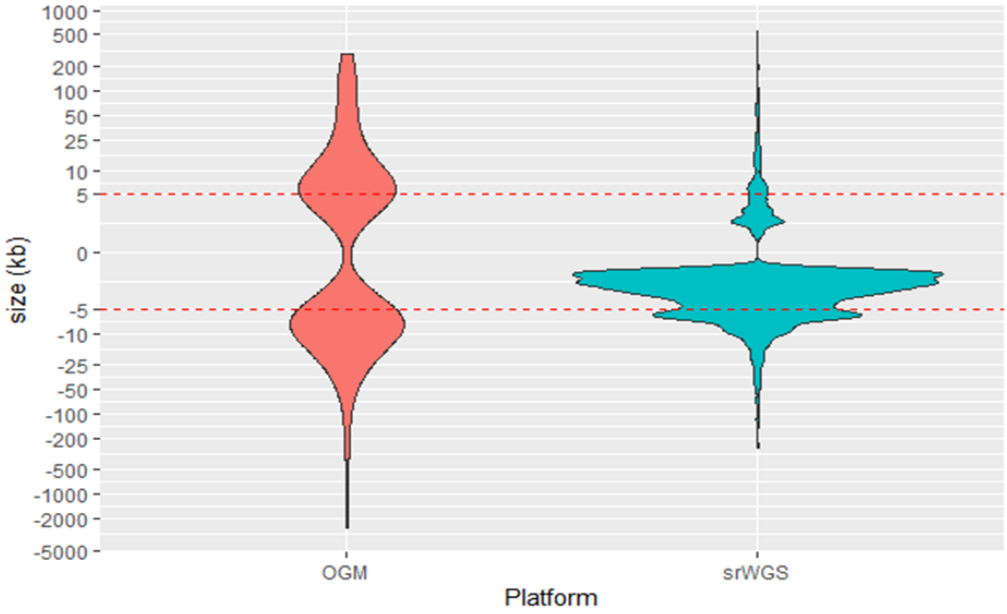


Figure 3.5. Size distribution of CNVs identified with OGM and srWGS

As one of my main aims was to optimise SV detection in srWGS, I considered 446 CNVs overlapping with OGM as the best representation of a gold standard and investigated the presence of these CNVs in the short-read WGS data using two different CNV tools (Figure 3.6). Under 5kb (N=60), nearly half of the calls that have been taken as a truth set were ascertained by just dysgu-sv. Above 5kb (N=386), particularly calls more than 20kb, CNVRobot has the highest percentage of unique detection of calls. Notably, the overlap between CNVRobot and dysgu-sv revealed complementary detection capabilities: dysgu-sv excelled at detecting CNVs below 5 kb, whereas CNVRobot outperformed in identifying larger CNVs (>5kb), with many calls above 5 kb detected by both tools. This overlap underscores the potential for integrating these methods to enhance overall sensitivity, especially given CNVRobot's alignment with OGM's proficiency for larger SVs. These findings suggest that CNVRobot is the more robust

tool for detecting larger CNVs in srWGS, while a combined approach leveraging both tools could maximise detection across the different sizes.

Table 3.4. No. of variants detected by WGS and OGM within different size ranges, including their overlaps.

	Deletions			Duplications		
	WGS	**OGM**	**Overlapped**	**WGS**	**OGM**	**Overlapped**
1-2.5 kb	3485	0	**0**	291	0	**0**
2.5-5 kb	1643	48	**48**	226	145	**12**
5-10 kb	1232	311	**210**	110	292	**31**
10-20 kb	198	111	**76**	51	79	**19**
>20 kb	145	94	**37**	73	116	**13**

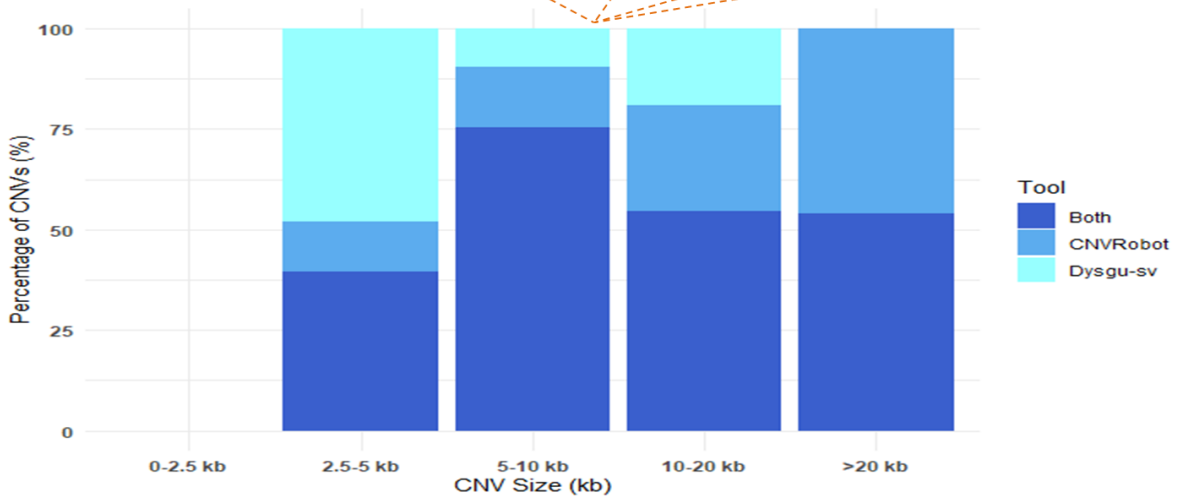


Figure 3.6. The proportion of overlapping CNVs between WGS and OGM detected by CNVRobot, Dysgu-sv or both.

3.4 Discussion

In this chapter, we paved the way for detecting SVs in srWGS data and investigated the effect of recently developed OGM technology on SV detection. To do so, I combined different bioinformatic tools and platforms into an integrated workflow and applied this to both simulated and real genomic data.

Whereas srWGS is becoming more available and is superseding srWES in clinics, it still remains a challenge to use it for the detection of all variation types, particularly SVs, which have been

shown to remarkably increase diagnostic yield (Shieh et al., 2021). So, I aimed to find the most efficient way of performing SV detection in srWGS data by combining tools using different SV signatures as suggested in the literature (van Belzen et al., 2021). I firstly used simulated *in silico* srWGS data including a large and diverse set of deletions and duplications. A number of recently published, highly-cited and commonly used SV detection tools were applied to this data. CNVRobot, which uses read depth for SV detection, was found the most sensitive and specific tool for the detection of deletions and duplications. Breakpoint accuracy was less good for this tool, as expected since it does not incorporate split reads. Both dysgu-sv and a combination of Lumpy, DELLY and MANTA were able to detect SV breakpoints at single base resolution. Additionally, split read and read pair signatures enable us to ascertain balanced SVs (inversions and translocations) even though sensitivity of most methods are not as high as for CNVs.

In addition to studying SV detection accuracy, I also measured the run time for each tool as this is critical in using these tools on large srWGS datasets. Dysgu-sv was the fastest with 1 hour for 2 samples, in contrast to sv-callers which was the slowest, taking 9 times longer using the same compute infrastructure. Taken together, I decided to combine CNVRobot and dysgu-sv and exclude SV-callers since this combination enables us to detect all SV types with high breakpoint detection resolution and sensitivity in a reasonable time.

It is important to note that while our simulation assumed that SVs were present in all cells (germline), the real samples may contain variants present in only a fraction of cells (somatic mosaicism). This is a significant consideration, as mosaicism can lead to lower read support and more complex signatures that are challenging for standard tools to detect. However, as our study is mainly focused on reliably detecting germline SVs, this simulation model is appropriate for that specific purpose. To further optimise SV detection in real samples and also evaluate OGM technology for SV detection, 9 samples from our male infertility cohort were sequenced with srWGS and analysed by OGM. We processed the WGS data using CNVRobot and dysgu-sv, detecting approximately 924 SVs per genome. This number is roughly 11-fold lower than the average of 11,844 SVs per genome reported in large-scale SV studies (Collins et al., 2020). It is probably because we filtered out SVs less than 1kb whereas their threshold is 50 bp (the median SV size is 306 bp). Unlike in the *in-silico* analysis, SV ranges and numbers detected by the two tools were very different probably due to noise in the data, and signatures tools used. As it might not be an efficient way of confirming thousands of variants

by using gold standard methods such as qPCR, we used OGM calls overlapped with WGS as a gold standard. Performing OGM also allowed us to investigate how effective this recently developed technology is for SV detection.

Mantere et al., 2021, concluded that OGM is better at identifying large SVs (>10kb) with high specificity due to its ability to analyse long DNA molecules, though it may miss smaller variants, resulting in lower sensitivity compared to srWGS. In contrast, srWGS provides higher sensitivity for detecting a broader range of CNVs, including smaller events, but struggles with large SVs in repetitive regions (Pei et al., 2024). So, these technologies are not alternatives to each other, they may be complementing each other. Additionally, the existing OGM data analysis tools still are not sufficient to interpret the data, and we were limited to options in Bionano Access software. Given OGM's reported low false-positive rate (Dremsek et al., 2021), we used CNVs overlapping between OGM and srWGS as a 'gold standard' to evaluate the performance of two srWGS SV callers, CNVRobot and dysgu-sv. Our analysis revealed that combining these tools enhanced detection sensitivity for these presumed true CNVs. However, without a detailed true positive/false positive assessment against OGM, we cannot conclusively evaluate specificity or false discovery rate.

Incorporating long-read sequencing could further enhance SV identification by resolving complex and repetitive regions beyond the capabilities of srWGS, providing improved breakpoint precision and detection of larger insertions (Ebert et al., 2021; Chaisson et al., 2019). Similarly, integrating emerging tools could enhance detection sensitivity and specificity for SV identification.

3.5 Conclusion

I concluded that read depth-based approaches are the most sensitive and specific way of detecting copy number variation and split-read-based approaches enable us to precisely detect SV breakpoints. By combining tools that implement different signatures such as read depth, split reads and read pairs we can further increase the accuracy of SV detection. Also, the computing requirements and processing time should be taken into account when choosing a combination of tools for SV detection. Additionally, the OGM technology may be better at identifying larger SVs, although we did not independently validate any of the SVs identified. For now, SV detection on srWGS data using a combination of CNVRobot and dysgu-sv seems to give most reliable data and will be used further in this thesis.

Chapter 4. Overview of SVs in Idiopathic NOA and Severe Oligozoospermia Cohort

4.1 Introduction

SVs, defined as genomic alterations larger than 50 base pairs, encompass a diverse range of DNA rearrangements such as deletions, duplications, insertions, inversions, translocations, and complex rearrangements (Weischenfeldt et al., 2013). These variants play a crucial role in human genetic diversity and have significant implications for genome function and disease susceptibility (Feuk et al., 2006). SVs, because of their size, have a disproportionate impact on human health relative to their frequency in the genome. They can disrupt gene function, alter gene dosage, or affect regulatory elements, leading to various phenotypic consequences (Conrad et al., 2010). For instance, rare and *de novo* SVs have been implicated in neurodevelopmental disorders like autism spectrum disorder and schizophrenia (Pinto et al., 2010).

Evaluating the relationship between SVs and diseases involves comprehensive genomic analyses. However, detecting SVs poses significant challenges due to their size, complexity, and the mechanisms by which they arise (Alkan et al., 2011). Traditional detection methods like karyotyping and microarray analyses offer limited resolution and may miss smaller or more complex SVs (Pinto et al., 2011). The detection limit of these technologies restricts our ability to fully understand the contribution of SVs to various conditions, including male infertility.

Advancements in sequencing technologies have revolutionised our ability to detect and analyse SVs. WGS, particularly with high coverage, allows for a more comprehensive and detailed view of the genome (Chaisson et al., 2019). It enables the identification of SVs that were previously undetectable with conventional methods. Moreover, the development of sophisticated computational tools and scalable analytical pipelines has facilitated the processing of large genomic datasets, enhancing our capacity to discover and interpret SVs on a population scale (Abel & Duncavage, 2013).

Current technological progress has contributed significantly to the creation of population SV catalogues. Large-scale initiatives like the Genome Aggregation Database (gnomAD) have aggregated genomic data from tens of thousands of individuals (Collins et al., 2020). These resources provide insights into the prevalence and distribution of SVs across different

populations. These resources serve as reference databases that aid in the classification of variants, helping researchers differentiate between likely pathogenic variants and benign polymorphisms. (Karczewski et al., 2020).

Here, in this project I investigate the role of SVs in the specific form of male infertility by using current technologies and databases. In this chapter, I present an overview of our analysis of structural variants in 216 patient-parent trios, laying the ground for detailed examinations in the upcoming chapters.

4.2 Aims

This chapter aims to provide an overview of SVs identified in WGS data of 216 patient-parent trios.

4.3 Results

WGS data from 216 patients affected by azoospermia or severe oligozoospermia and their parents were analysed with the optimised pipeline to identify SVs (see chapter 3). The two different tools, CNVRobot and dysgu-sv, were combined (see chapter 2.3.3) which increased the sensitivity and specificity for calling SVs. Detailed clinical information was collected for all families in our cohort. While comprehensive multi-generational pedigrees were not available for all cases, the collected data includes key details that suggest a heritable component. For example, 16 probands showed a family history, with 13 having affected male siblings and 4 having evidence of male factor infertility in maternal uncles. This clinical information provides crucial support for our analysis.

To ensure the integrity of our data, we performed rigorous quality control to confirm the relationships within each trio. The genomic data from each sample was examined for sex, ancestry, and relatedness using peddy (Pedersen & Quinlan, 2017) by Dr. Miguel Xavier. This process was critical for confirming paternity for all trios and for identifying potential parental consanguinity. If any discrepancies were identified, such as incorrect sex or a lack of relatedness, we investigated for sample swaps. All samples that could not be resolved were excluded from the study. Two families were flagged for a high possibility of consanguinity, which was accounted for in our downstream analyses.

4.3.1 Overview of SVs in Probands

The SV calling from the WGS probands data identified a total of 237,213 SVs (28,218 by both tools, 89,292 by CNVRobot, and 119,703 by dysgu-sv) with four outlier patients based on Rosner`s test analysis (Figure 4.1. A), with an average of 1,099 SVs identified per sample. While SVs identified by both CNVRobot and dysgu-sv can be considered the high-confidence set, our approach was to retain all SVs and carefully interpret them in later chapters by continuously reviewing each candidate using IGV and CNVRobot plots. This allowed us to avoid overly restrictive filtering and ensured a comprehensive assessment of potential SVs.

The outliers were examined further to assess their validity, and an increased number of deletions were found throughout all chromosomes. There was not any enrichment on any chromosome or region. All four outliers presented with greater noise, demonstrated by Figure 4.2. despite the quality metrics (median of read depth and Derivative Log Ratio Spread (DLRS)[3]) of the data shown to be within an acceptable range (>30x coverage and <0.3 DLRS). Fluctuation in read depth is one possible explanation, resulting from poor sample quality or sequencing. It is notable that with the increased data noise, there is an increase in false positive calls for the smaller sized SVs. A focused examination of the SVs based on size distributions highlighted that SVs smaller than 5kb were observed twofold more frequently among the outliers, with SVs ≥5kb found to be comparable to the cohort average (Figure 4.1. B). Therefore, all SVs smaller than 5kb from the 4 outlier samples were excluded from further analysis and caution was applied while interpretation of ≥5kb SVs for these cases, given the increased probability of false positive calls.

[3] DLRS refers the standard deviation (SD) of the disparities among Log R Ratio (LRR) values for probes arranged based on genomic position, which is then divided by the square root of two (Dennis et al., 2021).

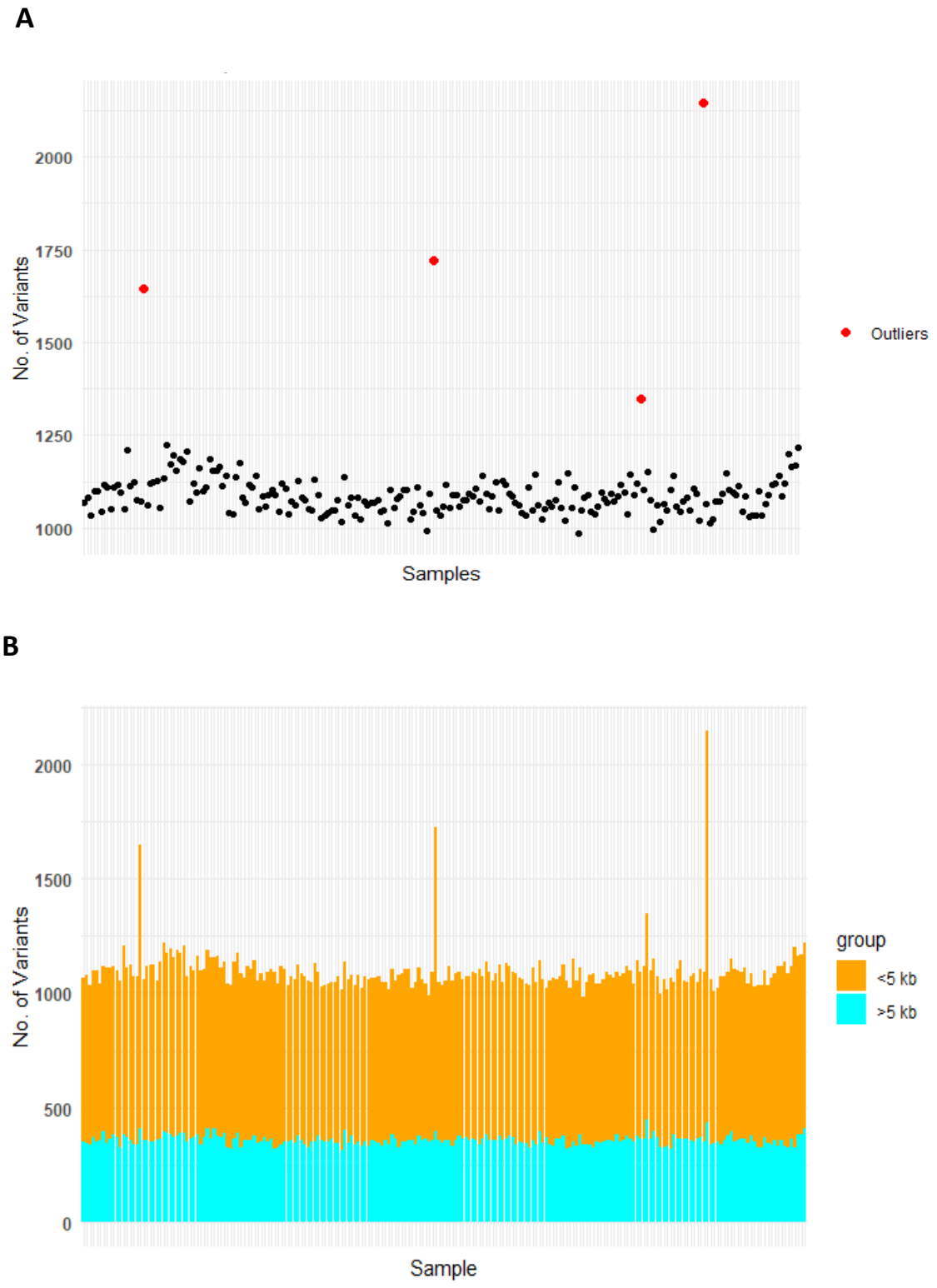


Figure 4.1. Total number of SVs per sample detected in 216 probands with azoospermia or severe oligozoospermia. A. Each dot represents one patient. Red dots indicate the four outlier patients. **B.** Total number of SVs per sample grouped and coloured by size, where orange represents SVs ≤5kb and turquoise highlights SVs ≥ 5kb.

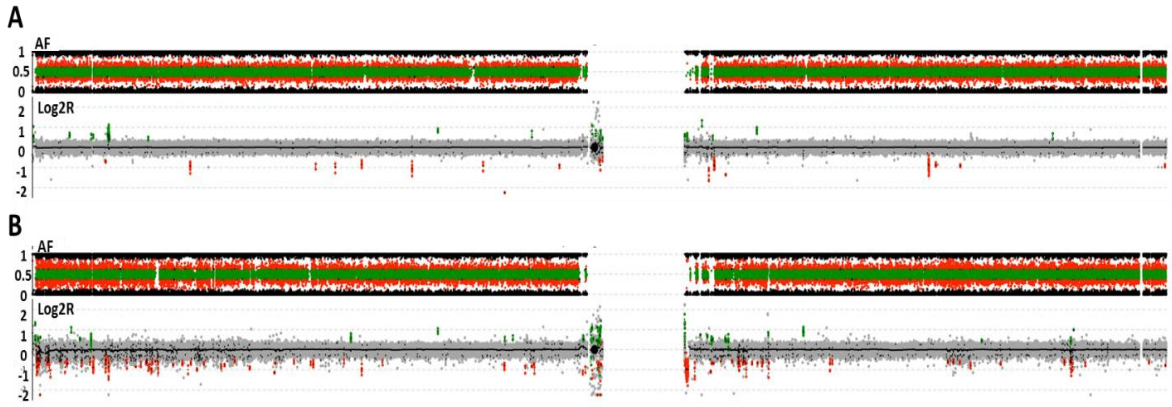


Figure 4.2. Examples of chromosome 1 plots from samples NIJ_MI_02080P (**A**, a non-outlier sample) and NIJ_MI_02247P (**B**, an outlier sample) generated by CNVRobot. The presence of noise in copy number ratios and SNP zygosity is more evident in B when compared to A.

Next, we studied the distribution of SVs over the autosomal chromosomes. As expected, the longer chromosomes contained most SVs (Figure 4.3. A). After normalisation for chromosome length by calculating the number of SVs per Mb using the effective size[4] of the chromosomes, SVs appear to be evenly distributed among most chromosomes, except for chromosomes 19 to 22 (Figure 4.3. B). A potential explanation for this higher rate is that these chromosomes are known to be particularly gene-dense, which could lead to increased rates of transcription-driven SV formation. In line with this hypothesis, our analysis confirmed a significant positive correlation between the absolute number of genes and the total number of SVs across the autosomes ($r = 0.65$, $p = 0.001$). However, to test the link to gene density more directly, we performed a final normalisation of the SV count by the number of genes per Mb (Figure 4.3. C). This analysis revealed a more complex pattern, where the gene-dense chromosomes 19 and 22 no longer stood out. Instead, chromosomes such as 2, 4, and 13 showed a disproportionately high SV burden relative to their gene content, indicating that factors beyond simple gene density are also at play.

[4] The effective genome size was determined by calculating denoised intervals utilising control data. Denoised intervals refer to refined genomic regions obtained through the removal of noise.

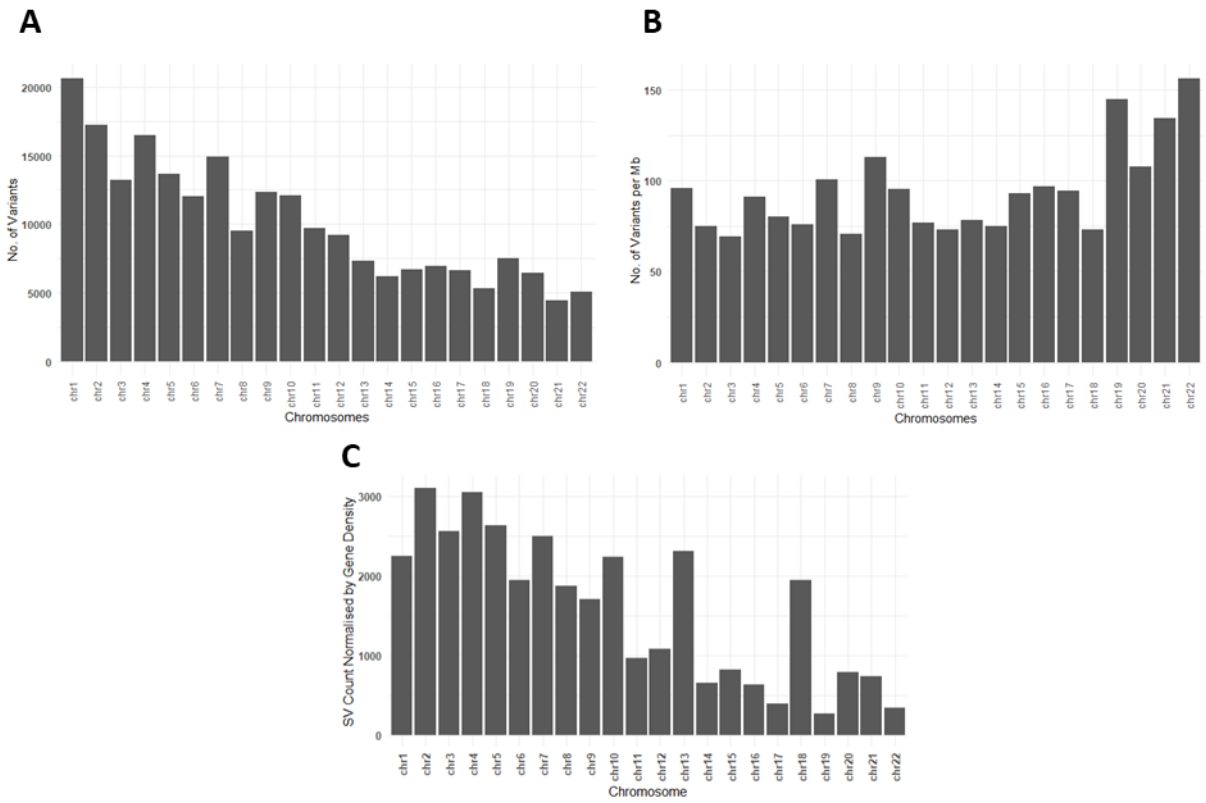


Figure 4.3. Distribution of all SVs identified in 216 probands with azoospermia or severe oligozoospermia by chromosomes. A. Distribution of the total number of SVs identified by autosomal chromosome. Distribution is as expected based on chromosome length. **B.** Number of SVs identified per Mb by autosomal chromosome (calculation was done by using the length of the effective size of the chromosomes). **C.** SV count normalised by the number of genes per Mb.

Additionally, we observed that CNVs were dispersed differently along the autosomal chromosomes. Both the subtelomeric and pericentromeric regions had larger proportions of CNVs, however their confidence assignments were lower than anticipated because the sequences in these regions are frequently complicated and challenging to assay accurately (Figure 4.4.). Figure 4.4 illustrates a single chromosome as an example, with chromosome 15 chosen specifically due to its sub-centromeric region. This region contains low-copy repeats that are known to be hotspots for structural rearrangements, a feature directly relevant to genomic disorders such as Prader-Willi/Angelman Syndrome.

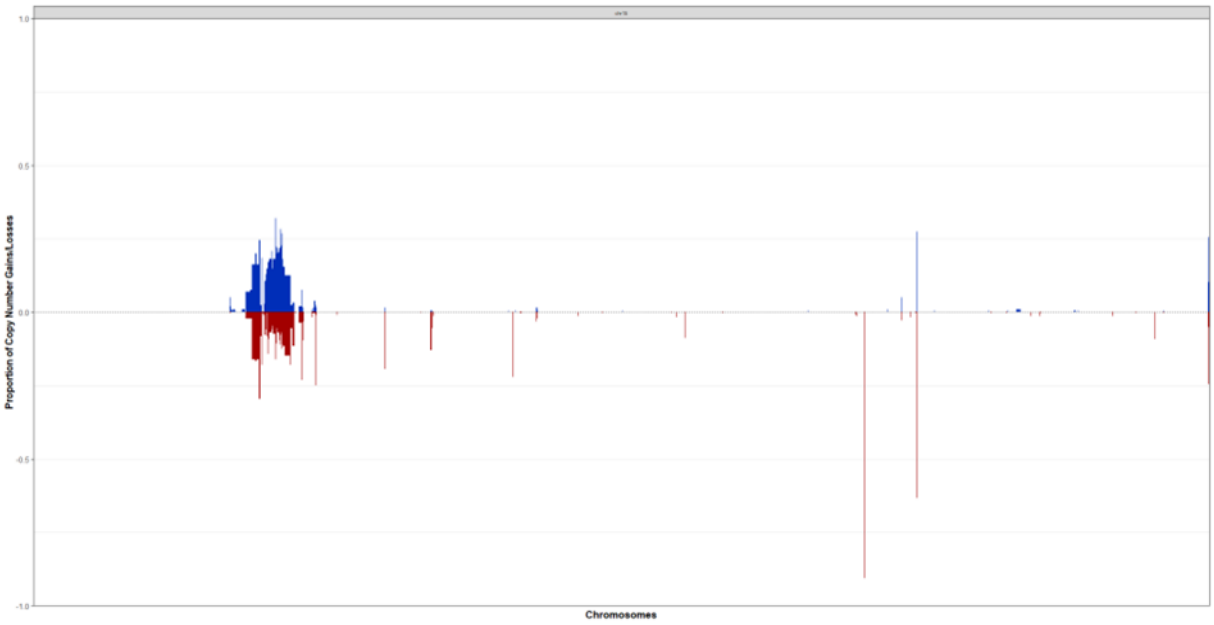


Figure 4.4. **Proportion of SVs identified along the chromosome 15**. It can be clearly seen that the subtelomeric region had larger proportions of SVs. Red colour represents deletions and blue duplications.

We also looked at the size distribution of SVs across all chromosomes. The majority of SVs were small in size, with most measuring less than 10kb. Nearly half of these were even smaller, measuring less than 2.5kb (Figure 4.5. A). Approximately 176,000 deletions were identified, which is four times more than the number of duplications found. Among the identified SVs, 161,348 (70%) were common (>1% in population databases), while 70,024 (30%) were rare (<1% in population databases). (Figure 4.5. B). To refine the analysis, we mostly focused on rare SVs that were present in less than 1% of the samples from the population databases for further investigation. When examining the genomic landscape of these rare SVs, it becomes apparent that the vast majority are situated within intergenic regions (Figure 4.5. C).

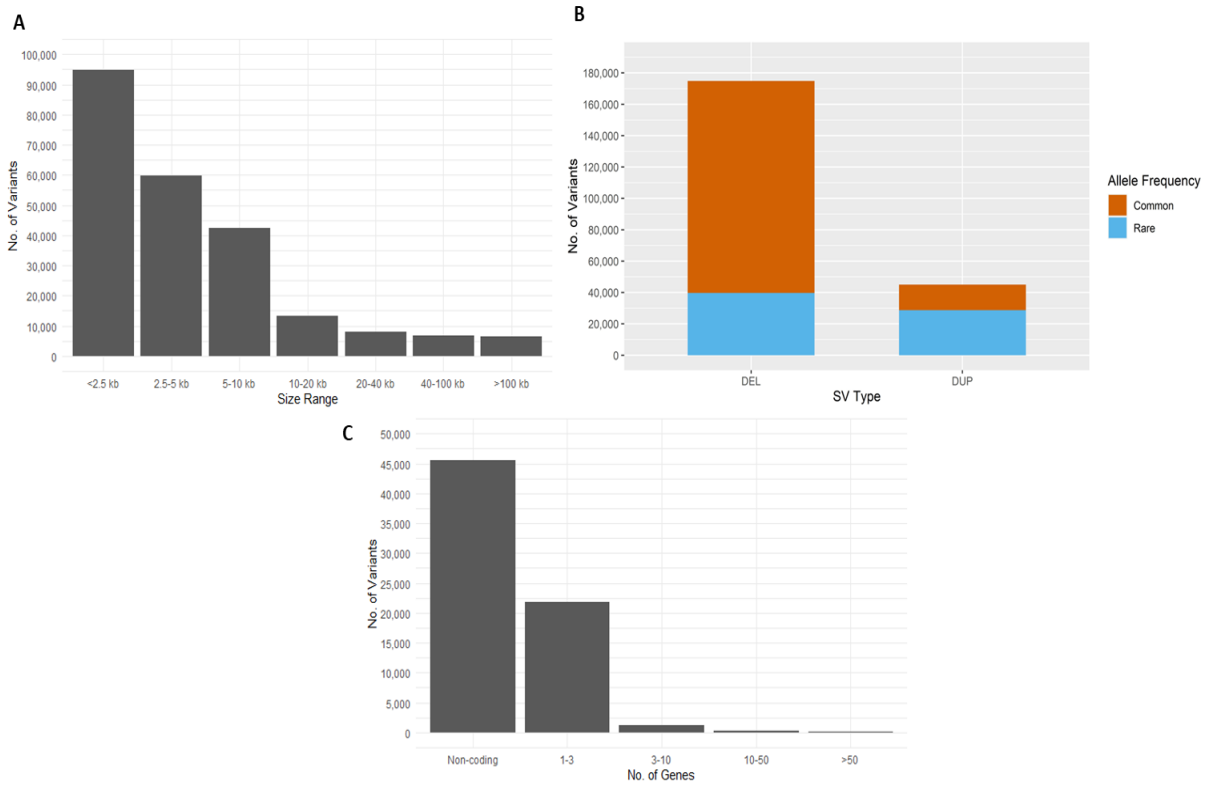


Figure 4.5. Overview of SVs identified in 216 probands. A. Size distribution of SVs identified in 216 probands with azoospermia or severe oligozoospermia. **B.** Number of SVs identified in 216 probands based on population allele frequency (rare: present in <1% of the samples of the population databases). The common SVs outnumbered the rare SVs as expected. **C**. The number of genes overlapped with rare SVs was identified in 216 probands.

4.3.1.1 SVs on The Sex Chromosomes

To examine SVs on the sex chromosomes, a new version of CNVRobot, which includes a significant improvement in CNV calling on sex chromosomes, has been applied, in combination with dysgu-sv (see chapter 2.3.3). A total of 6,437 SVs were identified on these chromosomes (152 by both tools, 2,240 by CNVRobot, and 4,045 by dysgu-sv), with an average of 30 SVs per patient. The very low (2.4%) numerical overlap is not unexpected and highlights the tools' different methodologies, as well as the inherent challenge of SV detection, particularly on gonosomes. SVs detected only by dysgu-sv were substantially smaller (mean 3 kb) than those detected only by CNVRobot (mean 13 kb). This confirms that most SVs missed by CNVRobot were very small and, therefore, represented a small proportion of the chromosomes despite dominating the SV count.

Based on Rosner's test analysis, there were 5 outliers, 4 with low numbers (around 8) and 1 with a high number (85) compared to the average number. The proband with the high number of SVs was analysed cautiously. Of the identified SVs, 625 (10%) were rare, 519 on the X

chromosome (38 by both tools, 400 by CNVRobot, and 81 by dysgu-sv) and 106 on the Y chromosome (3 by both tools, 91 by CNVRobot, and 12 by dysgu-sv).

519 rare SVs, ranging from 1kb to 1.6Mb, on the X chromosome were evenly distributed across the chromosome, except for the regions at the beginning and end. (Figure 4.6.). Since SVs on the X chromosome are inherited from the mother, a detailed analyses of these are provided in Chapter 6.

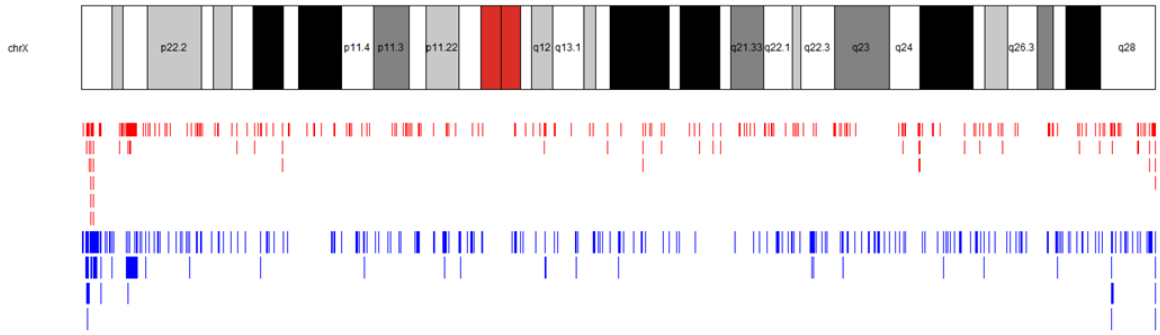


Figure 4.6. Distribution of 519 rare CNVs on the X chromosome in 216 probands with azoospermia or severe oligozoospermia (deletions in red and duplications in dark blue).

4.3.1.1.1 Inherited CNVs on Chromosome Y

A total of 106 rare SVs, ranging from 1kb to 2Mb, on chromosome Y were identified across the entire cohort (3 by both tools, 91 by CNVRobot and 12 by dysgu-sv). The larger SVs were inspected for overlap with known AZF deletions, and none were found, as expected, since screening for these deletions was used as an exclusion criterion at the beginning of the study. Among these, 34 SVs that were observed fewer than three times in the cohort across chromosome Y are illustrated in Figure 4.7. Since severe male infertility cannot be inherited from fertile fathers through spontaneous fertilization, paternally inherited SVs on the Y chromosome were excluded from further analysis.

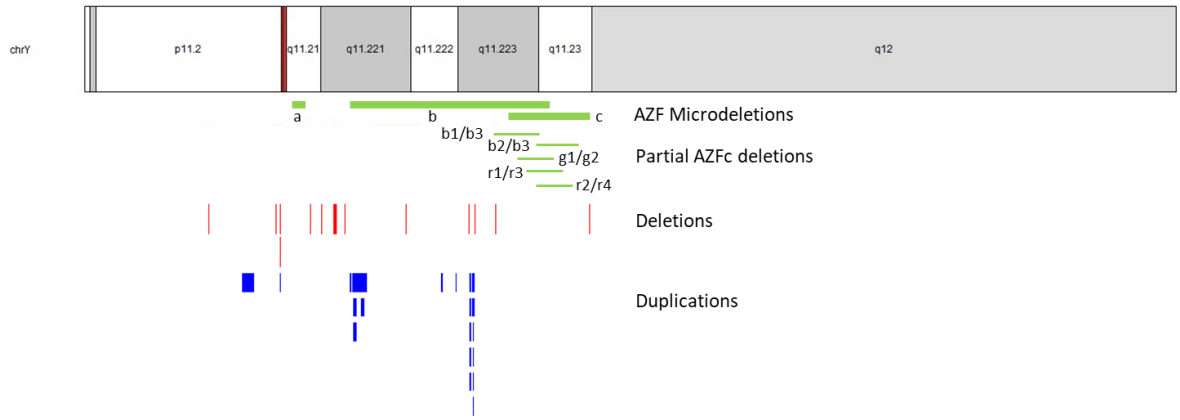


Figure 4.7. Distribution of 34 rare CNVs on the Y chromosome in 216 probands with azoospermia or severe oligozoospermia (deletions in red and duplications in dark blue). AZF microdeletions and partial AZFc deletions are also depicted in green. CNVs detected in our cohort did not include known Y chromosome microdeletions causing male infertility.

It is acknowledged that this stringent filtering approach has known caveats. For instance, it would fail to identify a pathogenic variant transmitted from a sub-fertile father with oligozoospermia who was able to conceive naturally. Furthermore, this filter cannot account for paternal germline mosaicism, where a father could pass on a mutation present in his germline while appearing unaffected in somatic tests. Consequently, it is possible that some true causative variants were inadvertently excluded by this analytical step.

4.3.1.2 Balanced Structural Variations

Balanced SVs were identified by the dysgu-sv tool that uses split reads and discordant read pairs enabling detection of inversion and translocations. Overall, 8485 inversions and 7839 translocations were ascertained in the 216 probands. Of the 8485 inversions, 568 were rare (6.6 %) and 143 of those were inherited solely from the father which were excluded from further investigation. After filtering out inversions observed in more than three samples or any father within the entire control group, the remaining count of inversions that included genes and underwent visual inspection was 11. Of those 5 were real while one originated from the paternal side and was inaccurately predicted as probably *de novo* by the algorithm (Figure 4.8., Table 4.1.). Since these 4 inversions are inherited from mothers, detailed functional annotations are provided in Chapter 6.

Of the 7,839 translocations detected by dysgu-sv, 4,272 were predicted to be inherited solely from the father which were excluded from further investigation. After filtering out translocations observed in more than three samples or any father within the entire control

group, the remaining count of translocations that included genes and underwent visual inspection was 57. None of those seems to be real.

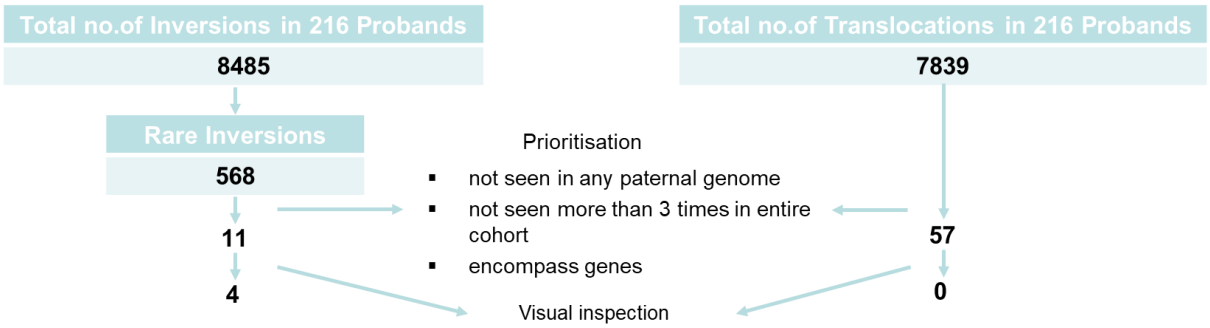


Figure 4.8. Total and breakdown of the number of inversions and translocations identified in 216 probands with azoospermia or severe oligozoospermia.

Table 4.1. The identified inversions in 216 probands with azoospermia or severe oligozoospermia after systematic filtration.

Proband	Genomic Location	Size (bp)	Genes
NIJ_MI_01490P	chr16:10731977-11390943	658966	*CIITA -CLEC16A-DEXI-NUBP1-PRM1-PRM2-PRM3-RMI2-SOCS1-TNP2-TVP23A*
NIJ_MI_00986P	chr16:88976931-88982526	5595	*CBFA2T3* (Involving 5' UTR)
NIJ_MI_01166P	chr10:115483969-115486405	2436	*ATRNL1* / (within intron)
IND_MI _004P	chr6:123290449-123292571	2122	*TRDN* (within intron)

4.3.2 Overview of SVs in Parents

The infertile patients in our research provided a blood sample in the clinics where they were referred, while the parents received a saliva collection kit at home, and srWGS was then performed. Since saliva-derived DNA yields srWGS data with generally lower coverage than blood-derived DNA, we sequenced most of these samples twice to obtain at least 30X coverage and merged their data.

A total of 204,314 SVs were identified in 216 mothers (26,552 detected by both tools, 42,757 uniquely by CNVRobot, and 135,005 uniquely by dysgu-sv), averaging 946 SVs per sample. In contrast, 259,678 SVs were identified in fathers (28,661 by both tools, 100,277 uniquely by CNVRobot, and 130,740 uniquely by dysgu-sv), averaging 1,203 SVs per sample (Figure 4.9.A.). The observed difference in SV counts per individual between fathers (which is similar with probands) and mothers primarily derived from the higher number of CNVRobot calls in fathers. No enrichment on any particular chromosome including gonosomes was observed.

So, this difference was not due to SVs on sex chromosomes. When considering only autosomal SVs, a total of 250,860 SVs were in fathers (28,486 detected by both tools, 96,774 uniquely by CNVRobot, and 125,600 uniquely by dysgu-sv) and 196,978 SVs in mothers (25,583 detected by both tools, 41,255 uniquely by CNVRobot, and 129,140 uniquely by dysgu-sv). Therefore, this disparity may arise because female controls, being fully diploid, enable identification and exclusion of more bias-prone genomic regions, reducing false positive calls. Alternatively, sex-specific differences in normalisation (such as increased noise from hemizygous X and Y chromosome regions in males) could increase CNVRobot's call rate in fathers, as normalisation assumes a diploid baseline that may be less stable across male genomes.

The size distribution of SVs was similar between probands and parents, with the majority being small, predominantly less than 10 kb in length. Regarding the distribution of SVs across autosomal chromosomes, longer chromosomes harboured the higher number of SVs, as expected due to their larger size. After normalising for chromosome length by calculating the number of SVs per Mb based on the effective chromosome size, SVs were found to be evenly distributed across most autosomes. However, chromosomes 19 to 22 exhibited a higher density of SVs in both fathers and mothers (respectively Figure 4.9.B., Figure 4.9.C.), a pattern also observed in probands.

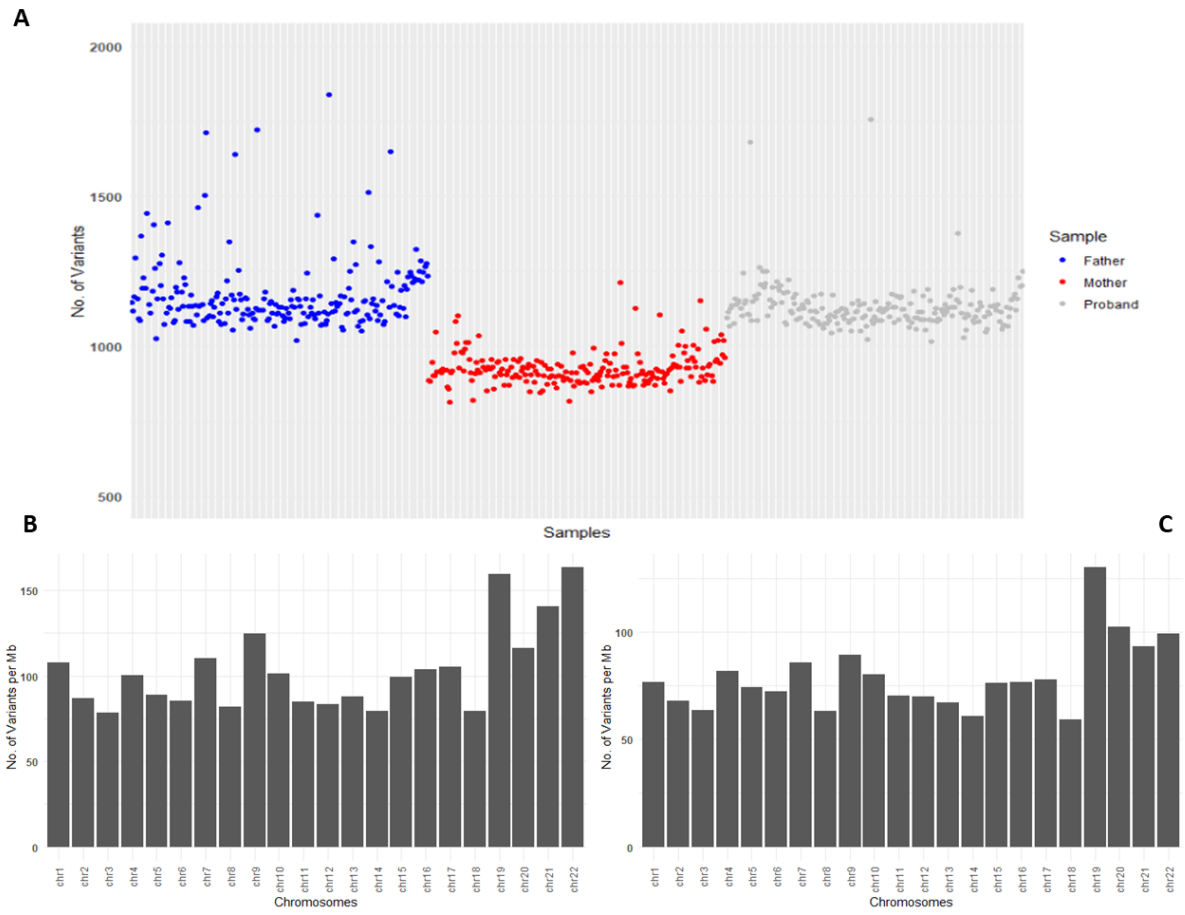


Figure 4.9. Overview of SVs identified in parents. **A.** Total number of SVs per sample identified in parents. Red dots indicate mothers, grey probands and blue fathers. **B.** Number of SVs identified per Mb in fathers by autosomal chromosome (calculation was done by using the length of the effective size of the chromosomes). **C.** Number of SVs identified per Mb in mothers by autosomal chromosome (calculation was done by using the length of the effective size of the chromosomes).

4.3.2.1 Comparing Probands with Parents

To determine if a significant difference exists in the mean number of SVs between mothers, fathers, and probands, a linear mixed-effects model (LMM) was used. This statistical approach was selected to properly control for the non-independent nature of data within family trios. The analysis was performed on the set of SVs on autosomes that were identified by both detection tools.

The LMM, with SV count as the dependent variable, sample type (father, mother and proband) as a fixed effect, and family ID as a random effect, revealed a significant effect of sample type on the number of SVs ($p < 2.2e\text{-}16$). A post-hoc analysis with Tukey's p-value adjustment showed that mothers had a significantly lower mean number of SVs compared to both fathers (mean difference = 13.5, $p < .0001$) and probands (mean difference = 11.1, $p < .0001$). The

difference between fathers and probands was not significant at the $p < 0.05$ level (mean difference = 2.4, $p = 0.090$) (Figure 4.10.).

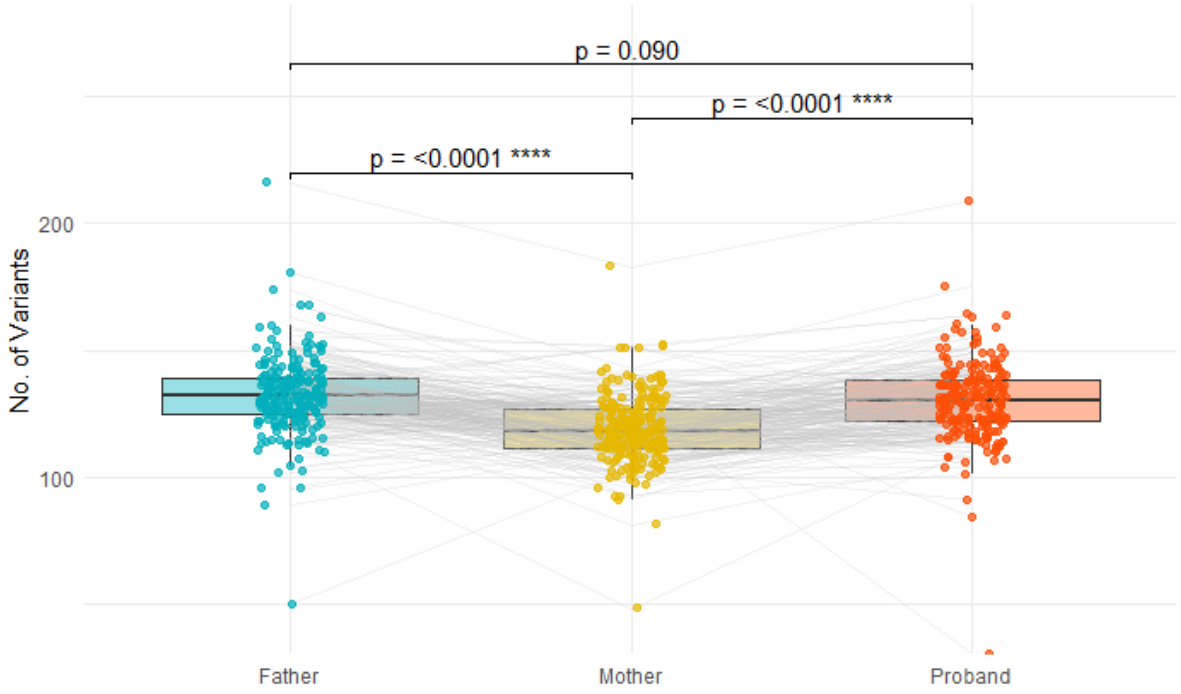


Figure 4.10. Average no. of the SVs identified by both tools in father, mother and probands on the autosomes. Mothers had significantly fewer variants than both fathers and probands ($p < .0001$). The father-proband comparison was not significant. Grey lines connect family members.

4.4 Discussion

In this chapter I present an overview of analysis of SVs in 216 patient-parent trios, laying the ground for detailed examinations in the upcoming chapters.

On average, 1099 SVs were identified per sample. Unlike single nucleotide variants, there is no average number of SVs in the human genome due to the influence of numerous factors, such as sequencing platforms and analysis tools. A recent large population-based study reported an average of 4,400 germline SVs per individual (Abel et al., 2020), while the GnomAD-SV estimated an average of 7,400 germline SVs per individual (Collins et al., 2020). Both of these studies utilised Illumina short-read sequencing technology. The lower number of SVs identified in our study can likely be attributed to our threshold, which is 1kb, whereas the aforementioned studies used a threshold of 50 bp. Additionally, recent comprehensive multi-platform studies have reported higher numbers of SVs, such as 13,000 SVs in a trio (Zook et al., 2020) and 27,000 SVs per germline genome (Chaisson et al., 2019). The discrepancy in

SV counts across studies highlights the impact of varying methodologies and sequencing approaches.

Four outlier samples with significantly increased numbers of SVs, especially deletions, were observed in our study. The primary cause of such outliers can often be attributed to sequencing artifacts, poor DNA sample quality, or technical issues during library preparation (Khayat et al., 2021). Proper filtering of these outliers was critical to avoid biases and misinterpretations during downstream analyses. Therefore, all SVs smaller than 5kb from the 4 outlier samples were excluded from further analysis.

When examining the distribution of SVs across autosomes, I found that the number of SVs was generally consistent with chromosome length, as expected. However, when normalised by effective chromosome size, an uneven SV distribution was observed, with chromosomes 19-22 showing a higher number of CNVs compared to chromosomes 1-18. For chromosomes 1-18, the distribution was relatively even. A similar pattern has been observed in parents as well as in research on CNV mapping of the human genome using large-scale samples (Zarrei et al., 2015). Similar to our observation, the highest proportion was found in chromosomes 19 and 22, while the lowest proportion was found in chromosomes 5, 8, and 18. One hypothesis for this observation is that these chromosomes are known to be particularly gene-dense, which could lead to increased rates of transcription-driven SV formation.. Our data supports the initial premise of this hypothesis, showing a significant positive correlation between the absolute number of genes and the number of SVs per chromosome ($r = 0.65$, $p = 0.001$). However, a more nuanced picture emerged when the SV count was normalised directly by gene density. This analysis revealed that, contrary to the initial hypothesis, chromosomes 19 and 22 have a relatively low SV burden for their high density. Conversely, gene-poor chromosomes, notably chromosome 18, exhibited a disproportionately high number of SVs relative to their gene content. This suggests that while gene content could be a contributing factor, it is not the sole determinant of SV rates, and that other chromosome-specific features are major drivers of genomic instability.

For the distribution along the chromosomes, as we observed, Zarrei et al., (2015) and Wong et al., (2013) also noted that CNVs are dispersed differently, with SVs appearing to be enriched in subtelomeric and pericentromeric regions. It may explain uneven distribution in chromosome 19-22. It could be due to the complexity of these regions and the challenges of accurately assaying them. The technical limitations in calling SVs in these regions should be

taken into consideration when interpreting results. It should also be kept in mind that they likely also represent genuine SVs arising from non-homologous allelic recombination mediated by low copy repeats during meiosis.

Regarding the size distribution of SVs, we found that most variants were less than 10kb, with a significant proportion being even smaller than 2.5kb. This is consistent with previous studies indicating that smaller SVs are more common and generally involve less functional genomic elements (Abel et al., 2020; Collins et al., 2020). In our analysis, approximately 176,000 deletions were identified, which is fourfold greater than duplications. This skew towards deletions, which is probably due to detection bias as deletions are easier to detect than duplications using read depth information, has also been reported in other WGS studies (Zarrei et al., 2015). Of all identified SVs, a substantial proportion (70%) were common in the population. Additionally, rare SVs were predominantly located in intergenic regions suggest that they are unlikely to have a direct impact on protein-coding genes, although their effects on regulatory elements should not be disregarded.

Analysis of sex chromosome SVs, using the latest version of CNVRobot (v4.2) and dysgu-sv, revealed a total of 6,437 SVs, of which 625 were rare inherited variants. The use of the updated CNVRobot provided improved sensitivity for detecting SVs on sex chromosomes, which are known to be more challenging due to their repetitive nature and lower sequence coverage (Sudmant et al., 2015; Massaia & Xue, 2017; Miga et al., 2020). The analysis of sex chromosomes is particularly important in the context of male infertility (Massaia & Xue, 2017; Rogers, 2021). Recurrent microdeletions on the Y chromosome in Azoospermia Factor (AZF) regions are the most commonly known genetic anomalies in men with azoospermia (Zhang et al., 2013). Additionally, Riera-Escamilla et al., (2022) conducted research with large-scale analyses of the X and revealed 21 recurrently mutated genes associated with spermatogenic failure. SVs demonstrate a largely uniform distribution pattern across the X chromosome, like autosomes, with notable exceptions in the sub telomeric regions. While these sub telomeric enrichments may partially reflect technical limitations in variant detection due to repetitive sequences, they likely also represent genuine SVs arising from non-homologous allelic recombination mediated by low copy repeats during meiosis.

The detection of balanced SVs, inversions and translocations, was conducted using the dysgu-sv tool, which relies on split reads and discordant read pairs mapping. A total of 8,485 inversions were identified, with an average of 39 per proband. Of these inversions, 568 (6.7%)

are rare varying between 1kb and 218Mb in size. Previous studies have shown that the number of inversions per human genome varies but is generally comparable. For instance, Auton et al., (2015) and Collins et al., (2020) reported an average of approximately 10 and 14 inversions per individual, respectively, in their analysis, which is less than our finding but within a similar order of magnitude. The ability to detect and validate balanced SVs is crucial, as they may have functional implications in noncoding regions or affect regulatory interactions, despite not resulting in copy number changes (Weischenfeldt et al., 2013, Spielmann et al., 2018).

Translocations are very rare in germline genomes of healthy individuals. On average, a human genome harbors fewer than one translocation, meaning it is most common to find zero translocations in an individual's genome (Sudmant et al., 2015; Chaisson et al., 2019). This rarity is due to the fact that translocations can disrupt genomic integrity and are often deleterious, leading to negative selection against them (Weischenfeldt et al., 2013; Redin et al., 2017).

For the parents, saliva-derived DNA was used. Saliva-derived DNA generally yields lower coverage compared to blood-derived DNA due to lower DNA concentration and higher levels of bacterial contamination (Yao et al., 2020). Additionally, previous work by a former PhD student in our group demonstrated that saliva-derived DNA does not produce systematic coverage biases in WES, although lower average coverage was noted. Our approach of sequencing parental samples twice and merging the data allowed us to achieve comparable coverage. The higher average number of SVs detected in fathers (1,203 per sample) compared to mothers (946 per sample) appears driven by CNVRobot's increased call rate in males, potentially due to sex-specific normalisation artifacts from hemizygous X and Y regions or reduced filtering of bias-prone regions when using female diploid controls. In fact, a more robust statistical analysis using a LMM confirmed this disparity, revealing that mothers had significantly fewer high-confidence variants than both fathers ($p < .0001$) and probands ($p < .0001$). This finding suggests that the transmission of SVs is not uniform. Instead, the significant deficit of SVs in mothers, a pattern not seen in their probands who inherit from both parents, strongly supports the hypothesis that a systematic technical bias, rather than stable biological transmission, is influencing the variant counts in this cohort.

4.5 Conclusion

In conclusion, this chapter presents the demographics of SVs in 216 patient-parent trios. I discussed the methodological and technological differences in SV detection, the viability of saliva-derived DNA for WGS, SV distribution across and within chromosomes, SVs on sex chromosomes, the detection of balanced SVs, and SV burdens in parents and probands. Overall, this chapter lays the foundation for exploring the potential pathogenic effects of SVs, which may contribute to the genetic basis of male infertility in the studied cohort.

Chapter 5. *De Novo* SVs in Idiopathic NOA and Severe Oligozoospermia

5.1 Introduction

Male infertility affects approximately 7% of men worldwide, representing a significant public health concern (Krausz & Riera-Escamilla, 2018). Genetic factors are known to cause the most severe forms of isolated male infertility, however, around 40% of cases remain idiopathic despite extensive evaluation (Tüttelmann et al., 2018; Houston et al., 2021; Kasak & Laan, 2021). This persistence of male infertility at a high frequency in the population poses a paradox, given the negative selection pressure against mutations that impair reproductive fitness (Krausz & Riera-Escamilla, 2018). This paradox suggests that infertility is not merely inherited recessively but that *de novo* dominant as well as dominant maternally inherited models (on autosomes as well as the X chromosome) significantly contribute to the disease's prevalence.

De novo mutations, which arise spontaneously in the germline, are prominent contributors to various fitness-impairing diseases (Acuna-Hidalgo et al., 2016). It is hypothesised that a similar scenario exists for male infertility, where *de novo* variants could account for a significant portion of unexplained cases. In line with this hypothesis, the two most frequent genetic causes of severe male infertility, Y chromosome microdeletions and the presence of an additional X chromosome in Klinefelter syndrome, are both *de novo* events. These chromosomal abnormalities explain up to 5.4% and 3.5% of cases in large patient cohorts, respectively (Punab et al., 2016; Olesen et al., 2017; Tüttelmann et al., 2018). Beyond these known chromosomal anomalies, the contribution of *de novo* variants, particularly SVs, to the aetiology of male infertility remains largely unexplored. Recent studies have started to shed light on this area. Most relevant, Oud et al., (2021) performed WES on infertile men and identified novel genetic factors, suggesting that *de novo* mutations play a role in male infertility.

De novo SVs are generally rare in the general population (Acuna-Hidalgo et al., 2016). Early studies using microarray-based techniques estimated a *de novo* CNV rate of approximately 0.01 per generation, with higher rates observed in cohorts with autism spectrum disorder (Sebat et al., 2007). However, detection rates have historically been influenced by methodological limitations and sample sizes (Collins et al., 2020). More recent WGS studies have reported higher rates of *de novo* structural variations, suggesting that previous estimates

may have been conservative (Kloosterman et al., 2015; Belyeu et al., 2021). For instance, (Belyeu et al., 2021) identified *de novo* SV rates of 0.160 and 0.206 per generation in unaffected and autism families, respectively, indicating that patient cohorts may be enriched for such variants. In addition, Jung et al., (2024) found an overall mutation rate of 0.13 events per genome in their recently published study.

De novo point mutations are generally passed down through the paternal lineage, accounting for approximately 80% of such mutations (Goldmann et al., 2016; Jónsson et al., 2017). This bias is attributed to the higher number of DNA replication cycles during male gametogenesis, leading to increased opportunities for replication errors. Similar paternal bias has been observed in studies of *de novo* CNVs (Hehir-Kwa et al., 2011). However, in the context of male infertility, we might anticipate a deviation from this pattern due to negative selection against pathogenic variants affecting spermatogenesis genes in the male germline, potentially resulting in a higher proportion of *de novo* variants of maternal origin. By investigating the parent-of-origin of *de novo* SVs in male infertility patients, we can test this hypothesis and gain insights into the mechanisms underlying the disease and the potential role of maternal transmission.

In this chapter, I present the results of analysis of *de novo* SVs in WGS of 216 azoospermia and severe oligozoospermia patient-parent trios. I aim to identify *de novo* SVs and determine their contribution to the aetiology of male infertility. Additionally, I explore the parent-of-origin patterns of these *de novo* SVs to test the hypothesis of maternal bias in transmission within this patient population. By integrating our findings with recent advancements in the field, I seek to enhance the understanding of the genetic underpinnings of male infertility and highlight the importance of *de novo* SVs as potential pathogenic factors.

5.2 Aims

This chapter aims to

- Identify and validate *de novo* SVs from WGS data of 216 patient-parent trios.
- Determine the parent of origin of *de novo* SVs.
- Interpret the *de novo* SVs detected.
- Identify novel candidate dominant and X-linked male infertility genes.

5.3 Results

WGS data from 216 patients affected by azoospermia or severe oligozoospermia and their parents were processed with the optimised pipeline to identify SVs. As detailed in Chapter 3, CNVRobot and dysgu-sv were combined for SV detection. CNVRobot includes an integrated feature for predicting inheritance in trio studies, while for dysgu-sv, a custom script was developed to assess inheritance in trios. A total of 25,425 SVs were predicted as *de novo*, comprising 665 identified by both tools (with predictions based on CNVRobot), 8,937 uniquely by CNVRobot, and 15,823 uniquely by dysgu-sv. Identified SVs predicted as *de novo* were inspected visually using CNVRobot plots and the IGV. William Coppock, an undergraduate student in our group, systematically reviewed all SVs predicted as *de novo* by dysgu-sv, utilising HTML report and linked IGV snapshots. After inspection, a total of ten *de novo* SVs were identified, corresponding to 0.046 *de novo* SVs per patient. The ten *de novo* SVs were further confirmed with Sanger and qPCR validation, confirming exact breakpoints and the expected number of copies in their parents (see chapter 2). All ten *de novo* SVs passed validation, consisting of eight *de novo* deletions and two *de novo* duplications (Table 5.1, Table 5.2, respectively).

Table 5.1. The identified and validated *de novo* deletions, detailing proband ID, genomic coordinates, size, region and affected genes.

Proband	Genomic Location (GRCh38)	Size	Region	Genes
NIJ_MI_0584P	chrX:109522132-109564264	42kb	Intragenic (Exonic)	*NXT2*
NCL_MI_0090P	chr11:11845751-11848086	2kb	Intragenic (Intronic)	*USP47*
NIJ_MI_01258P	chr17:74424749-74426589	2kb	Intergenic (Enhancer)	*GPRC5C*
NIJ_MI_02080P	chr3:5586136-5842695	256kb	Intergenic	
IND_MI _053P	chr6:149093143-149106110	13kb	Intergenic	
NIJ_MI_00352P	chr5:35571276-35580327	9kb	Intergenic	
NIJ_MI_01166P	chr3:55346873-55354443	7kb	Intergenic	
NCL_MI_0167P	chr18:70448845-70453694	5kb	Intergenic	

Table 5.2. The identified and validated *de novo* duplications, detailing proband ID, genomic coordinates, size, region and affected genes.

Proband	Genomic Location (GRCh38)	Size	Region	Genes
NCL_MI_0054P	chr4:133053579-133086452	33kb	Intergenic	
NIJ_MI_00118P	chr12:37298001-37303000	5kb	Intergenic	

5.3.1 De Novo Deletions

A *de novo* deletion was identified on chromosome X (chrX:109522132-109564264), with the exact size of the deletion determined as 42,133 bp using split and discordant reads[5] (Figure 5.1.), which is approximately 35kb larger than previously determined using WES data (7kb). The *de novo* deletion in NIJ_MI_0584P must be of maternal origin since fathers do not transmit the X chromosome to male offspring. The deletion encompasses the entire *NXT2* gene (pLI=0.71). It also encompasses the distal enhancer of the *KCNE5* gene (GeneHancer id: GH0XJ109536[6]) (pLI=0.12). From population genetic data, the observed/expected (o/e) fraction for *NXT2's* loss-of-function (LoF) variants is zero, with an upper bound LoF o/e fraction (LOEUF) of 0.51 (gnomAD, v2.1.1, data not yet available for v4.1). Such low LoF o/e fractions

[5] A sharp decrease in read depth, discordant reads around the breakpoints, and soft-clipped reads—where parts of the reads misalign due to belonging to another part of the genome—indicate the exact breakpoints (Figure 5.1. B-C).

[6] *KCNE5* gene is around 88kb away from GH0XJ109536. Interactions between gene and GH0XJ109536 is evidenced by GTEx eQTLs and C-HiC experiments.

are uncommon and indicate intolerance to LoF variants, highlighting the biological importance of the encoded protein and the selective pressure on genetic variants that affect its function. For instance, *TEX11* is another X-chromosomal gene with a LoF o/e fraction of zero and is one of the most well-established genes associated with male infertility (Yatsenko et al., 2015; Wyrwoll et al., 2023). Further research was conducted with our collaborators to better understand the role of the gene in spermatogenesis, and the results will be discussed in the discussion section.

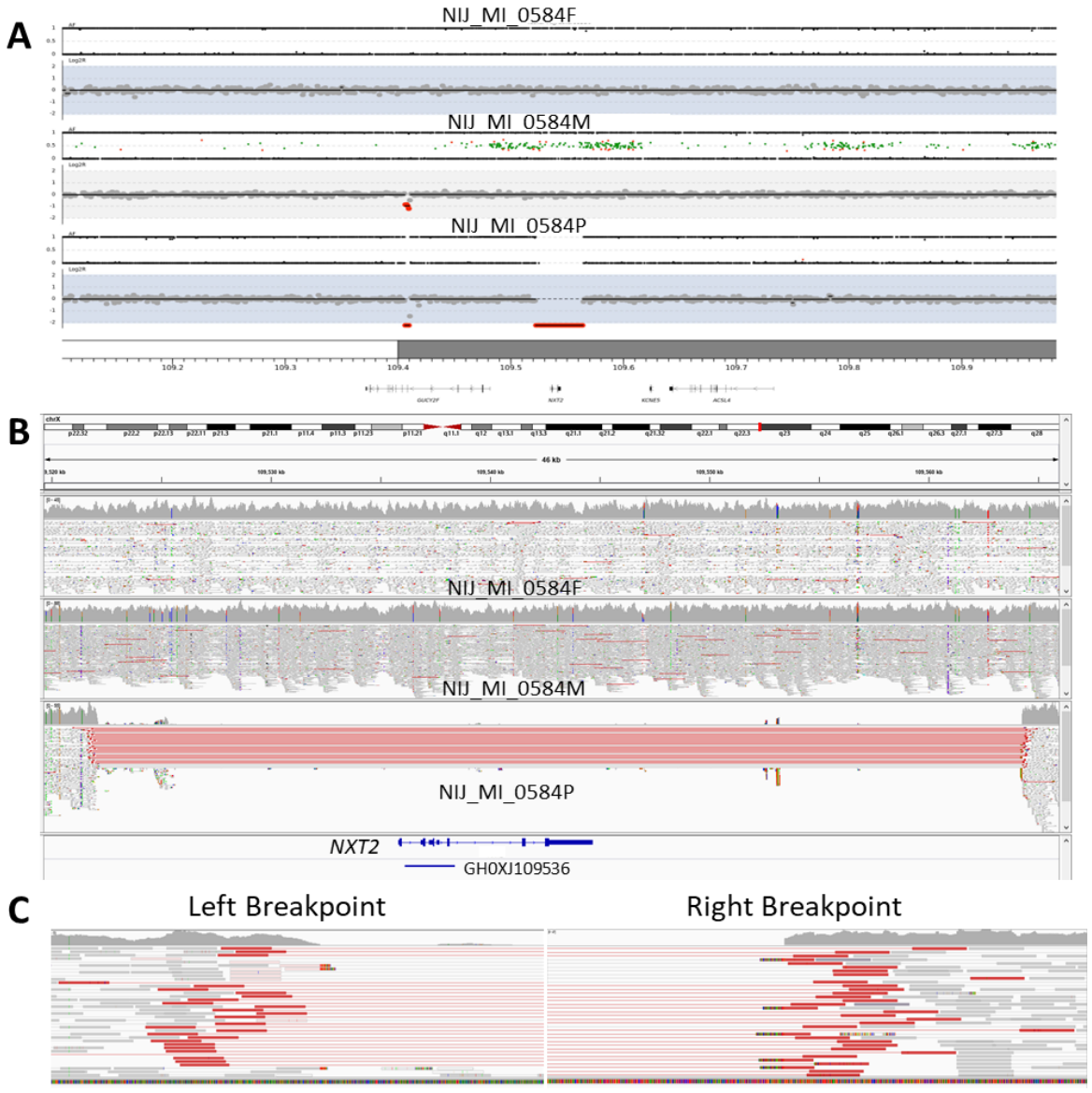


Figure 5.1. CNVRobot and IGV plots of the *de novo* deletion on chromosome X A. CNVRobot plot of the *de novo* deletion on chromosome X identified in azoospermic patient NIJ_MI_0584P. The deletion is present only in the proband. In proband, the MAF track confirms the deletion by absence of SNPs zygosity data within the affected region. Also, a log2R value of -2 indicates hemizygous deletion. **B.** IGV plot of the same *de novo* hemizygous deletion. Red lines represent ties of the discordant read pairs **C.** Zoom-in of Figure B, showing

both breakpoints of the deletion indicated by the presence of split and discordant reads (in red) in patient NIJ_MI_0584P.

A *de novo* deletion was detected on chromosome 11 of patient NCL_MI_0090P with azoospermia. The CNVRobot and IGV plots (Figure 5.2.) clearly show the heterozygous deletion in the proband and the absence of this event in both the mother and the father. This region has not been previously reported as deleted in the population databases. This 2,335 bp heterozygous deletion occurred within the first intron of *USP47* gene (pLI=1) and removed part of the Promoter/Enhancer regulatory element (6%) of the *USP47* gene (GeneHancer id: GH11J011840)[7]. The *USP47* gene is predicted to be likely dominant by DOMINO (Quinodoz et al., 2017). There was no informative SNP to determine whether the deletion arose from the maternal or paternal allele.

Additional investigation of the *USP47* gene was conducted by examining SNVs across the entire cohort and CNVs in 234 individual cases. One heterozygous *de novo* SNV (*USP47*(NM_017944.4):c.40-776A>G) in the first intron in NIJ_MI_00602P with azoospermia and a 20kb heterozygous deletion (chr11:11,932,899-11,952,966) encompassing eleven exons in MAN_MI_0013P with oligozoospermia were identified (Figure 5.3). Even though *USP47* (NM_017944.4):c.40-776A>G is not present in the population frequency databases, it has a likely benign classification due to it not being located in an evolutionary conserved region.

[7] The deletion removed the distal-enhancer like signature (ENCODE ID: EH38E1521013) of Promoter/Enhancer regulatory element (GeneHancer id: GH11J011840).

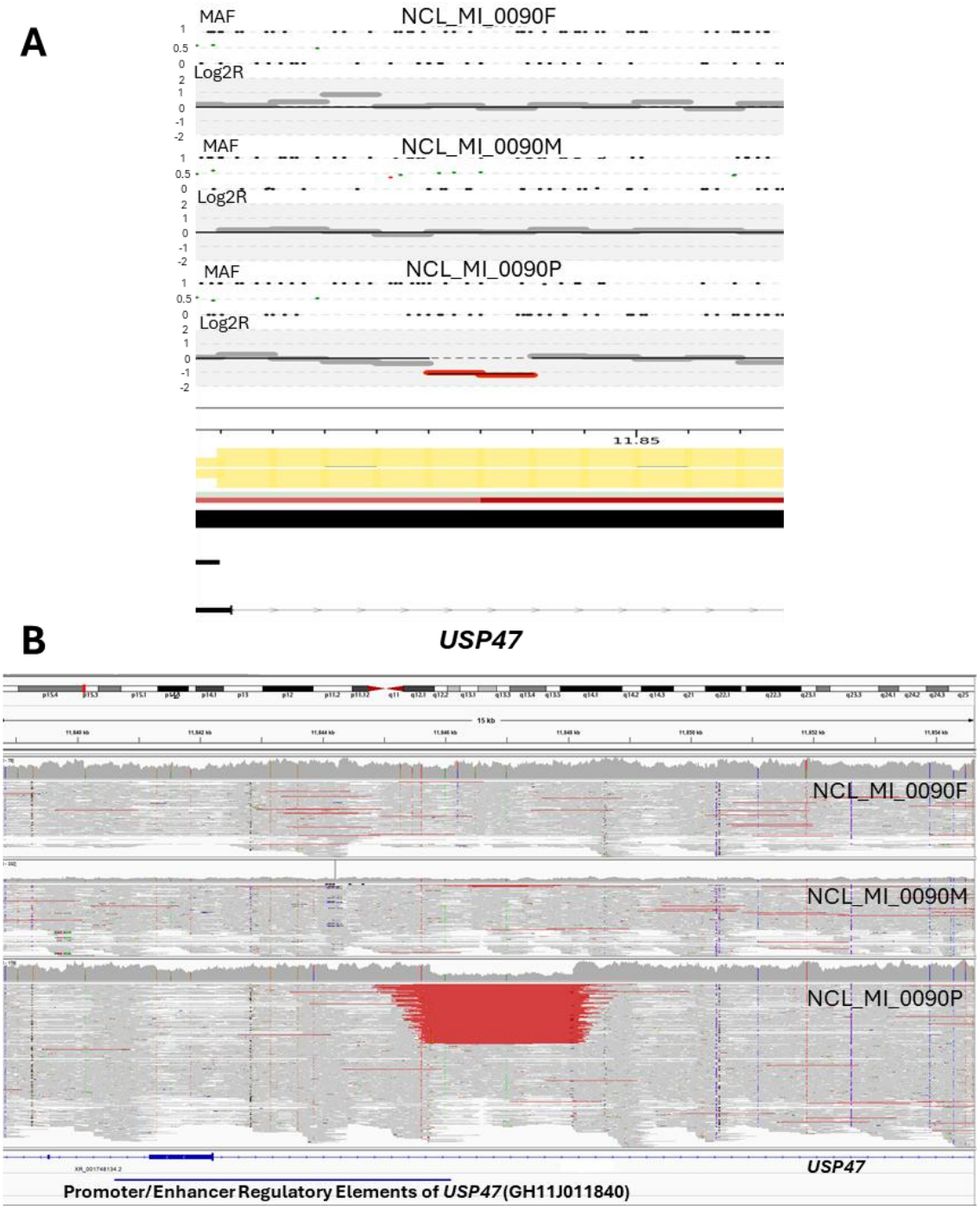


Figure 5.2. CNVRobot and IGV plots of heterozygous *de novo* deletion on chromosome 11 identified in azoospermic patient NCL_MI_0090P. A. CNVRobot plot of the heterozygous *de novo* deletion. The deletion is present only in the proband. In proband, no heterozygous SNP was observed (MAF=0.5) within the deleted region. Also, a log2R value of -1 indicates heterozygous deletion. **B.** IGV plot of the heterozygous *de novo* deletion detected in NCL_MI_0090P. Red lines represents ties of the discordant read pairs.

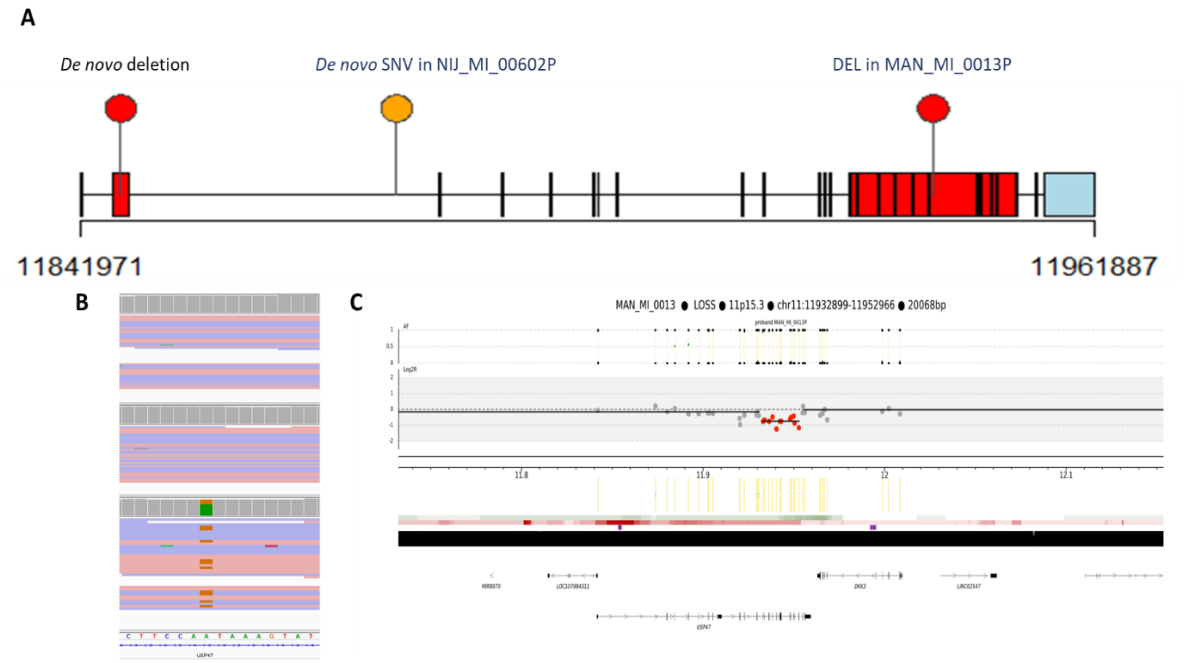


Figure 5.3. A. Illustration of identified CNVs and SNV across the *USP47* gene. Rectangles represent exons, red indicates deletions and yellow SNV. **B.** IGV plot of the *de novo* SNV detected in NIJ_MI_00602P. **C.** CNVRobot plot of the heterozygous deletion detected in patient MAN_MI_0013P by WES.

The other *de novo* deletion was detected on chromosome 17 of NIJ_MI_01258P with severe oligoasthenozoospermia. The CNVRobot and IGV plots (Figure 5.4.) show the heterozygous deletion in the proband and the absence of this event in both the mother and the father. It was a novel deletion and was not observed in the population databases. This 1841 bp *de novo* deletion occurred within the regulator region of *GPRC5C* gene (pLI=0) and removed the part of the Promoter/Enhancer regulatory element (3%) of the *GPRC5C* gene in one allele (GeneHancer id: GH17J074426)[8]. There was no informative SNP to determine whether the deletion arose from maternal or paternal allele.

[8] The deletion removed the four distal-enhancer like signatures of Promoter/Enhancer regulatory element (GeneHancer id: GH17J074426).

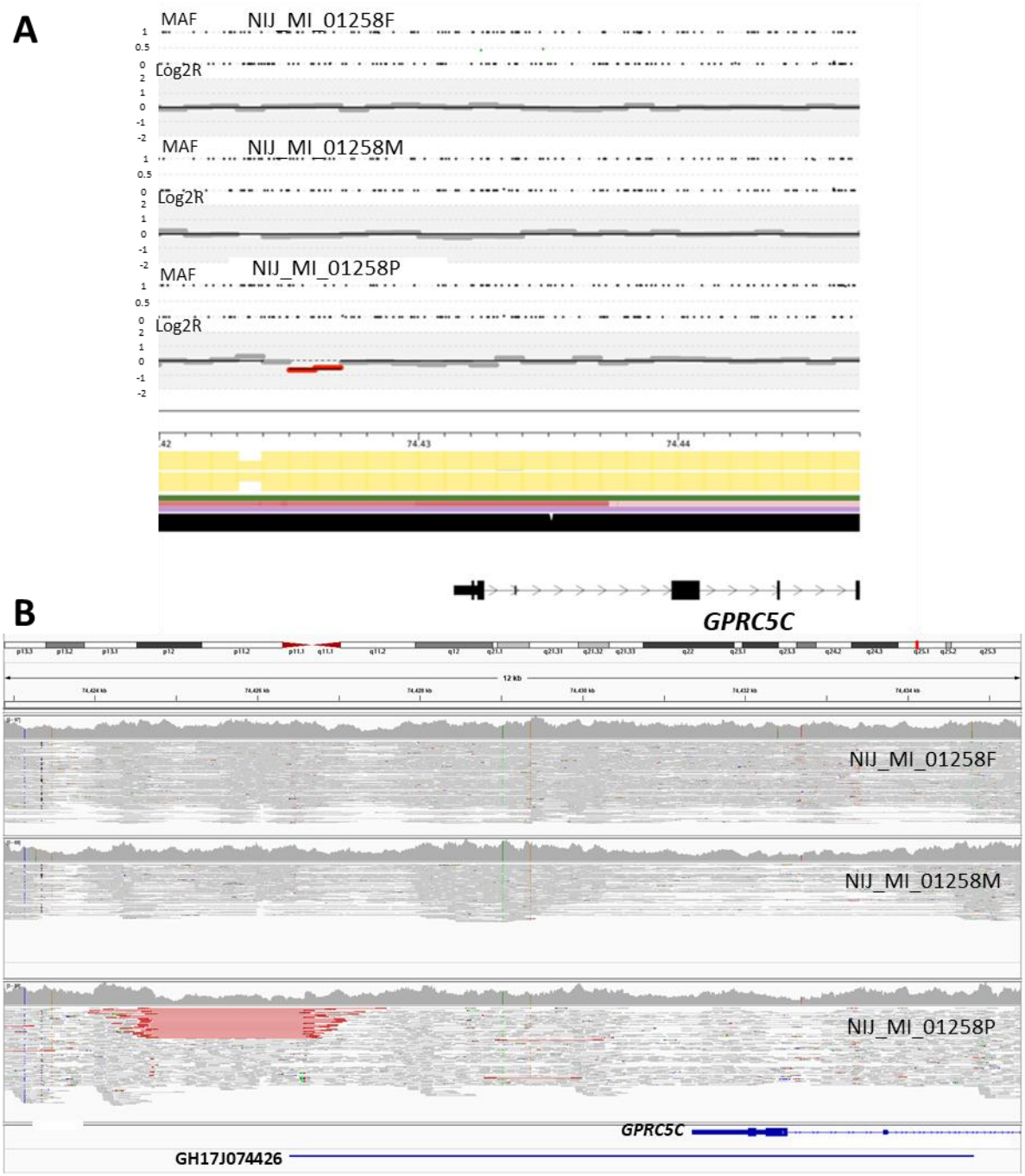


Figure 5.4 CNVRobot and IGV plots of the heterozygous *de novo* deletion on chromosome 17 detected in patient NIJ_MI_01258P with severe oligoasthenozoospermia. A. CNVRobot plot of the heterozygous *de novo* deletion. The deletion is present only in the proband. In proband, a log2R value of -1 indicates heterozygous deletion. **B.** IGV plot of the heterozygous *de novo* deletion detected in NIJ_MI_01258P, also showing disruption of the promoter. Red lines represent ties of the discordant read pairs.

The largest (256kb) *de novo* deletion was detected on chromosome 3 of NIJ_MI_02080P with azoospermia. The CNVRobot and IGV plots (Figure 5.5.A, 5.5.B) clearly show the heterozygous deletion in the proband and the non-occurrence of this event in both the mother and father. A 24kb smaller heterozygous deletion (232kb) was previously reported in the same region in the gnomAD database (gnomAD ID: DEL_3_29481, Allele Frequency = 0.00004610). It was identified in a healthy female who is from African/African American genetic ancestry group. The intergenic 256,560 bp *de novo* deletion arose from the maternal allele which was determined by using informative SNPs within the region.

There are 51 candidate cis regulatory elements and a Topologically Associated Domain (TAD) boundary[9], according to the ENCODE database, within the deleted region. This deletion potentially affects the regulation of *EDEM1* (pLI=0)*, ARL8B* (pLI=0.13)*, BHLHE40* (pLI=0.99)*, GRM7* (pLI=1, pHaplo=0.90), and two lncRNA genes; ENSG00000226022 and ENSG00000229642 (Figure 5.5.C), by either removing regulatory elements or disrupting the TAD boundary.

[9] These boundaries come from experiment (ENCODE ID: ENCSR834DXR) in which Hi-C was performed on human neuroblastoma cell line. Unfortunately, no data are available where TAD boundaries can be retrieved for the testis, but there might still be a TAD boundary at this region in the testis.

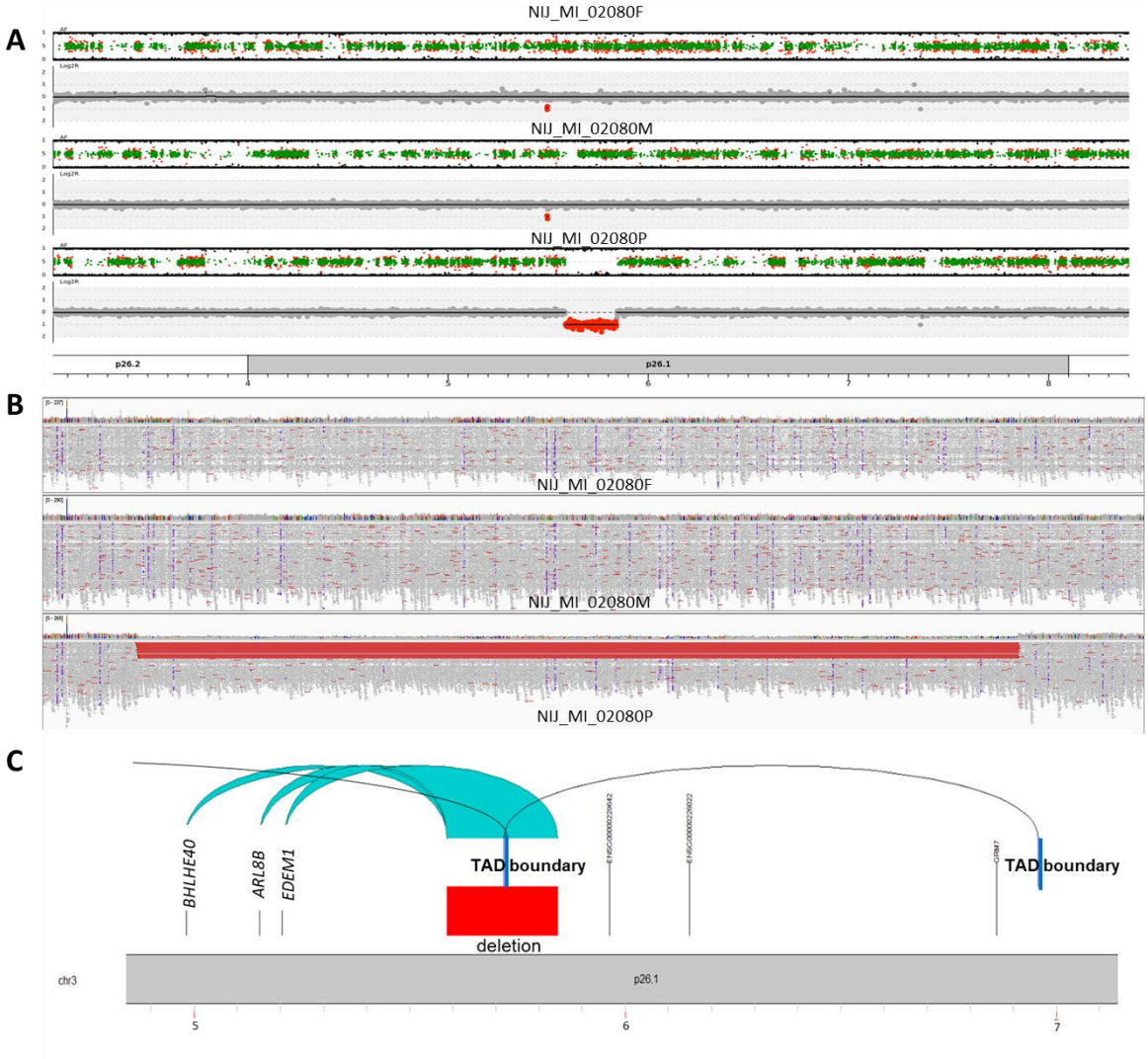


Figure 5.5. CNVRobot and IGV plots of the 256kb heterozygous *de novo* deletion on chromosome 3 identified in patient NIJ_MI_02080P with azoospermia. A. CNVRobot plot of the 256kb heterozygous *de novo* deletion. The deletion is present only in the proband. In proband, no heterozygous SNP was observed (MAF=0.5) within the deleted region. Also, a log2R value of -1 indicates heterozygous deletion. **B.** IGV plot of the 256kb heterozygous *de novo* deletion detected in NIJ_MI_02080P. Red lines represent ties of the discordant read pairs **C.** Functional annotation of the heterozygous *de novo* deletion.

The remaining four rare heterozygous *de novo* deletions, ranging from 5kb to 13kb, occurred in four different patients in the intergenic regions of autosomal chromosomes (Table 5.1.). The parent-of-origins of *de novo* SVs were determined where possible by using informative SNPs (see Table 5.3. for phasing of all *de novo* SVs). Among these, no overlap was found with any cCRE, and therefore, they were not examined further.

5.3.2 De Novo Duplications

Two heterozygous *de novo* tandem duplications were identified in 216 probands. Both occurred within intergenic region, and one was very close to the centromeric region of chromosome 12, making it challenging to investigate. The second *de novo* duplication was identified on chromosome 4 of NCL_MI_0054P with azoospermia. The CNVRobot and IGV plots (Figure 5.6) clearly present the heterozygous tandem duplication in NCL_MI_0054P. A heterozygous deletion was previously reported in the same region in gnomAD database (gnomAD ID: DEL_4_50236, Allele Frequency = 0.00004610). It was identified in a healthy female who is from European genetic ancestry group. This 32,873 bp *de novo* duplication occurred within the intergenic region of the genome and arose from the paternal allele which was determined by using informative SNPs within the region. There is one cCREs, according to the ENCODE database, within the duplicated region which is close (60kb away) to the *PCDH10* (pLI=0.89, pTriplo=0.49) gene.

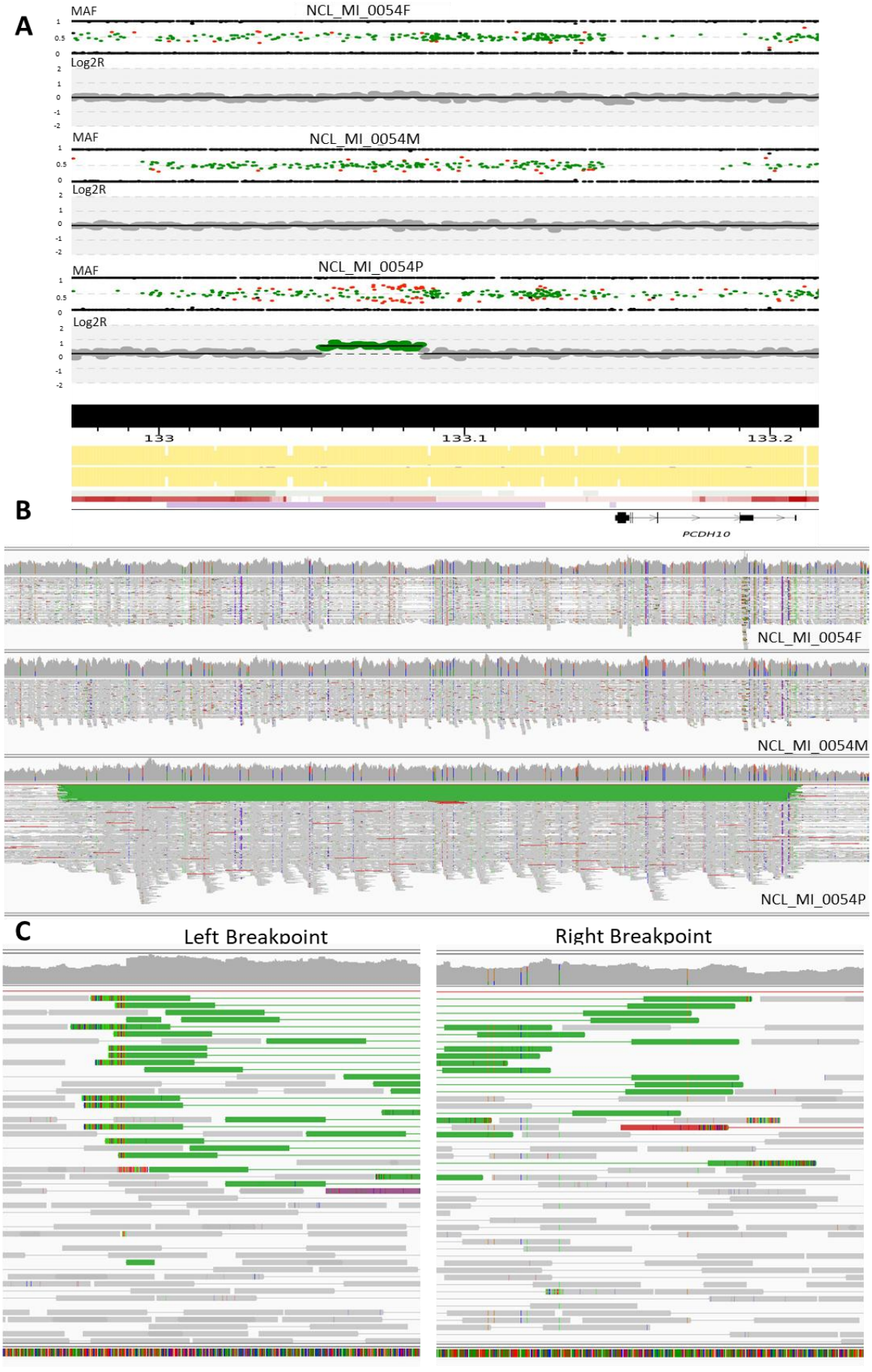

A
MAF
NCL_MI_0054F
Log2R
MAF
NCL_MI_0054M
Log2R
MAF
NCL_MI_0054P
Log2R
133
133.1
133.2
PCDH10
B
NCL_MI_0054F
NCL_MI_0054M
NCL_MI_0054P
C
Left Breakpoint
Right Breakpoint

Figure 5.6. CNVRobot and IGV plots of *de novo* duplication on chromosome 4 identified in patient NCL_MI_0054P. A. CNVRobot plot of the *de novo* duplication. The duplication is present only in the proband. MAF values deviate from the standard position of 0, 0.5, and 1, instead, intermediate values are observed, indicating additional copy. Also, an expected log2ratio value of approximately 0.58 indicates a heterozygous duplication. **B.** IGV plot of the *de novo* duplication identified in NCL_MI_0054P. **C.** Detailed breakpoints of the duplication showing split and discordant reads (in green) in patient NCL_MI_0054P.

5.3.3 Replication Study of Candidate Genes

The genes (*USP47, GPRC5C, BHLHE40*, and *PCDH10*), which may be affected by *de novo* SVs and potentially contribute to the observed patient phenotypes, were further investigated in replication cohorts of infertile patients as well as fertile controls (see Chapter 2.5). I sought for potentially pathogenic or likely pathogenic LoF SNVs in these genes within the Genomics of Male Infertility Group cohort, the MERGE study from Germany, and the fertile control cohort. One potentially pathogenic frameshift SNV (NM_003670.3:c.900_904dup (p.Thr302IlefsTer7)) was identified in the *BHLHE40* gene in singleton NIJ_MI_2291P, a patient with azoospermia from Nijmegen. No pathogenic LoF SNVs were identified in the other genes or in the fertile control cohort.

5.3.4 Phasing of *De Novo* SVs

I was able to determine the parent of origin for 7 out of 10 *de novo* SVs (see chapter 2). Of these, 5 arose from the paternal allele (71.43%), while 2 arose from the maternal allele (28.57%) (Table 5.3).

Table 5.3. The parent of origin of identified *de novo* SVs.

Proband	Genomic Location - SV Type (GRCh38)	Parent of Origin
NIJ_MI_0584P	chrX:109522132-109564264 - DEL	Maternal
NCL_MI_0090P	chr11:11845751-11848086 - DEL	Unable to determine
NIJ_MI_01258P	chr17:74424749-74426589 - DEL	Unable to determine
NIJ_MI_02080P	chr3:5586136-5842695 - DEL	Maternal
IND_MI _053P	chr6:149093143-149106110 - DEL	Paternal
NIJ_MI_00352P	chr5:35571276-35580327 - DEL	Paternal
NIJ_MI_01166P	chr3:55346873-55354443 - DEL	Unable to determine
NCL_MI_0167P	chr18:70448845-70453694 - DEL	Paternal
NCL_MI_0054P	chr4:133053579-133086452 - DUP	Paternal
NIJ_MI_00118P	chr12:37297735-37303000 - DUP	Paternal

5.3.5 A *De Novo* Deletion Identified in German Trios

I performed CNV detection in WES data of 29 patient-parent trios using CNVRobot (see chapter 2 for cohort description and CNV analysis in WES data). The identified CNVs were analysed by our undergraduate student Thomas Giles, under the supervision of myself, Dr. Giles Holt and Prof. Joris Veltman.

A total of 221 CNVs were predicted to be *de novo* in the 29 patient-parent trios. Following visual inspection, a 525 kb *de novo* deletion on chromosome 16 was identified in proband 1280 (chr16:29663628-30188773) (Figure 5.7). This deletion was previously catalogued in the gnomAD database (gnomAD ID: GD_16P11.2-BP4-BP5_DEL, allele frequency = 0.0002071). The ClinGen expert panel has concluded that there is sufficient evidence to classify this deletion as pathogenic, with evidence of incomplete penetrance. The deleted region is predominantly linked to severe developmental phenotypes, including intellectual disability and developmental delay. However, (Auwerx et al., 2024) reported that CNVs in this region have since been associated with a broad spectrum of phenotypic alterations, characterised by high variability in expressivity and incomplete penetrance. No fertility phenotype was described for patients with this deletion. This *de novo* deletion resulted in the loss of one copy of 29 genes. Among these, *MAZ* emerged as a candidate gene due to its role in male reproductive development, potentially contributing to the observed phenotype. MAZ, a transcriptional regulator within the Wnt signalling pathway, has been previously implicated in genitourinary birth defects and urogenital development as well as kidney, and bladder abnormalities (Punjani et al., 2021).

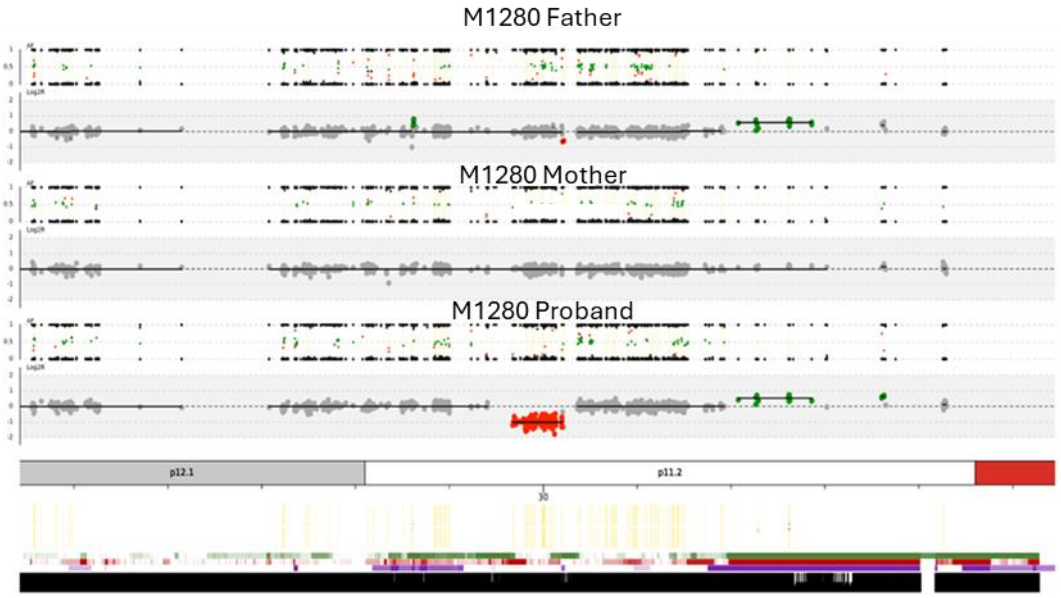


Figure 5.7. CNVRobot plot of the heterozygous *de novo* deletion on chromosome 16 detected in patient M1280. The deletion is present only in the proband. In proband, a log2R value of -1 indicates heterozygous deletion.

5.4 Discussion

The study presented in this chapter aimed to investigate the presence, frequency and role of *de novo* SVs outside the AZF regions in severe male infertility. WGS data from 216 patients affected by azoospermia or severe oligozoospermia and their parents were analysed. A total of ten *de novo* SVs were identified in 216 patient-parent trios, corresponding to approximately 0.046 *de novo* SVs per patient.

This rate is higher than the *de novo* CNV rates of approximately 0.01 per generation reported in early investigations of the general population using microarray technologies (Sebat et al., 2007). On the other hand, it is lower than the reported *de novo* SV rates of 0.160 by Belyeu et al., 2021, based on WGS studies in general population. It is indeed difficult to compare our rate with the literature since the detection rates of *de novo* SVs have been influenced by methodological constraints and the sizes of the cohorts studied (Collins et al., 2020). It has also been shown that the rate of *de novo* SVs is higher in individuals with certain rare diseases, such as intellectual disability, than in the general population (Sebat et al., 2007; Vissers et al., 2016; Jung et al., 2024). Therefore, while a direct comparison of the *de novo* SV rate is challenging, there does not seem to be an increased burden of *de novo* SVs in severe male infertility. However, it is important to note that patients with *de novo* SVs on the Y chromosome were excluded in our study, thus, our *de novo* SV rate for male infertility patients is likely underestimated. To determine the true rate of *de novo* SV occurrence in this specific form of male infertility, an unbiased analysis with larger cohorts and control sets is required.

The only exonic *de novo* deletion identified occurred on the X chromosome. This X-linked *de novo* deletion has not been previously reported in the population databases and removes the only copy of the *NXT2* gene (pLI=0.71) and the distal enhancer of the *KCNE5* gene (pLI=0.12). In the Protein Atlas database, *NXT2* exhibits moderate protein expression levels and high RNA expression levels in the testis. Additionally, the gene is evolutionarily conserved in eutherian mammals, and a previous study has indicated that the gene plays a non-essential role in mouse fertility (Khan et al., 2018). However, in mice, gene expression is not predominantly found in the testis unlike in humans, making the mouse not a viable comparative model for further investigating this gene. To better understand the role of the *NXT2* gene in spermatogenesis, our group collaborated with the International Male Infertility Genetics Consortium (IMIGC) and Dicke et al., 2024, conducted a study investigating the specific function of *NXT2* in nuclear RNA export within the human testis. This study aimed to explore

the molecular interactions of *NXT2* and its role in male fertility, focusing on patients with NOA, a condition in which sperm production is severely impaired. Our collaborators first investigated the interaction partners of *NXT2* in testicular tissue through immunoprecipitation and mass spectrometry. The analysis revealed that *NXT2* interacts with nuclear export factors *NXF1, NXF2*, and *NXF3*, as well as nucleoporins *NUP93* and *NUP214*. These findings support the hypothesis that *NXT2* plays a role in a testis-specific RNA export pathway that differs from the ubiquitous *NXT1-NXF1* pathway. To further validate these interactions, co-immunoprecipitation experiments confirmed that *NXT2* directly binds to *NXF2* and *NXF3*, both of which are testis-specific proteins involved in spermatogenesis. Additionally, gene ontology analysis revealed that *NXT2* and its interaction partners are enriched in biological processes related to nuclear export and mRNA transport. In parallel, they screened exome and genome sequencing data from more than 2,700 infertile men to identify mutations in *NXT2* that might be linked to male infertility. We identified two additional patients with rare, deleterious nucleotide variants in *NXT2*. These cases as well as our case highlight the critical role of *NXT2* in spermatogenesis. The first case, M3065, presented with a hemizygous single nucleotide duplication (*NXT2*(NM_018698.5):c.354dup(p.Asp119Ter)), resulting in a premature stop codon in the *NXT2* gene. This truncating mutation is predicted to lead to nonsense-mediated decay (NMD) of the *NXT2* mRNA, effectively abolishing the production of functional *NXT2* protein. Testicular biopsy revealed a SCO phenotype, where germ cells were absent in most seminiferous tubules. The patient had elevated FSH levels, indicating impaired spermatogenesis. Importantly, the same *NXT2* mutation was present in two of his infertile brothers, further reinforcing the genetic link between this mutation and non-obstructive azoospermia. No sperm could be retrieved through TESE, highlighting the severe impact of this mutation on fertility. The second case, M2004, involved a missense mutation in *NXT2* (*NXT2*(NM_018698.5):c.268G>T (p.Ala90Ser)), resulting in an alanine-to-serine substitution at position 90. This mutation affects the NTF2-like domain, which is crucial for *NXT2's* interaction with *NXF2* and *NXF3*. Functional assays revealed that this mutation causes abnormal splicing, producing both correctly and aberrantly spliced transcripts. The resulting protein, lacking part of the NTF2-like domain, showed impaired binding to its interaction partners. Testicular biopsy from M2004 revealed some spermatogenic activity, with a few seminiferous tubules containing spermatogonia and spermatocytes. However, no sperm could be retrieved for fertility treatments, indicating that while this mutation does not completely eliminate *NXT2* function, it severely disrupts spermatogenesis. The results from these cases suggest that *NXT2*

is essential for normal germ cell development in humans. Loss-of-function mutations in *NXT2* lead to severe defects in sperm production, ranging from a complete absence of germ cells to the production of abnormal sperm. These findings demonstrate that *NXT2* is not only involved in the export of mRNA in testicular cells but also plays a critical role in the early stages of germ cell development, likely during foetal and embryonic stages. The absence of *NXT2* cannot be compensated by its paralog *NXT1*, highlighting its unique and essential role in male fertility. In conclusion, this study establishes *NXT2* as a key player in human spermatogenesis and male fertility. The identification of *NXT2* mutations in infertile men underscores its importance in the nuclear export machinery within the testis. Further research is needed to identify the specific mRNAs transported by *NXT2* and to understand the precise molecular mechanisms through which *NXT2* supports spermatogenesis. Specifically, RNA immunoprecipitation followed by sequencing (RIP-Seq) on testicular tissue could be used to identify the direct mRNA cargo bound to *NXT2*, providing a clear list of its transport targets.

The other possibly affected gene, *KCNE5*, may be impacted by the disruption of a distal enhancer. The *KCNE5* gene encodes a member of a family of single-pass transmembrane domain proteins that function as ancillary subunits to voltage-gated potassium channels. A recent literature review concluded that potassium production is crucial for male fertility, not only during sperm capacitation but also throughout the processes of spermiogenesis and epididymal maturation (Delgado-Bermúdez et al., 2024). However, the *KCNE5* gene was found to have low expression in the testis and the gene has been associated with sudden cardiac death through gain-of-function mutations (Ohno et al., 2011). Also, it is noteworthy that this distal enhancer was identified in B and T cells rather than testis-specific data. Considering that the complete deletion of the *NXT2* gene is the cause of the patient's phenotype, it is unlikely that any potential disruption of this gene would be responsible for the phenotype.

Before moving to discussing non-coding *de novo* SVs, it worth noting that interpreting noncoding SVs presents significant challenges due to the complexity and limited functional annotation of the noncoding genome. Unlike coding regions, where the impact of mutations on protein function can be more directly assessed, SVs in noncoding regions, like those in gene enhancers, often have indirect and context-dependent effects on gene regulation, making pathogenicity predictions difficult (Castro et al., 2023). The vastness of the non-coding genome and the variability of regulatory element activity across different tissues and developmental stages add to the complexity (Ellingford et al., 2022).

One *de novo* deletion was identified within the first intron of the *USP47* gene in patient NCL_MI_0090P. The *USP47* gene encodes ubiquitin-specific protease that acts as a regulator of cell growth and genome integrity. The gene has low tissue specificity, based on the data from the Protein Atlas database. A *de novo* mutation in the Y chromosome gene *USP9Y*, which similarly encodes an ubiquitin-specific protease, has been reported in a man with azoospermia (Sun et al., 1999). Additionally, research conducted by Sinnar et al. revealed altered testicular gene expression in mice lacking polyubiquitin gene Ubb (Sinnar et al., 2011). In contrast, a study knocking out the *USP47* gene in mice did not report any effects on the fertility (Lei et al., 2022). Interestingly, I identified a 20kb heterozygous deletion affecting eleven exons of *USP47* gene in one of the singleton cases with oligozoospermia was identified. This deletion was classified as possibly pathogenic, as it removes part of the constrained *USP47* gene. While the functional impact of the *de novo* intronic deletion in the *USP47* gene remains to be elucidated, its significance is underscored by our identification of a more disruptive deletion in an additional patient in our cohort. A minigene splicing assay could be used to investigate whether this *de novo* intronic deletion leads to aberrant splicing. Also, to strengthen the potential link between *USP47* disruption and male infertility, a next step could be to identify additional cases through collaborations and by utilising gene-matching platforms such as GeneMatcher (Sobreira et al., 2015). If further individuals with relevant *USP47* variants and a similar phenotype are found, functional studies, such as knock out mice, would be essential to understand the role of *USP47* gene.

One of the other *de novo* deletions, in patient NIJ_MI_01258P with severe OAT, occurred within the regulator region of the *GPRC5C* gene (pLI=0) and removed part of the Promoter/Enhancer regulatory element (3%) of the *GPRC5C* gene. The *GPRC5C* gene encodes G-protein coupled receptor family C group 5 member C, belonging to the type 3 G protein-coupled receptor family, characterised by a signature seven-transmembrane domain motif. These seven-transmembrane receptors, also known as G protein-coupled receptors (GPCRs), reside on the cell membrane, translating extracellular signals into significant physiological effects (Insel et al., 2019). Extensively studied across various diseases such as cardiovascular disease, diabetes, obesity, depression, cancer, and Alzheimer's disease, GPCRs represent crucial targets for therapeutic interventions (Yang et al., 2021). According to the protein atlas database, GPRC5C exhibits its highest expression levels in peripheral tissues, notably the stomach, kidney, liver, pancreas, and prostate. Although its expression in the testis is low, it

demonstrates relatively higher levels in the epididymis and seminal vesicle. Zhang et al. (2020) provide an overview of the roles of various GPCRs, including ADGRG2, AT2 receptors, LGR4, GPER, and adenosine receptors, in the efferent ductules and epididymis. They highlight the critical involvement of several GPCR superfamily members in maintaining ion-water homeostasis in the epididymis, developing the efferent ductules, establishing the blood-epididymal barrier, and facilitating sperm maturation. They also highlighted the need for further investigation into the intricate signalling pathways of GPCR members within the epididymis and their specific physiological functions that influence male fertility. Additionally, *GPRC5C* has been shown to regulate olfactory cilia composition and length (Bhat et al., 2023). Despite the function of *GPRC5C* in the reproductive system yet to be discovered, disruptions in its regulation may impact sperm maturation and movement in patient NIJ_MI_01258P with severe OAT.

The largest (256kb) *de novo* deletion was identified on chromosome 3 of NIJ_MI_02080P with azoospermia. There are 51 candidate cis regulatory elements and a TAD boundary, according to the ENCODE database, within the deleted region. This deletion potentially affects the regulation of *EDEM1* (pLI=0)*, ARL8B* (pLI=0.13)*, BHLHE40* (pLI=0.99)*, GRM7* (pLI=1), and two lncRNA genes; ENSG00000226022 and ENSG00000229642 (Figure 5.5.C) by either removing regulatory elements or disrupting TAD boundary. Two out of four coding genes, which are located close to the deletion, are classified as dosage sensitive, as exhibit high pLI scores (pLI ≥ 0.99). Of those *BHLHE40* gene encodes a transcription factor which has been attributed to play a role in Sertoli cell signalling and Sertoli gene activation (Fice & Robaire, 2023). Although the role of *GRM7* in spermatogenesis is yet to be discovered, it was showed up-regulated in spermatogenesis (Herati et al., 2017). Moreover, ENSG00000226022 and ENSG00000229642 are highly expressed in testis and prostate compared to other tissues based on RNA-Seq Expression Data from GTEx (53 tissues, 570 donors). The exact role of lncRNA in spermatogenesis and male infertility is still under investigation but when the current studies taken into consideration (Joshi & Rajender, 2020; Kyrgiafini et al., 2022), possible disruption in the regulation of these genes might cause inappropriate regulation of spermatogenesis. Additional investigation of the *BHLHE40* genes was conducted in replication cohorts as well as fertile cohort control. One heterozygous frameshift variant, classified as likely pathogenic, was identified in *BHLHE40* gene in the replication cohort, but none in the fertile controls. However, as this case is a singleton, we cannot determine whether the variant occurred *de novo* or was

inherited, and if inherited, from which parent. Nevertheless, this represents the second case with the same phenotype carrying a potentially pathogenic LoF mutation in *BHLHE40*, highlighting the need for more cases and further functional studies. To this end, *BHLHE40* could be submitted to a matchmaking database such as GeneMatcher (Sobreira et al., 2015). While not commonly used in the male infertility research field, this platform could help identify recurrent cases in independent cohorts. To investigate the gene's function, CRISPR-Cas9 could be used to knock down *BHLHE40* in a Sertoli cell line, followed by RNA-sequencing to identify its downstream gene targets and clarify its role in Sertoli cell signalling.

One of the two *de novo* duplications occurred in the intergenic region of the chromosome 4 in patient NCL_MI_0054P. There is one cCREs, according to the ENCODE database, within the duplicated region which is close by *PCDH10* (pLI=0.89, pTriplo=0.49) gene. This gene is previously reported as disrupted by a *de novo* balanced insertional chromosome translocation (2p24;4q28.3q31.22) in a patient with azoospermia (Tzschach et al., 2009). Even though the disruption of the regulation of the *PCDH10* gene by this duplication needs further understanding, it could still possibly affect the gene and therefore cause the infertility in this patient.

In proband 1280 from German cohort, a 525 kb *de novo* deletion on chromosome 16 encompassing 29 genes, including *MAZ*, was identified. The region is linked to severe phenotypes like intellectual disability but with incomplete penetrance (Auwerx et al., 2024). The lack of severe phenotypes in this patient may reflect incomplete penetrance, though *MAZ* haploinsufficiency could contribute to fertility issues. Alternatively, this deletion may be unrelated to the patient’s infertility, suggesting other genetic factors at play.

Lastly, I determined the parent of origin for 7 out of 10 *de novo* SVs, with 5 (71.43%) arising from the paternal allele and 2 (28.57%) from the maternal allele. These findings are consistent with previous studies that have shown a paternal bias in the transmission of *de novo* mutations. For *de novo* point mutations, they are known to be transmitted predominantly through the paternal lineage, accounting for approximately 80% of such mutations (Goldmann et al., 2016; Jónsson et al., 2017). This paternal bias has also been observed for *de novo* CNVs (Hehir-Kwa et al., 2011). However, in the context of male infertility, it was hypothesised that this typical paternal bias might not hold due to the potential for negative selection against pathogenic variants affecting spermatogenesis genes in the male germline. Such selection could prevent deleterious mutations from being passed on to offspring, possibly resulting in a

higher proportion of *de novo* variants being maternally derived. My results, however, do not indicate a significant deviation from the expected paternal bias, as the majority of *de novo* SVs (71.43%) were still of paternal origin. In addition, my colleagues Holt et al., 2022, phased and determined the parent-of-origin for 77 of 109 *de novo* SNVs (71%) from research conducted on NOA patients with WES by Oud et al., 2021, using LRS, with 64 of these (83%) being of paternal origin. The results revealed that 75% of those classified as likely pathogenic occurred on the paternal allele. The consistent paternal bias across both SVs and SNVs raises questions about how these potentially pathogenic variants escape negative selection in the paternal germline and are transmitted to offspring, despite their potential to impair fertility. Three potential mechanisms were suggested to explain this phenomenon (Oud et al., 2021).. First, the *de novo* variant may arise after the temporal window in which the fertility gene is active, thereby bypassing its negative effects on fertility. Second, the variant may affect a gene that is not directly involved in spermatogenesis but is essential for the germline in the offspring. Third, the phenomenon of spermatogonial cells sharing mRNA and proteins within cysts could mask the deleterious effects of a mutation in one cell, allowing the variant to be passed on via fully functioning sperm. These mechanisms could explain how *de novo* pathogenic variants of paternal origin contribute to male infertility. Further research is needed to elucidate these mechanisms and to investigate the timing and selection pressures acting on *de novo* variants in the context of male infertility.

5.5 Conclusion

In this chapter, I explored the role of *de novo* SVs in male infertility by analysing WGS data from 216 patient-parent trios affected by azoospermia or severe oligozoospermia. Ten *de novo* SVs were identified, corresponding to approximately 0.046 *de novo* SVs per patient. One *de novo* deletion was found on the X chromosome, removing the *NXT2* gene, a key player in spermatogenesis. Collaborative research revealed that *NXT2* functions in a testis-specific RNA export pathway, and mutations in this gene were linked to impaired sperm production. Additionally, a possible regulatory disruption of genes by *de novo* SVs suggests that noncoding SVs also play a significant role in male infertility. Particularly, a *de novo* deletion in the intronic region of the *USP47* gene is likely to be causative, as another more disruptive deletion in the same gene was identified in a different patient within our cohort. Moreover, a 525 kb *de novo* deletion, which is linked to severe phenotypes, could contribute to infertility by disrupting the

MAZ gene. These results suggest that both coding and noncoding *de novo* SVs outside the Y chromosome contribute to the aetiology of male infertility.

Chapter 6. Rare Maternally Inherited SVs in Idiopathic NOA and Severe Oligozoospermia Cohort

6.1 Introduction

In the past chapter, the role of *de novo* SVs was examined in 216 patient-parent trios affected by idiopathic NOA and severe oligozoospermia. Building upon these findings, this chapter explores the potential contribution of rare maternally inherited (MI) SVs to the aetiology of male infertility in the same cohort. SVs from the mother, whether inherited from her family or arising as a *de novo* event, can disrupt genes critical for spermatogenesis. This may cause infertility in her male offspring without affecting her own fertility. This is biologically plausible due to the fundamental differences between gametogenesis: male spermatogenesis is a continuous process, whereas female oogenesis is finite. Therefore, a mutation in a gene essential solely for spermatogenesis is unlikely to impact female fertility but can cause infertility when passed to a son.

Historically, several studies have documented MI translocations (Chandley et al., 1972; Chandley et al., 1975) and other chromosomal abnormalities (Smith et al., 1965) causing infertility in males but not in females. However, the role of individual genes was not examined in these studies due to the limited resolution of genomic techniques available at that time. More recently, Sazci et al., 2005, described a male patient with primary infertility carrying a balanced translocation inherited from his mother, who showed normal fertility. Additionally, Hodžić et al., 2021, identified two maternally inherited, predicted pathogenic, mutations in known male infertility genes, a heterozygous frameshift variant in the *FKBPL* gene and an in-frame deletion in the *UPF2* gene, through WES of 13 infertile males with idiopathic azoospermia and their parents. Moreover, Ji et al. (2021) described a family, where a pathogenic *TEX11* splicing variant was passed from an unaffected carrier mother to her four sons and two grandsons (via her carrier daughter), all of whom had azoospermia.

Sex chromosomes are of particular interest since both the X and Y chromosomes in males are enriched with genes specifically expressed in the testis (Mueller et al., 2008). The gene content of the X chromosomes is conserved among placental mammals (Mueller et al., 2013). Rare or *de novo* mutations in these chromosomes can have significant phenotypic effects due to the absence of a second compensatory allele. The research on sex chromosomes in male infertility is expanding, however, according to latest systematic review by (Houston et al., 2021) there

are only 3 X-linked genes (*AR, TEX11, and USP26)* for which there is sufficient clinical evidence to include them in diagnostic testing for NOA and oligozoospermia. Nevertheless, recently published large-scale analyses of the X chromosome in 2,354 infertile men revealed 21 recurrently mutated genes strongly associated with NOA such as *RBBP7* (Riera-Escamilla et al., 2022).

While variants in the large non-recombining, X-specific region are indeed maternally inherited in males, this rule does not apply to the pseudoautosomal regions (PARs). Genes in the PARs, such as *SHOX*, are inherited in an autosomal-like fashion due to obligatory X-Y recombination in male meiosis. This process is essential for male fertility and can even result in father-to-son transmission of PAR-based variants. Therefore, our trio-based approach, while not required for the X-specific region, is uniquely suited to investigate complex inheritance on the autosomes and to properly resolve the pseudoautosomal inheritance of the PARs, a previously underexplored area in male infertility research. This is particularly significant given the lack of knowledge about autosomal maternal inheritance and its contribution to male infertility. Our dataset provides an opportunity to explore dominant inheritance models, which have been largely overlooked in the context of male infertility. By studying *de novo* mutations, we have already identified evidence for dominant genetic causes of male infertility (see chapter 5 and Oud et al., 2021). This framework can now be extended to investigate whether the same genes implicated in *de novo* mutations also contribute to infertility through maternal inheritance. However, this will not be the case if these genes also play a role in oogenesis, as maternal inheritance of such variants would not be expected.

Mitochondrial DNA (mtDNA) is also maternally inherited. Recent evidence suggests this is ensured by what may be a quality-control step during spermatogenesis, where the protein mitochondrial transcription factor A (TFAM) is redirected from the mitochondria to the nucleus, leading to the elimination of paternal mtDNA (Lee et al., 2023). Consequently, the presence of mtDNA in ejaculated sperm, rather than mutations within it, is proposed as a potential biomarker for failed maturation and male infertility (Lee et al., 2023). This hypothesis aligns with research linking an increased mtDNA copy number to infertility, a condition potentially caused by the dysfunctional expression of nuclear-encoded regulators such as DNA polymerase γ (*POLG*) (Vahedi Raad et al., 2024). While these findings highlight a crucial interplay between the two genomes, an analysis of SVs in mtDNA was not performed in this study, as our focus was on the nuclear genome.

In this chapter, I investigated maternally inherited SVs in 216 male infertility trios. Unlike *de novo* SVs, which are extremely rare, approximately half of all SVs identified in each proband are expected to be maternally inherited. Prioritising potentially pathogenic SVs among these inherited variants poses a significant challenge due to the vast number of benign CNVs present in the human genome (Collins et al., 2020) . This chapter aims to explore the differences between maternally and paternally inherited SVs and to identify rare, potentially pathogenic maternally inherited SVs that may contribute to male infertility.

6.2 Aims

This chapter aims to

- Investigate the differences between maternally and paternally inherited rare SVs identified in the cohort of 216 patient-parent trios.
- Prioritise and interpret the likely pathogenicity of rare maternally inherited SVs.
- Identify novel candidate dominant, X- or Y-linked male infertility genes.

6.3 Results

In this chapter, a cohort of 216 patient-parent trios diagnosed with azoospermia and severe oligozoospermia was used to investigate the role of MI SVs in a specific form of male infertility. As detailed in Chapter 3, two different tools, CNVRobot and dysgu-sv were combined to identify SVs in this cohort. A general overview of SVs was provided in Chapter 4 followed by an analysis of *de novo* SVs in Chapter 5. In this chapter, we focus on rare MI SVs, where we investigated dominant inheritance patterns, as was done in the chapter 5.

6.3.1 Overview of Rare Autosomal Inherited SVs

Firstly, the number of rare paternally inherited (PI) and MI SVs were compared. To perform this comparison, *de novo* SVs, homozygous SVs, SVs for which inheritance was unclear, SVs on sex chromosomes, SVs with high population frequency (>0.01), SVs seen more than 10 times in the entire cohort and SVs with low confidence[10] were excluded. The remaining rare autosomal SVs with high confidence numbered 3,587, constituting 2% of all SVs detected. Of these, 1855 were MI SVs and 1732 were PI SVs. SVs that are rare in the population and span a significant part of the genome are frequently linked to disease phenotypes (Li et al., 2020). Therefore, we examined the size distribution of SVs across all categories for both groups (Figure 6.1). To assess whether a significant difference exists between MI and PI SVs across

[10] CNVs which are labelled as mappability low and control high noise by CNVRobot were considered as low confidence SVs.

these categories, a Pearson's Chi-squared test of independence was conducted[11], revealing a statistically significant association (p-value=0.04138). Residual analysis further indicated that this difference was primarily driven by an over-representation of <2.5 kb duplications in the PI group and >100 kb duplications in the MI group. These disparities are likely due to technical limitations, particularly in the identification of duplications, rather than reflecting underlying biological differences.

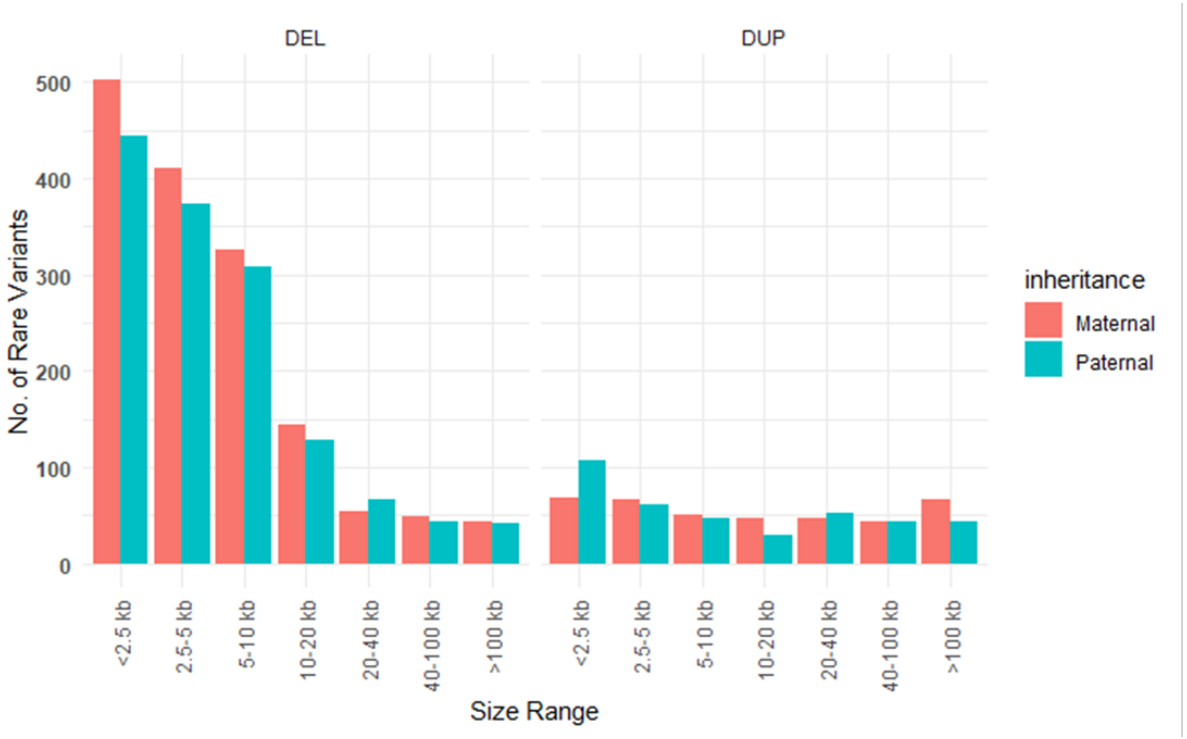


Figure 6.1. **The number of rare SVs on autosomes was categorised by size in 216 probands** with azoospermia or severe oligozoospermia. A Chi-squared test of independence revealed a statistically significant difference between maternally and paternally inherited SVs across these categories (p-value=0.04138). DEL, deletion; DUP, duplication.

6.3.2 Autosomal Maternally Inherited SVs

Inherited SVs are significantly more common than *de novo* SVs, making the prioritisation of likely pathogenic variants much more challenging. The MI SVs that might have a mild or no impact on female fertility but a significant effect on male fertility were focused on in this chapter. Such SVs could be passed to offspring without undergoing negative selection in

[11] The chi-square test was the appropriate because it evaluates whether there is an association between two categorical variables-in this case, the type of inheritance (MI vs. PI) and the category of SVs. By analysing the observed frequencies in each category, the chi-square test can determine if the distribution of SVs significantly differs between MI and PI groups.

mothers. There were 3,587 autosomal rare SVs with high confidence. Of these, 1,855 were MI SVs.

To investigate known disease genes, rare MI SVs that encompass at least one exon in the KCMIG with a score greater than 1 resulted in 5 SVs, and further filtration using a DOMINO (Quinodoz et al., 2017) score greater than 0.5 and a pLI score exceeding 0.9 narrowed it down to 1 SV (Table 6.1). Interestingly, following this filtration process for rare PI CNVs, no CNVs remained (Figure 6.2).

To reveal new candidate genes, rare MI CNVs larger than 5kb were filtered, leaving 866 CNVs, of which 67 encompassed at least one exon of genes with a pLI score exceeding 0.9, and 17 CNVs (3 deletions and 14 duplications) were absent in all paternal samples (Table 6.2 for Deletions, Table 6.3 for Duplications) (Figure 6.2). The second filtration step was applied to SVs inherited from the father, without filtering based on their presence in the paternal samples. However, no notable SVs associated with infertility were found. Consequently, further analysis was not conducted, as it is beyond the scope of this study.

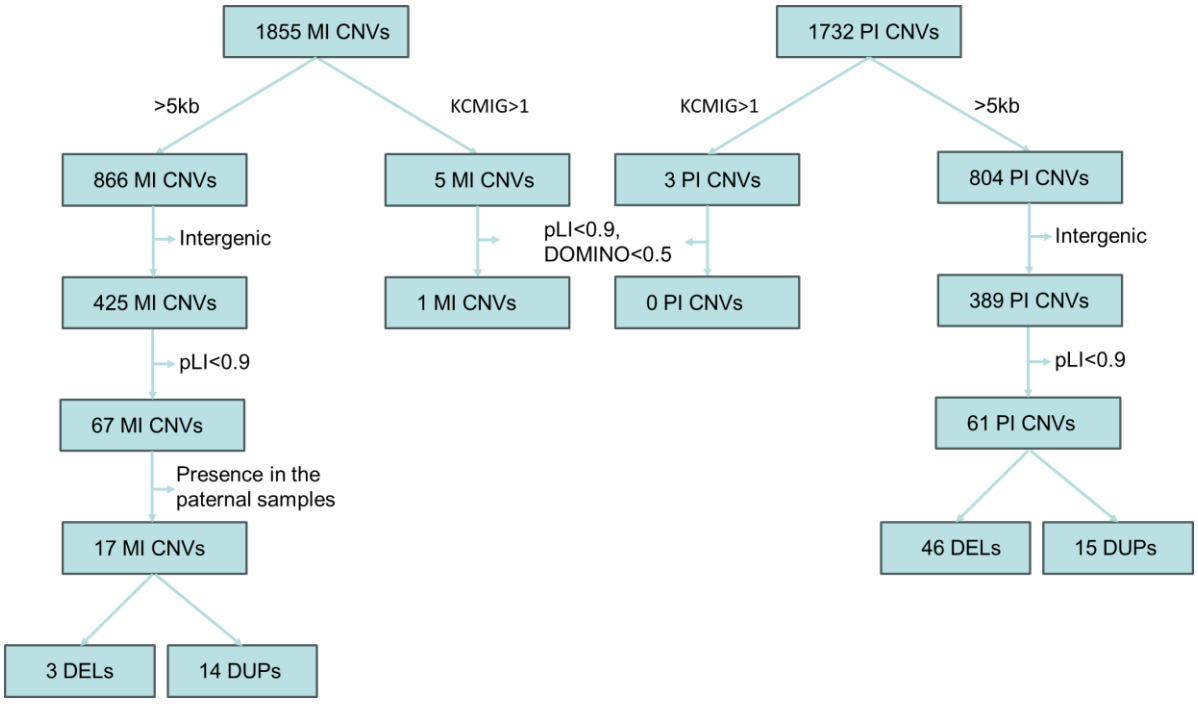


Figure 6.2. **The prioritisation steps for high-confidence inherited autosomal rare CNVs.** Both MI and PI CNVs were prioritised in two main steps: the first based on known and candidate genes, and the second employing an unbiased approach based on CNV size and constraint scores. KCMIG: Known and Candidate Male Infertility Genes, MI: Maternally Inherited, PI: Paternally Inherited, DEL: Deletion, DUP: Duplications

Table 6.1. Proband, genomic locations, size, and genes involved in the rare MI CNVs encompassing at least one exon in the KCMIG with >1 score, with a pLI score > 0.9 and a DOMINO score > 0.5 identified in the trio cohort.

Proband	Type	Genomic Location (GRCh38)	Size	Genes	Genes (pLI > 0.9, DOMINO > 0.5)
NIJ_MI_00347P	DUP	chr6:42680797-42713685	33kb	*UBR2 - PRPH2*	*UBR2*

A previously unreported MI tandem duplication found in chromosome 6 of patient NIJ_MI_00347P with extreme oligozoospermia encompasses the last 5 exons of the *UBR2* gene (pLI=1). From population genetic data, the observed/expected (o/e) fraction for *UBR2's* loss-of-function (LoF) variants is 0.39, with LOEUF of 0.47 (<0.6, gnomAD, 4.1.0). The exact size of the duplication was determined as 32,888 bp using split and discordant reads and the duplication was shown to be inherited from mother (Figure 6.3A-B). Although usually the presence of a breakpoint within the gene indicates that the gene structure is disrupted, when this tandem duplication was depicted (Figure 6.3C), it was observed that the duplication created a fusion gene without a promoter with the other affected gene *PRPH2* and that the *UBR2* structure remained intact. However, according to the ENCODE database, there are 2 distal enhancers (EH38E2467914 and EH38E2467917) in human testicular tissue within the duplicated region. In conclusion, although this tandem duplication did not disrupt the structure of the *UBR2* and *PRPH2* genes, and the fusion protein is predicted not to be expressed, the gene may still be affected by the duplication of distal enhancers within the duplicated region and potential changes in the local 3D genome structure.

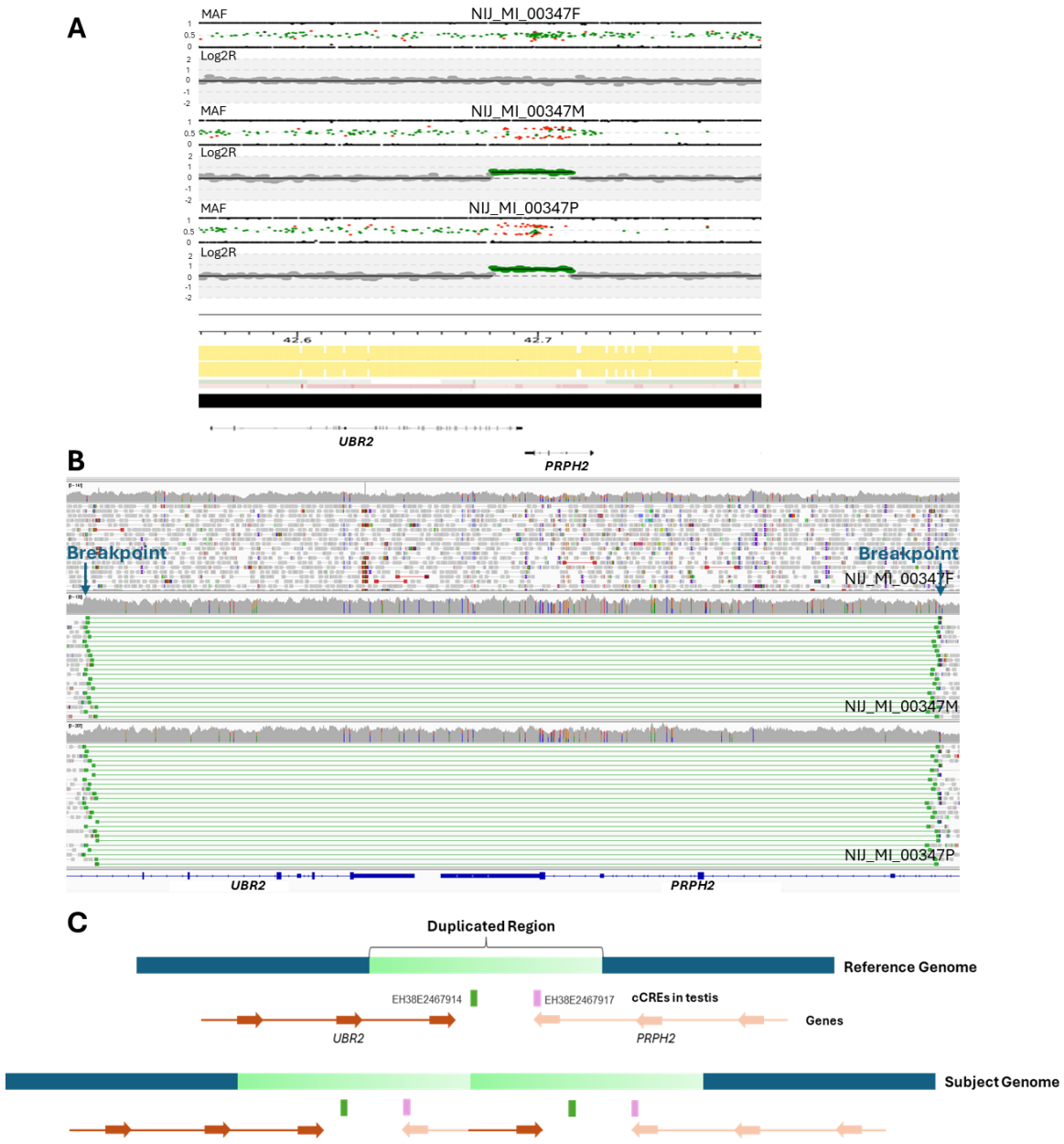


Figure 6.3. A. CNVRobot plot of the heterozygous MI tandem duplication detected in patient NIJ_MI_00347P. The duplication is present in both the mother and the proband. For both, MAF values deviate from the standard 0; instead, intermediate values are observed, indicating additional copy. Also, an expected log2ratio value of approximately 0.58 indicates a heterozygous duplication. **B.** IGV plot of the heterozygous MI tandem duplication identified in NIJ_MI_00347P. A sharp increase in read depth and split reads at the breakpoints reveals the exact breakpoints, with discordant reads in green representing the tandem duplication. It can also be seen that the breakpoints are in the intronic regions, with the last 5 exons of *UBR2* and the last 4 exons of *PRPH2* within the duplication. **C.** Depiction of the consequences of the tandem duplication on the genes. The duplication created a fusion gene without a promoter, involving the other affected gene, *PRPH2*, while the structure of *UBR2* remained intact. According to the ENCODE database, there are two distal enhancers (EH38E2467914 and EH38E2467917) in human testicular tissue within the duplicated region. (cCREs=candidate cis regulatory elements)

Table 6.2. Proband, genomic locations, size, and genes involved in the rare MI deletions >5kb identified in the trio cohort, encompassing at least one exon in the genes with a pLI score > 0.9 and not seen in any paternal genome.

Proband	Genomic Location (GRCh38)	Size	Genes	Genes with a pLI score > 0.9
NIJ_MI_00352P	chr14:104320302-105016836	697kb	*ADSS1 - AHNAK2 - AKT1 - C14orf180 - CDCA4 - CEP170B - CLBA1 - INF2 - PLD4 - SIVA1 - TMEM179 - ZBTB42*	*AKT1 - INF2 - CEP170B*
NIJ_MI_01564P	chr7:137454529-137463662	9kb	*DGKI*	*DGKI*
NIJ_MI_00151P	chr1:201856327-201862350	6kb	*IPO9*	*IPO9*

After the second filtration step to identify new candidate genes, three deletions remained, all of which were rare MI CNVs larger than 5 kb, encompassing at least one exon of the affected genes, with a pLI score exceeding 0.9, and absent in all paternal samples. (Table 6.2). All three deletions identified were unreported in the population databases. The 9kb deletion on chromosome 7 identified in patient NIJ_MI_01564P removes exon 28 of the *DGKI* gene (pLI=1) in one allele. Based on data from the Protein Atlas database, *DGKI* exhibits high expression levels in brain and endocrine tissues, whereas its expression in the testis is notably low. No abnormality was found in tested hormones in the patient (FSH and Inhibin B). This gene belongs to the type IV subfamily of diacylglycerol kinases. The exact function of the enzyme encoded by this gene remains unknown. Also, the *DGKZ* gene from the same family has been associated with increased immune responses against viruses and cancer, but not with spermatogenesis or infertility (Singh & Kambayashi, 2016). Therefore, this deletion does not seem to be related to the phenotype of the patient.

The largest deletion identified was detected on chromosome 14 of patient NIJ_MI_00352P with NOA. The SV was 697kb encompassing 12 protein coding genes. Three of which have high pLI score, *AKT1* (pLI = 0.98), *CEP170B* (pLI= 1) and *INF2* (pLI = 0.97) indicating likely intolerance to loss-of-function mutation (Figure 6.4). According to the Protein Atlas database, only *AKT1* has high expression at the protein level in the testis. Additionally, only intronic deletions have been reported for *AKT1* and *INF2* genes in population databases, whereas *CEP170B* is affected by deletions involving both introns and exons in 40 samples from the DGV Gold Standard. Also, the roles of *INF2* and *CEP170B* in humans have not yet been fully understood.

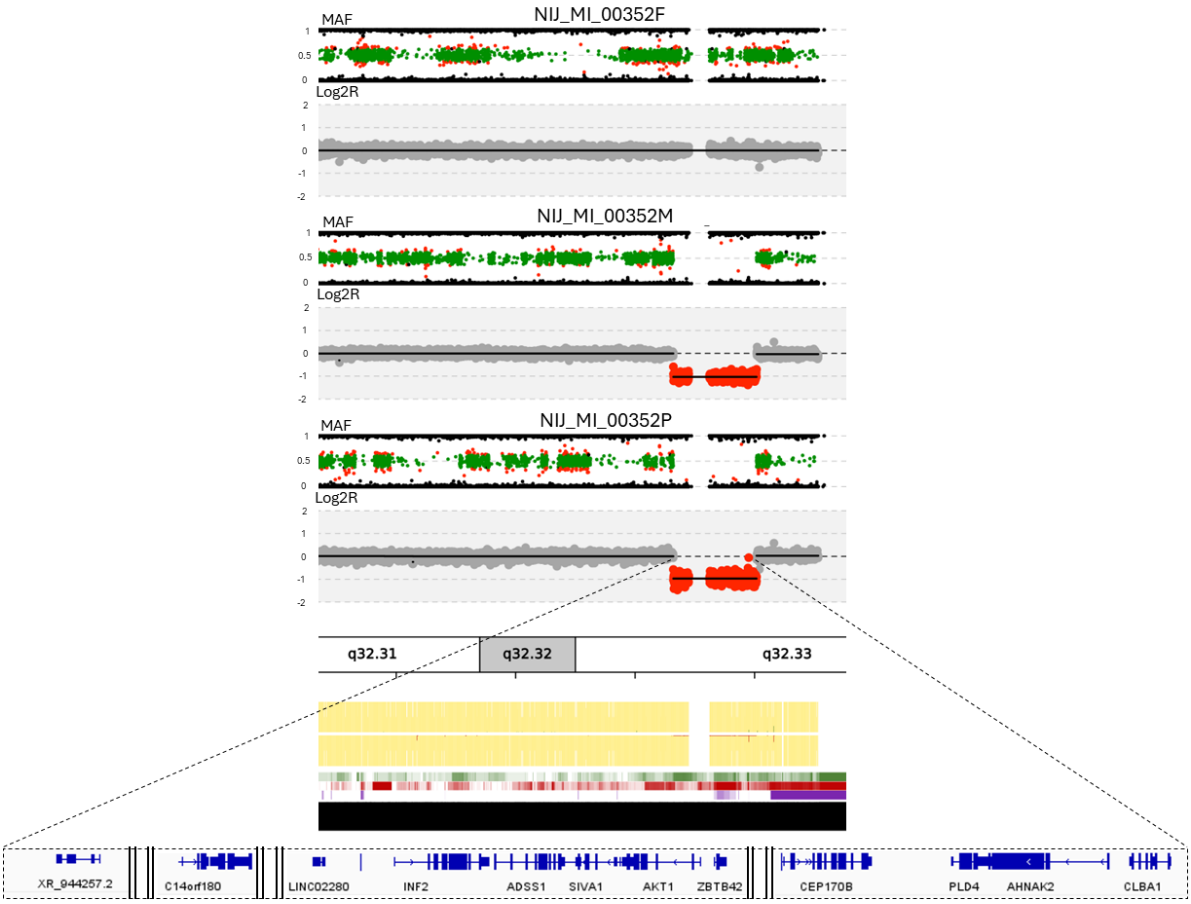


Figure 6.4. CNVRobot plot of the heterozygous MI deletion detected in patient NIJ_MI_00352P. The deletion removed 12 protein coding genes in one allele (Regions without gene removed and labelled with two consecutive lines "||" in gene track). The deletion is present in both the mother and the proband. For both, no heterozygous SNP was observed (MAF=0.5) within the deleted region. Also, a log2R value of -1 indicates heterozygous deletion. The blank region within the deletion is because of masked regions at the beginning of analysis due to complexity of genome.

The other deletion detected on chromosome 1 of patient NIJ_MI_00151P with NOA, omitting exon 11, exon 12 and exon 13 of *IPO9* (pLI=1) gene away in the one allele (Figure 6.5). From population genetic data, the observed/expected (o/e) fraction for *IPO9's* loss-of-function (LoF) variants is 0.22, with LOEUF of 0.31[12] (<0.6, gnomAD, 4.1.0). No deletions were reported in population databases in this region of the genome. The *IPO9* gene is expressed in all tissues almost equally including testis, according to Protein Atlas Database. GO annotations associated with this gene indicate that *IPO9* gene is involved in binding, obsolete protein transporter activity and protein import into nucleus. The improper regulation of importins from the same family was linked to various disorders, including infertility by disrupting miosis shown in mice (Hu et al., 2010; Navarrete-López et al., 2023).

[12] This LOEUF score indicates strong evolutionary constraint against loss-of-function variants in this gene, as evidenced by the significantly lower number of observed variants compared to expected (o/e = 0.22). The LOEUF score of 0.31 falls well below the suggested pathogenicity threshold of 0.6.

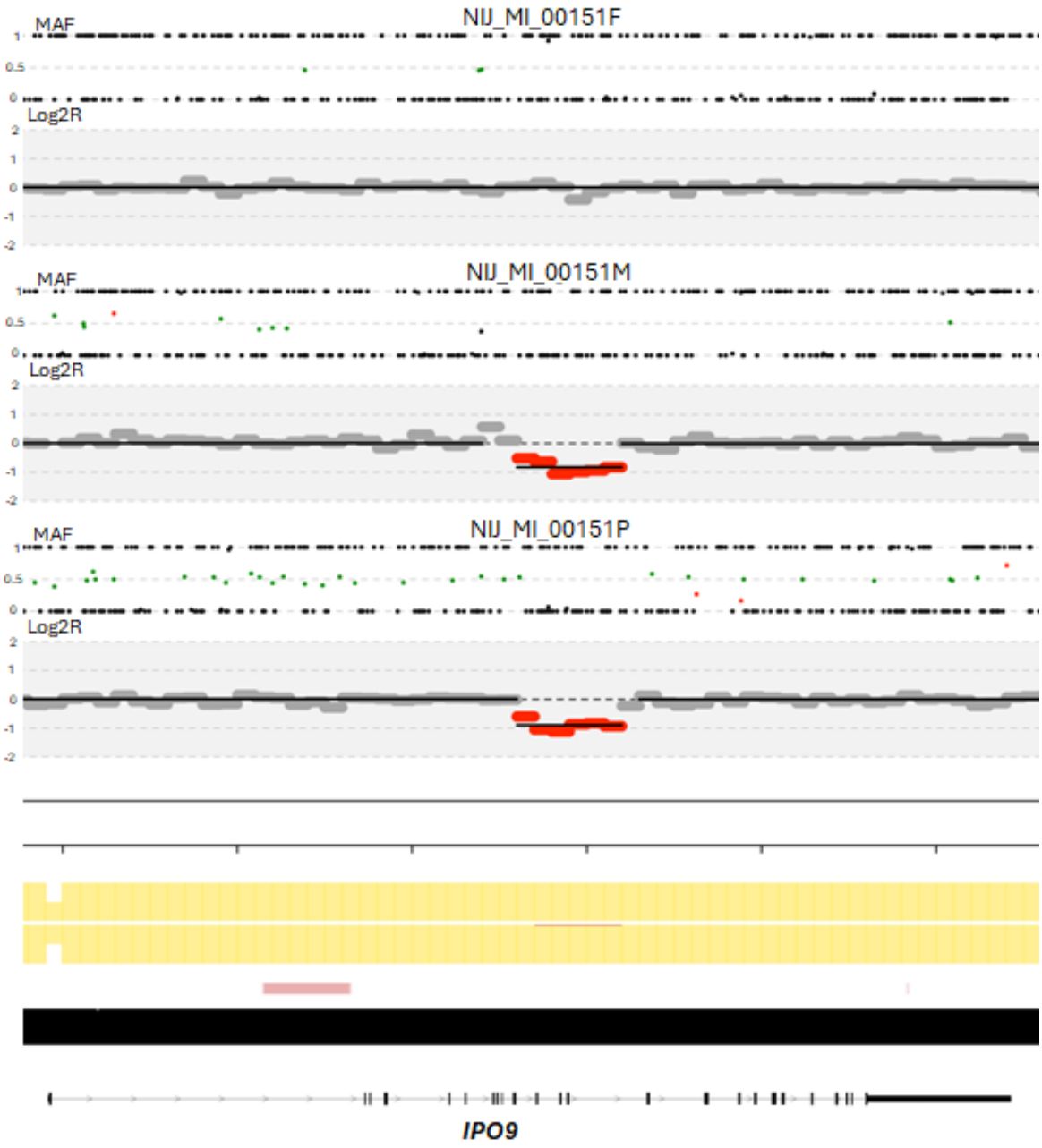


Figure 6.5. CNVRobot plot of the heterozygous MI deletion detected in patient NIJ_MI_00151P. The deletion removed 3 exons of *IPO9* gene in one allele. The deletion is present in both the mother and the proband. In proband, no heterozygous SNP was observed (MAF=0.5) within the deleted region. Also, a log2R value of -1 indicates heterozygous deletion.

Table 6.3. Proband, genomic locations, size, and genes involved in the rare MI duplications >5kb identified in the trio cohort, encompassing at least one exon in the genes with a pLI score > 0.9 and not seen in any paternal genome.

Proband	Genomic Location (GRCh38)	Size	Genes	Genes with a pLI score > 0.9	Genes with a pLI score > 0.9 disrupted by the CNV breakpoints
NIJ_MI_01487P	chr12:2136137-3499334	1.4Mb	*CACNA1C - CACNA1C - CACNA1C - CACNA1C - CACNA1C - CACNA1C - - CACNA1C-IT3 - FKBP4 - FOXM1 - ITFG2 - ITFG2-AS1 - LINC02417 - LOC100128253 - NRIP2 - PRMT8 - RHNO1 - TEAD4 - TEX52 - THCAT155 - TSPAN9 - TULP3*	*CACNA1C - PRMT8*	*CACNA1C - PRMT8*
NIJ_MI_01067P	chr22:23515803-24670975	1.1Mb	*ADORA2A - ADORA2A-AS1 - BCRP3 - C22orf15 - CABIN1 - CHCHD10 - DDT - DDTL - DERL3 - DRICH1 - GGT1 - GGT5 - GSTT2 - GSTT2B - GSTT4 - GSTTP2 - GUCD1 - GUSBP11 - IGLL1 - LRRC75B - MIF - MIF-AS1 - MMP11 - POM121L10P - POM121L9P - RGL4 - SLC2A11 - SMARCB1*	*SMARCB1*	-
NIJ_MI_00877P	chr1:77360689-77828694	468kb	*AK5 - MIGA1 - USP33 - ZZZ3*	*ZZZ3*	-
NIJ_MI_01697P	chr10:12477542-12842653	365kb	*CAMK1D*	*CAMK1D*	*CAMK1D*
NIJ_MI_02329P	chr7:70790547-71122651	341kb	*AUTS2*	*AUTS2*	*AUTS2*
NCL_MI_0105P	chr15:85214218-85601476	270kb	*AKAP13 - GOLGA6L3 - MIR7706*	*AKAP13*	*AKAP13*
NIJ_MI_01745P	chr20:56336516-56559903	224kb	*AURKA - CASS4 - CSTF1 - FAM209A - FAM209B - FAM210B - GCNT7 - RTF2*	*CSTF1*	-
NIJ_MI_01247P	chr11:36415088-36615333	200kb	*IFTAP - PRR5L - RAG1 - RAG2 - TRAF6*	*TRAF6*	-
NIJ_MI_01724P	chr10:88929859-89065015	136kb	*FAS - ACTA2*	*FAS - ACTA2*	-
NIJ_MI_01490P	chr15:90301192-90410396	109kb	*GABARAPL3 - IQGAP1 - ZNF774*	*IQGAP1*	*IQGAP1*
NIJ_MI_02145P	chr17:47917122-47997565	80kb	*CDK5RAP3 - PNPO - PRR15L - SP2 - SP2-AS1*	*SP2*	*SP2*
NIJ_MI_02251P	chr2:206433774-206489715	56kb	*ADAM23*	*ADAM23*	*ADAM23*
NIJ_MI_00960P	chr17:2021614-2047665	26kb	*DPH1 - OVCA2 - RTN4RL1*	*RTN4RL1*	*RTN4RL1*
NIJ_MI_02365P	chr19:18524752-18535724	12kb	*FKBP8*	*FKBP8*	*FKBP8*

The number of rare large duplications was higher than the number of deletions with similar characteristics, 14 in total. The detection and interpretation of duplications in WGS data is more complex than it is for deletions (Abel and Duncavage, 2013). The signatures in WGS data also allowed us to detect the type of duplication, whether it is tandem, dispersed, or inverted duplication. Here, 13 duplications were identified as tandem except for 1 inverted duplication

in patient NIJ_MI_02329P (Figure 6.6). When the region was examined in detail, it was noted that it was a complex SV which involved two inverted duplications and an inversion. The complex SV encompasses the last exon of *AUTS2* gene, which has a pLI score of 1. However, when this inverted duplication was depicted (Figure 6.6D), the *AUTS2* gene remained intact. Additionally, similar size duplications have been reported in population databases, including in males (DGV Gold= 6 in 13,628; gnomAD-SV= 12 in 126,092 with 4 males). Moreover, *AUTS2* gene is associated with Intellectual developmental disorder, autosomal dominant 26 (OMIM: 607270). Stimulatingly, this complex SV is thought to not have an effect on the patient's phenotype and no further analysis was carried out for this duplication.

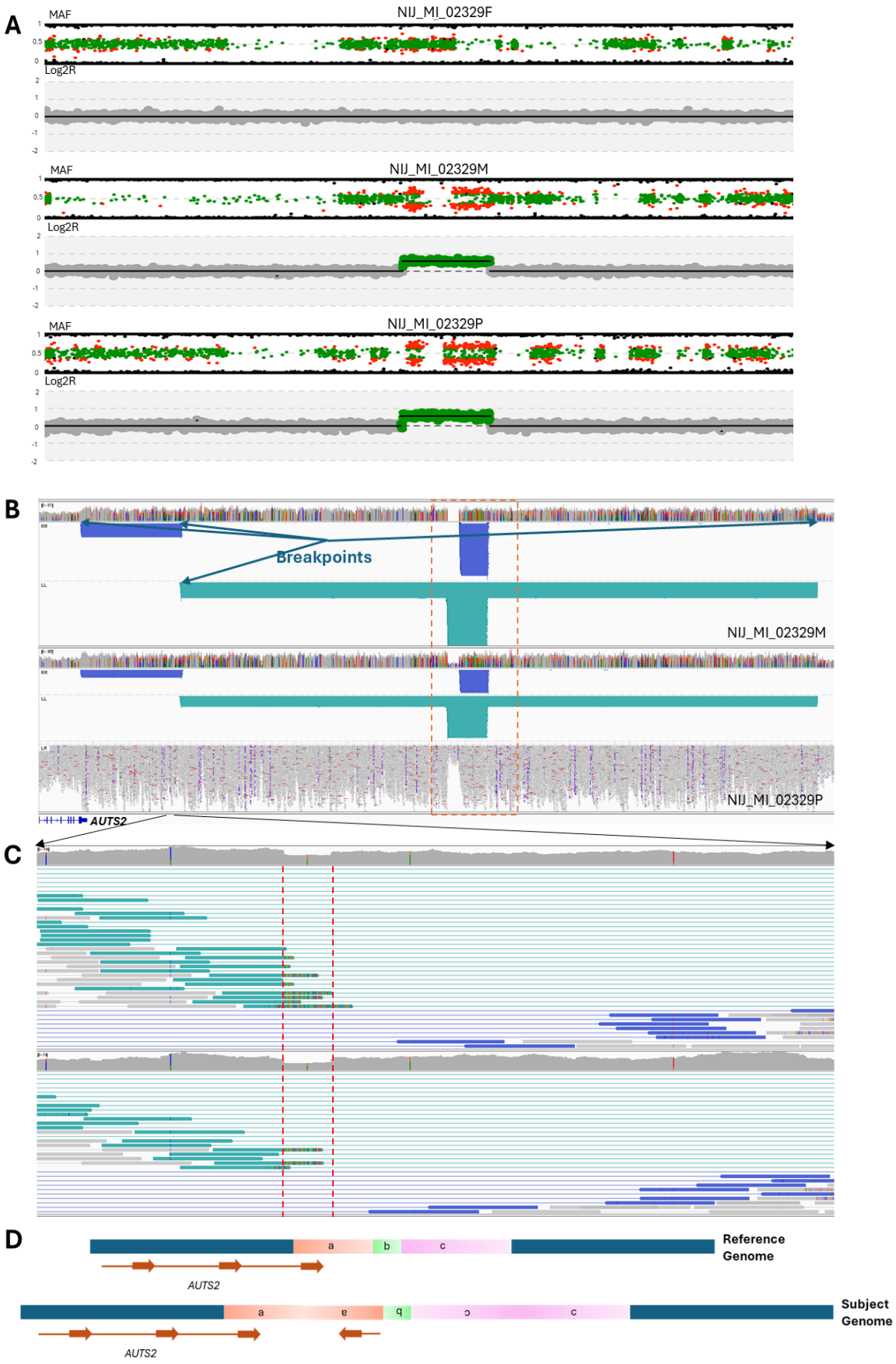

A
MAF
NIJ_MI_02329F
Log2R
MAF
NIJ_MI_02329M
Log2R
MAF
NIJ_MI_02329P
Log2R
B
Breakpoints
NIJ_MI_02329M
NIJ_MI_02329P
AUTS2
C
D
a
b
c
Reference Genome
AUTS2
Subject Genome
AUTS2

Figure 6.6. A. CNVRobot plot of the 341kb heterozygous MI complex SV identified in patient NIJ_MI_02329P. It is identified as a duplication because CNVRobot only considers the read-depth signature. The SV is present in both the mother and the proband. For both, MAF values deviate from the standard position of 0; instead, intermediate values are observed, indicating additional copy. Also, an expected log2ratio value of approximately 0.58 indicates a heterozygous duplication **B.** IGV plot of the heterozygous MI complex SV identified in NIJ_MI_02329P. A sharp increase in read depth and split reads at the breakpoints reveals four breakpoints, with discordant reads in green and dark blue representing the inversions. There are two inverted duplications with a 64 bp inversion in between (**C,** highlighted in red dashes). Additionally, there is a homozygous inverted deletion in the mother, which is in a heterozygous state in the proband within the right inverted duplicated region (highlighted with an orange dashed rectangle). **D.** Depiction of the consequences of the complex SV on the genes. The SV copies a portion of the *AUTS2* gene in the reverse direction, while the structure of *AUTS2* remains intact.

Studies have shown that loss-of-function intolerant genes may also be sensitive to dosage increases (Collins et al., 2020). Genes with high pLI scores may be intolerant to dosage increases when fully duplicated, or their genomic sequence could be disrupted if the duplication breakpoints occur within their coding regions (Collins et al., 2022). Therefore, the identified rare large duplications with high pLI scores were prioritised and interpreted in this study.

The two largest duplications, 1.4 Mb and 1.1 Mb, identified in patients NIJ_MI_01487P and NIJ_MI_01067P, respectively, encompass numerous genes, including those with high pLI scores. However, many duplications that cover the genes with high pLI scores, have also been reported in population databases. The remaining 11 duplications were also individually examined based on partial duplications listed in population databases within the duplicated region, as well as the involvement of potentially affected genes in spermatogenesis and their expression in the testis. For some of those, there is considerable number of partial duplications reported in population databases. For instance, 20 partial duplications are reported in the GnomAD-CNV database, with 8 of these occurring in males, within the 26kb duplication on chromosome 17 of patient NIJ_MI_00960P. This duplication encompasses three genes, including *RTN4RL1*, which has a pLI score of 1, and was disrupted by one of the breakpoints. For some others, a possible function in spermatogenesis could not be found. To exemplify, one of the breakpoints of 80kb duplication on chromosome 17 of patient NIJ_MI_02145P disrupt the *SP2* gene with a pLI score of 1. The gene encodes the Sp2 transcription factor, which binds to GC box promoter elements and selectively activates mRNA

synthesis from genes that contain functional recognition sites. There is currently no evidence of its involvement in spermatogenesis or male infertility, but its role as a transcription factor with specific promoter-binding capabilities suggests it could play a role in fertility regulation. Secondly, one of the breakpoints of the 56kb duplication on chromosome 2 in patient NIJ_MI_02251P with azoospermia disrupts the *ADAM23* gene (pLI=1). The gene is highly expressed in testis at the protein level according to Protein Atlas Database. While some ADAM family genes are known to play roles in various reproductive processes (Moreno et al., 2011), *ADAM23* was not found to have a direct connection with the process of spermatogenesis or male infertility in the literature. However, its high testicular expression and membership in the ADAM family make it interesting, suggesting the need for further studies to elucidate its precise role in spermatogenesis and so male fertility.

Following the elaborate investigations conducted in this study, the two duplications appeared to be of interest due to their relation to spermatogenesis, encompassing the *FAS* (pLI=1, ClinGen Triplosensitivity Score=0, pTriplo=0.33) and *ACTA2* (pLI=0.93, ClinGen Triplosensitivity Score= 0, pTriplo=0.55) genes in patient NIJ_MI_01724P (Figure 6.7), and the *FKBP8* (pLI=1, ClinGen Triplosensitivity Score is not available, pTriplo=0.89) gene in patient NIJ_MI_02365P (Figure 6.8). For the latter, given that one of the breakpoints of the tandem duplication occurs within the *FKBP8* gene, the gene structure remains intact but is partially duplicated (Figure 6.8 B). However, there is still a possibility of disruption of the gene's expression due to the structural change in the downstream region. All these 3 genes are predicted to be likely dominant by DOMINO. The *FAS* gene also plays a key role in spermatogenesis by regulating apoptosis through its interaction with the FAS ligand (*FASL*). Sertoli cells in the testes express *FASL*, which binds to the *FAS* receptor on germ cells, triggering apoptosis (Rajender, 2012). This process is essential for maintaining the balance between germ cells and Sertoli cells, ensuring that only a viable number of germ cells are supported, and defective cells are eliminated (M. Wang & Su, 2018; Sharma et al., 2023). The *ACTA2* gene encodes one of six highly conserved actin proteins, which are involved in cell motility, structure, integrity, and intercellular signalling. It has been identified as an RNA marker for human testicular peritubular myoid cells through scRNA sequencing and is expressed from birth onwards (Di Persio & Neuhaus, 2023). Studies showed that the loss or dysfunction of myoid cells can impair spermatogenesis by affecting the biochemical environment necessary for germ cell development (Zhou et al., 2019). The *FKBP8* gene encodes FK506-binding protein 8, which

plays a role in maintaining mitochondrial function by regulating mitochondrial dynamics and inhibiting apoptosis. It is closely related to mitophagy, the selective autophagic process that targets damaged or unnecessary mitochondria for degradation. Mitophagy is essential for maintaining mitochondrial quality and energy homeostasis during spermatogenesis, particularly in the processes where mitochondrial dynamics are critical, such as during spermiogenesis. These two duplications and their potential contribution to the phenotypes is later discussed in detail in the discussion of this chapter.

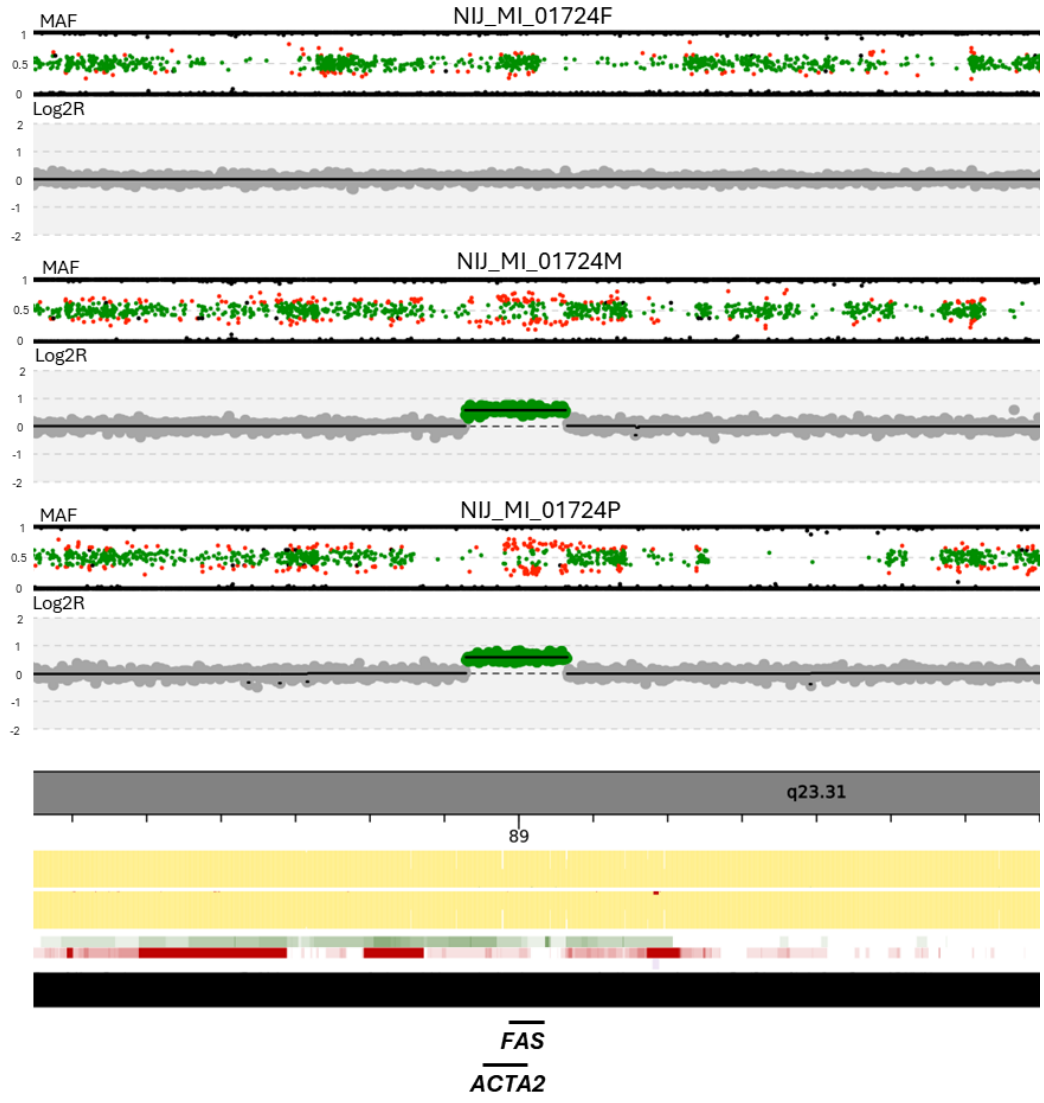


Figure 6.7. CNVRobot plot of the 136kb heterozygous MI tandem duplication identified in patient NIJ_MI_01724P. The duplication is present in both the mother and the proband, encompassing *FAS* and *ACTA2* genes. In both individuals, MAF values deviate from the standard position of 0, 0.5, and 1; instead, intermediate values are observed, indicating additional copy. Also, an expected log2ratio value of approximately 0.58 indicates a heterozygous duplication.

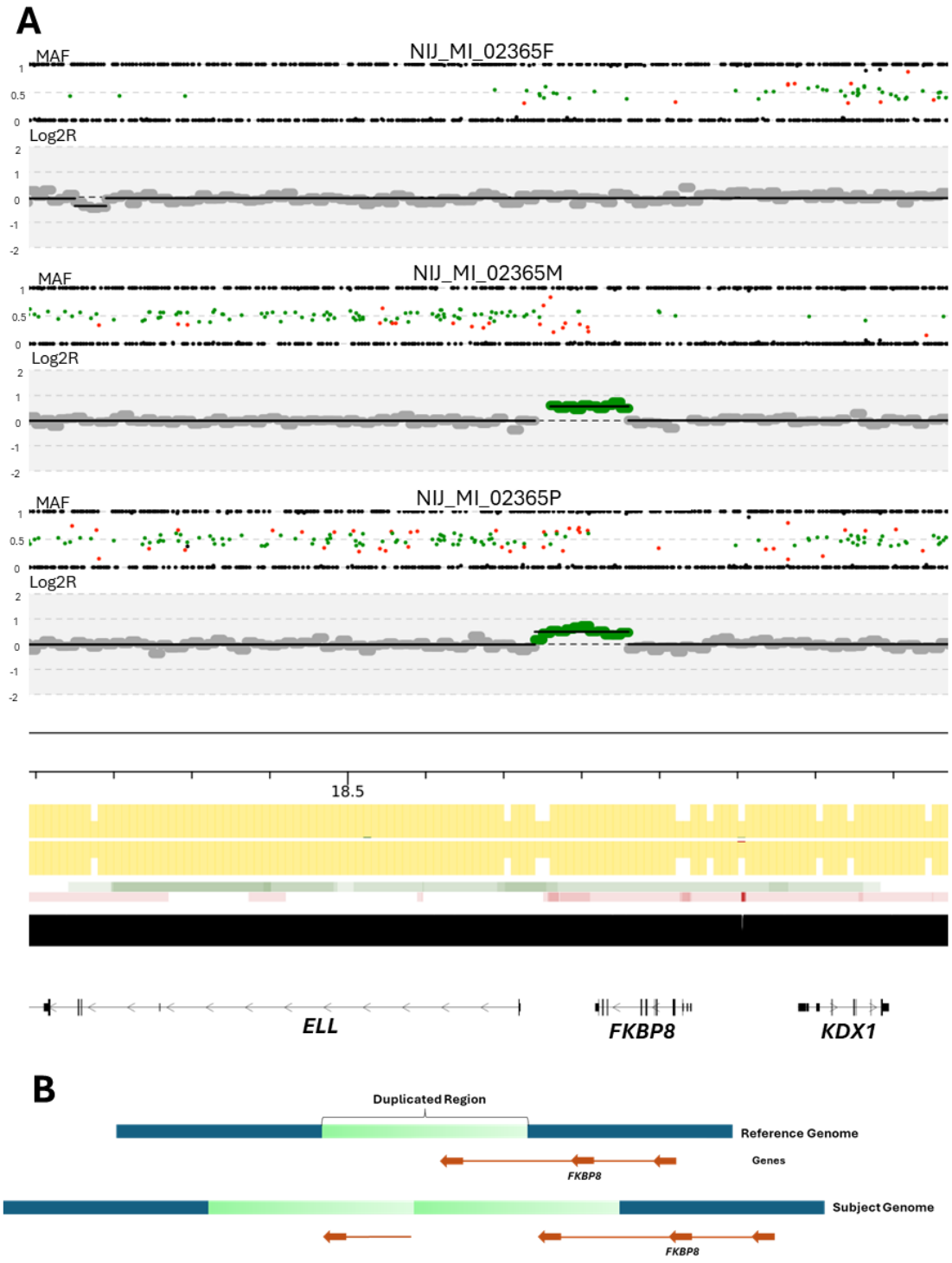


Figure 6.8. CNVRobot plot of the 12kb heterozygous MI tandem duplication identified in patient NIJ_MI_02365P. The duplication is present in both the mother and the proband, encompassing *FKBP8* gene. In both individuals, MAF values deviate from the standard position of 0, 0.5, and 1; instead, intermediate values are observed, indicating additional copy. Also, an expected log2ratio value of approximately 0.58 indicates a heterozygous duplication. **B**. Depiction of the consequence of the tandem duplication on the gene. It can be seen that the structure of *FKBP8* remained intact.

6.3.2.1 Maternally Inherited Inversions

The four inversions introduced in Chapter 4, Table 4.1, were all maternally inherited. Two of these inversions are located deep in the intronic region of the *ATRNL1* and *TRDN* genes, the other inversion is 6kb in length encompassing the 5' UTR of the *CBFA2T3* gene with a pLI score of 0.05 and was predicted to be recessive by DOMINO (Quinodoz et al., 2017). Further investigations were performed only on the large inversion of 659kb, which encompasses four protein-coding genes (*PRM1, PRM2, PRM3, TNP2*) exclusively expressed in the testes (Figure 6.9). Given the challenges in detecting complex SVs and the high probability of noise, this inversion was validated by PCR on both alleles the intact and inverted at both breakpoints (see chapter 2).

The inversion was identified on chromosome 16 of patient NIJ_MI_01490P, who showed severe oligoasthenoteratozoospermia (OAT), and was not observed in GnomAD-SV. There is no gene with a pLI score greater than 0.9 within the inverted piece of DNA or near this inversion. However, the four protein-coding genes (*PRM1, PRM2, PRM3, TNP*2) and one snoRNA gene (*SNORA482*) located within the inversion, 100kb away from the right breakpoint, as well as the protein-coding gene (*TEKT5*) near the inversion, 38kb away from the left breakpoint, were exclusively expressed in the testis. The inversion could impact gene expressions inside the inversion by altering gene positions, changing their orientation, or possibly modifying the chromatin architecture. Also, an enhancer of the *TEKT5* gene (GH16J010736) is located within the inversion region[13], which relocates the enhancer far from its original position and may therefore affect the transcription of the gene. It should be noted that gene-enhancer interaction was inferred from different tissues; however, when testis tissue data from the ENCODE database was examined, signals were observed in the same region where GeneHancer database defines the enhancer, indicating the presence of an enhancer in testis tissue as well.

[13] The enhancer of the *TEKT5* gene is defined by GeneCards using multiple sources, including EPDnew, ENCODE, and RefSeq.

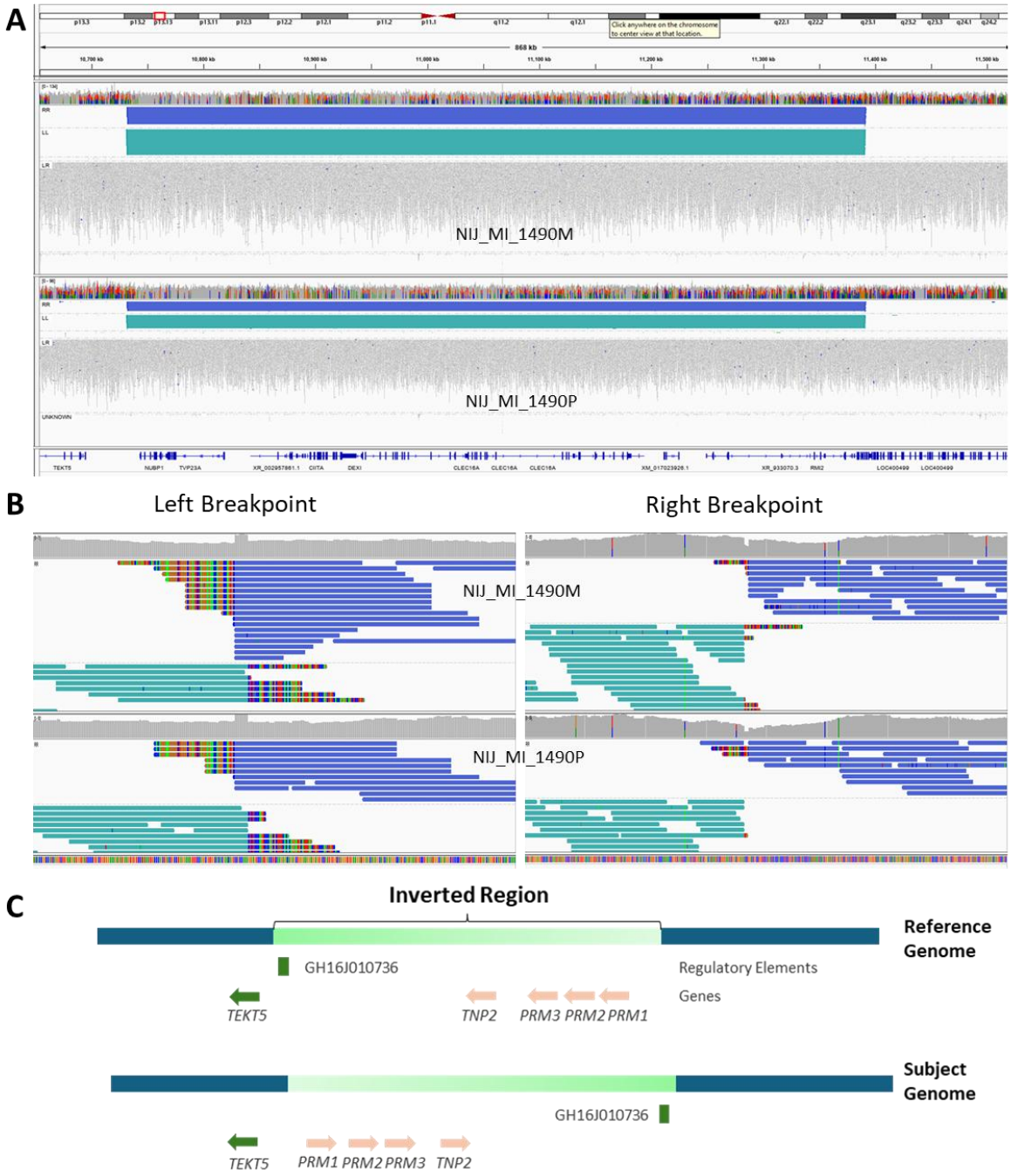


Figure 6.9. A. IGV plot of the 658kb heterozygous MI inversion identified in patient NIJ_MI_02365P. The duplication is present in both the mother and the proband. The green and dark blue lines represent discordant reads in forward-forward and reverse-reverse directions respectively, indicating an inversion. **B.** Split reads reveal the exact breakpoints, with a small deletion on one side and a duplication on the other. **C.** Depiction of the consequences of the inversion. The inversion alters gene positions and changes their orientations (only the genes exclusively expressed in testis are depicted). It also relocates the enhancer of the *TEKT5* gene far from its original position.

6.3.3 Inherited SVs on Chromosome X

The overview of SVs on sex chromosomes were introduced in Chapter 4. As all SVs on chromosome X were maternally inherited, except for one *de novo* deletion, they are described in detail here.

In total, 6,437 SVs were identified on sex chromosomes. Of these, 625 (10%) were rare inherited SVs, with 519 on the X chromosome.

To reveal the possibly pathogenic SVs inherited through X chromosome, two different filtration strategies were applied; (i) To investigate known disease genes on chromosome X, rare MI SVs which encompass at least one exon in the KCMIG with >1 score and not seen in any paternal genome were filtered resulting in no SVs. (ii) To identify new candidate genes, rare MI CNVs larger than 5kb, encompassing at least one exon of KCMIG with a score of ≤1 or genes with a pLI score exceeding 0.9, absent in all paternal samples, were kept. This resulted in 3 duplications, 1 complex SVs (CS) and 1 deletion (Table 6.4).

Table 6.4. Proband, genomic locations, size, and genes involved in the rare MI CNVs larger than 5kb identified on chromosome X in the trio cohort, encompassing at least one exon of KCMIG with a score of ≤1 or genes with a pLI score exceeding 0.9, and absent in any paternal samples. (CS: Complex SV)

Proband	Genomic Location (GRCh38)	Type	Size	Genes	Genes disrupted by the CNV breakpoints
NIJ_MI_00992P	chrX:105890803-106431990	DUP	541kb	*NRK - PWWP3B - SERPINA7*	*NRK*
NIJ_MI_00625P	chrX:16757558-16779385	DUP	22kb	*SYAP1*	*SYAP1*
NCL_MI_0001P	chrX:136487329-136507555	DUP	20kb	*BRS3 - HTATSF1*	*HTATSF1*
NIJ_MI_00151P	chrX:37834387-37847052	CS	13kb	*DYNLT3*	*DYNLT3*
NIJ_MI_01662P	chrX:17375201-17381669	DEL	6kb	*NHS*	*NHS*

The largest duplication identified in the patient NIJ_MI_00992P with severe OAT was reported in GnomAD-SV database, but only in two women (GnomAD-SV ID:DUP_CHRX_2BEEA917, AF= 0.00002085). In ClinVar, this duplication is shown in two individuals: one entry is classified as pathogenic with reported developmental delay (Variation ID: 58665), while another is classified as a VUS also presenting with developmental or morphological phenotypes (Variation ID: 153119). Furthermore, the DECIPHER database includes nine patients (6 males, 3 females) with overlapping duplications who present with a range of severe phenotypes,

including intellectual disability, autism, short stature, and congenital abnormalities such as renal agenesis (DECIPHER Patient IDs: 267826, 252154, 249139, 343974, 340015, 301482, 294850), although these variants are predominantly classified as VUS, with one exception listed as likely benign. In contrast to these database entries, no other significant developmental or morphological phenotypes were reported in the proband as well as in the mother.

One of the breakpoints disrupts *NRK* gene (pLI=0.99) and 2 other genes, *PWWP3B* (pLI=0.48, ClinGen Triplosensitivity Score is not available) and *SERPINA7* (pLI=0, ClinGen Triplosensitivity Score is not available), encompassed by the duplication (Figure 6.10). Given that this is a tandem duplication, the original structure of the *NRK* gene remains intact but is partially duplicated. The other two genes are fully encompassed within the duplicated segment, resulting in an extra copy of each. According to the Protein Atlas database, the *PWWP3B* gene is expressed equally in the ovary and testis, while *SERPINA7* gene is exclusively expressed in the liver and *NRK* gene exclusively in the ovary. The roles of the *PWWP3B* and *NRK* genes are unclear in spermatogenesis. Nevertheless, Tüttelmann et al., 2011,, presented an oligospermic man with a duplication not found in normozoospermic controls across several cases, encompassing the same 3 genes. It is unclear what the functions of these genes are, and it is also unclear what the effect is of the duplication on gene regulation, if any. Interestingly, this duplication was only observed in two females in the control population database, no men have been reported with this duplication. However, in our case and the case reported by Tüttelmann et al., both individuals who have the duplication were diagnosed with oligozoospermia.

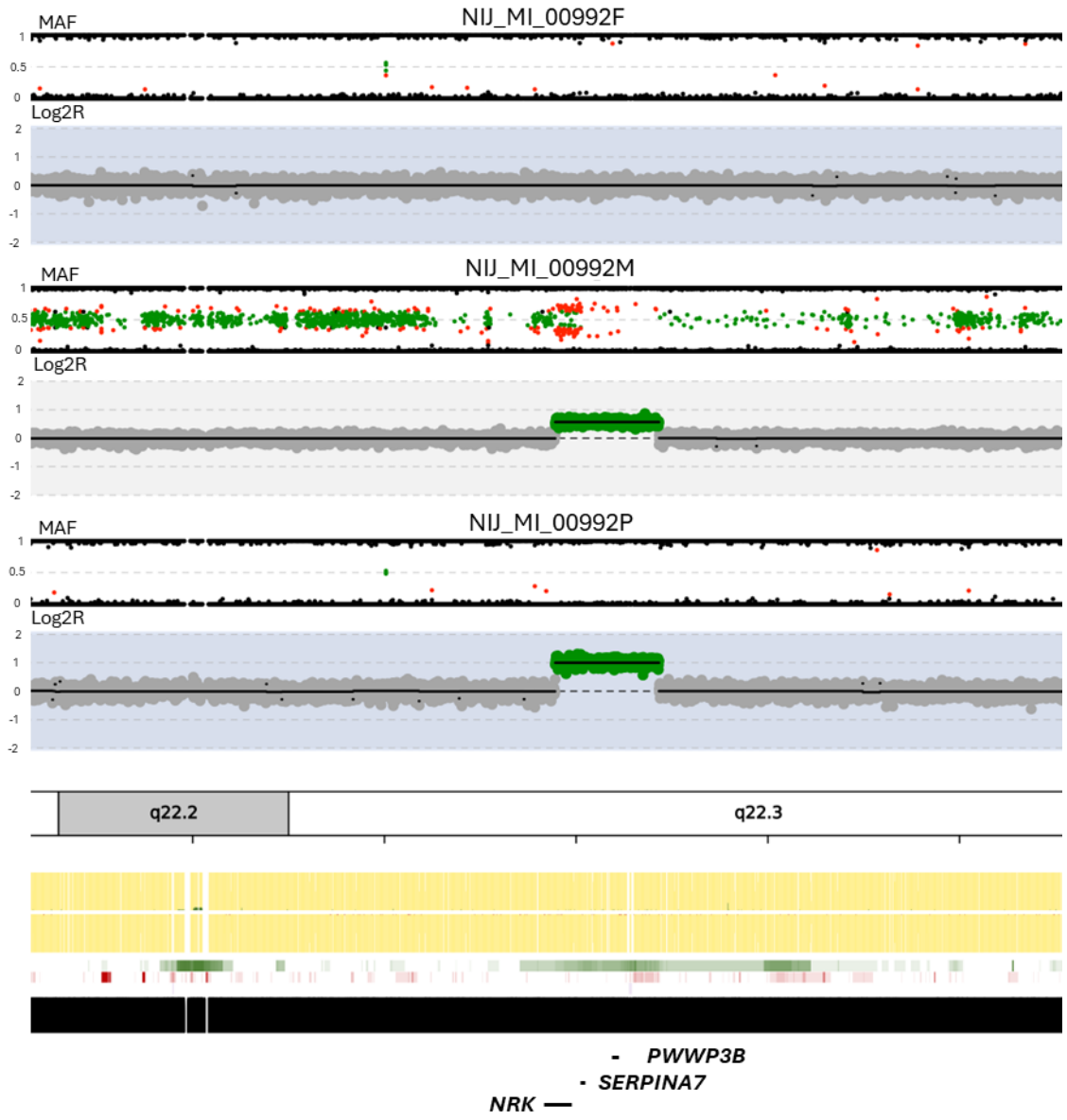


Figure 6.10. CNVRobot plot of the 541kb hemizygous MI tandem duplication identified on chromosome X in patient NIJ_MI_00992P. The duplication is present in both the mother and the proband, encompassing *3* protein coding genes and having a breakpoint in the *NRK* gene. In the proband, the Log2R score is 1 for the duplicated region since only one copy of the X chromosome exists in males. In the mother, an expected log2ratio value of approximately 0.58 indicates a heterozygous duplication.

The other duplication, identified in patient NIJ_MI_00625P with NOA, has one breakpoint that disrupts the *SYAP1* gene (pLI=0.94) (Figure 6.11). The pathology report for this patient indicated a SCO phenotype. The same size duplication was reported in 76 individuals in the GnomAD-SV database, but interestingly again all individuals were women (GnomAD-SV ID: DUP_CHRX_256179F5, AF=0.0007714). According to the Protein Atlas database, the *SYAP1* gene is expressed in all tissues with high expression in testis. The gene encodes Synapse Associated Protein 1 which plays a role in adipocyte differentiation by promoting mTORC2-mediated phosphorylation of *AKT1* at 'Ser-473' after stimulation of the growth factor. The role of the gene in spermatogenesis is unclear. Additionally, Tüttelmann et al., 2011, reported a man with SCOS who had a duplication encompassing the *SYAP1* gene, that was not found in normozoospermic controls across several cases. To sum up, two men with SCOS were identified, both having a duplication encompassing the *SYAP1* gene. Notably, one of these duplications is known to be carried by 74 women in population databases. This suggests that the region may be predisposed to duplications and more commonly carried by women potentially due to negative selection in men.

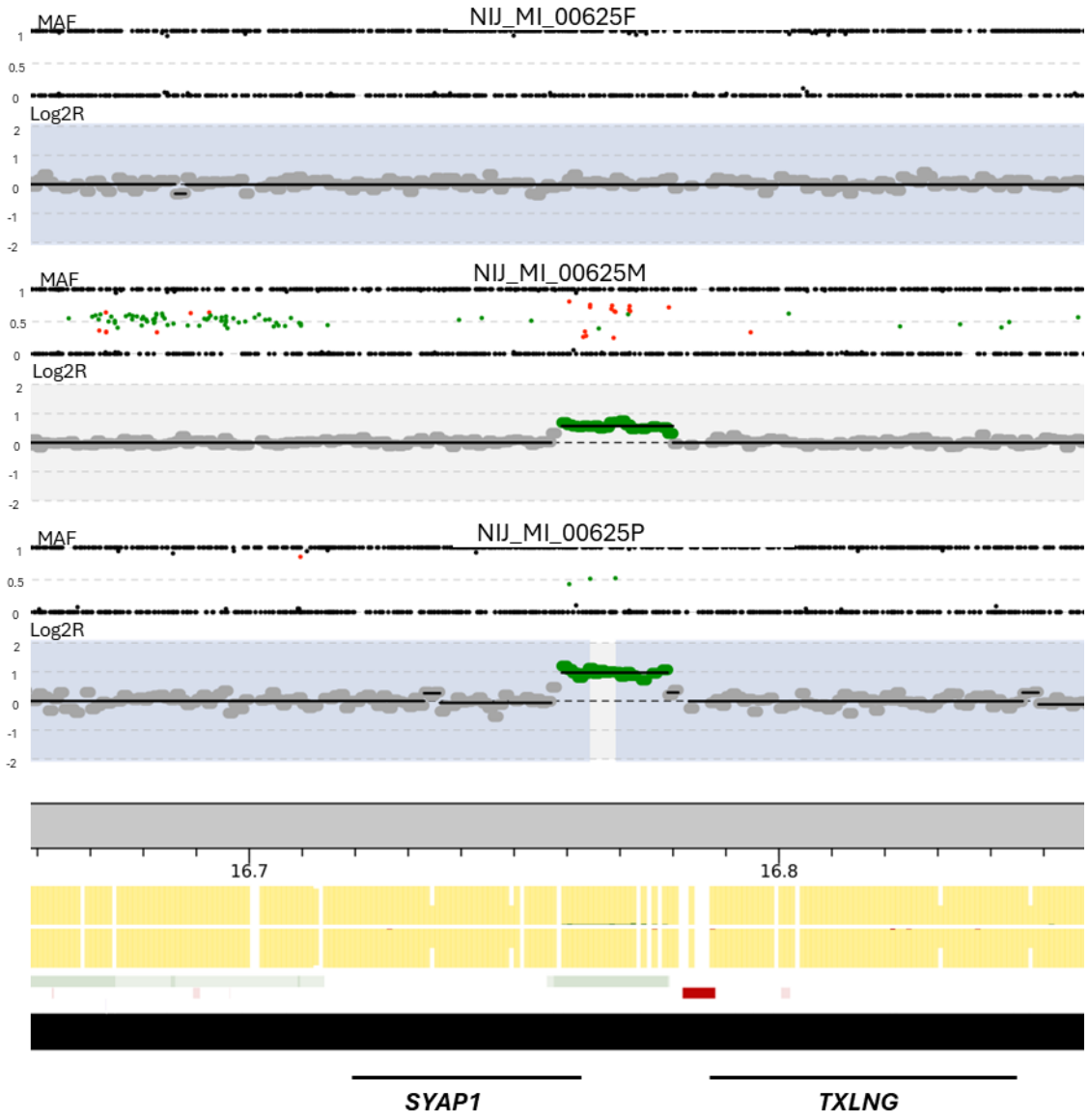


Figure 6.11. CNVRobot plot of the 22kb hemizygous MI tandem duplication identified on chromosome X in patient NIJ_MI_00625P. The duplication is present in both the mother and the proband, having a breakpoint in the *SYAP1* gene. In the proband, the Log2R score is 1 for the duplicated region since only one copy of the X chromosome exists in males. In the mother, an expected log2ratio value of approximately 0.58 indicates a heterozygous duplication.

The third duplication has one breakpoint that disrupts the *HTATSF1* gene (pLI=1) and encompasses the *BRS3* gene (pLI=0.95), detected in patient NCL_MI_0001P with NOA. The same-sized duplication was reported in 51 individuals in the GnomAD-SV database, 6 of whom are men (GnomAD-SV ID: DUP_CHRX_D706F1C5, AF=0.0005322). The *BRS3* gene is exclusively expressed in the epididymis at the RNA level, according to the Protein Atlas Database, and plays a role in sperm cell division, maturation, and function. The *HTATSF1* gene is expressed in all tissues and is involved in RNA and protein binding processes, though its function in spermatogenesis remains unclear. However, proteomic analysis showed that this gene is differentially expressed at the protein level, and it was associated with meiotic arrest in cattle-yak hybrids (Wu et al., 2023).

The previously unreported complex SV (CS), including one duplication and one deletion, was found in patient NIJ_MI_00151P with NOA, where the breakpoints disrupt the *DYNLT3* gene (pLI=0.72) (Figure 6.12). The *DYNLT3* gene encodes Dynein light chain, Tctex-type 3 protein, is a member of the cytoplasmic dynein DYNLT light chain family and was reported to have a potential role in chromosome congression during human mitosis. This gene is expressed in all tissues, according to Protein Atlas Database. Additionally, Huang et al., (2011) demonstrated that *DYNLT3* is essential for chromosome alignment and homologous chromosome segregation in mice.

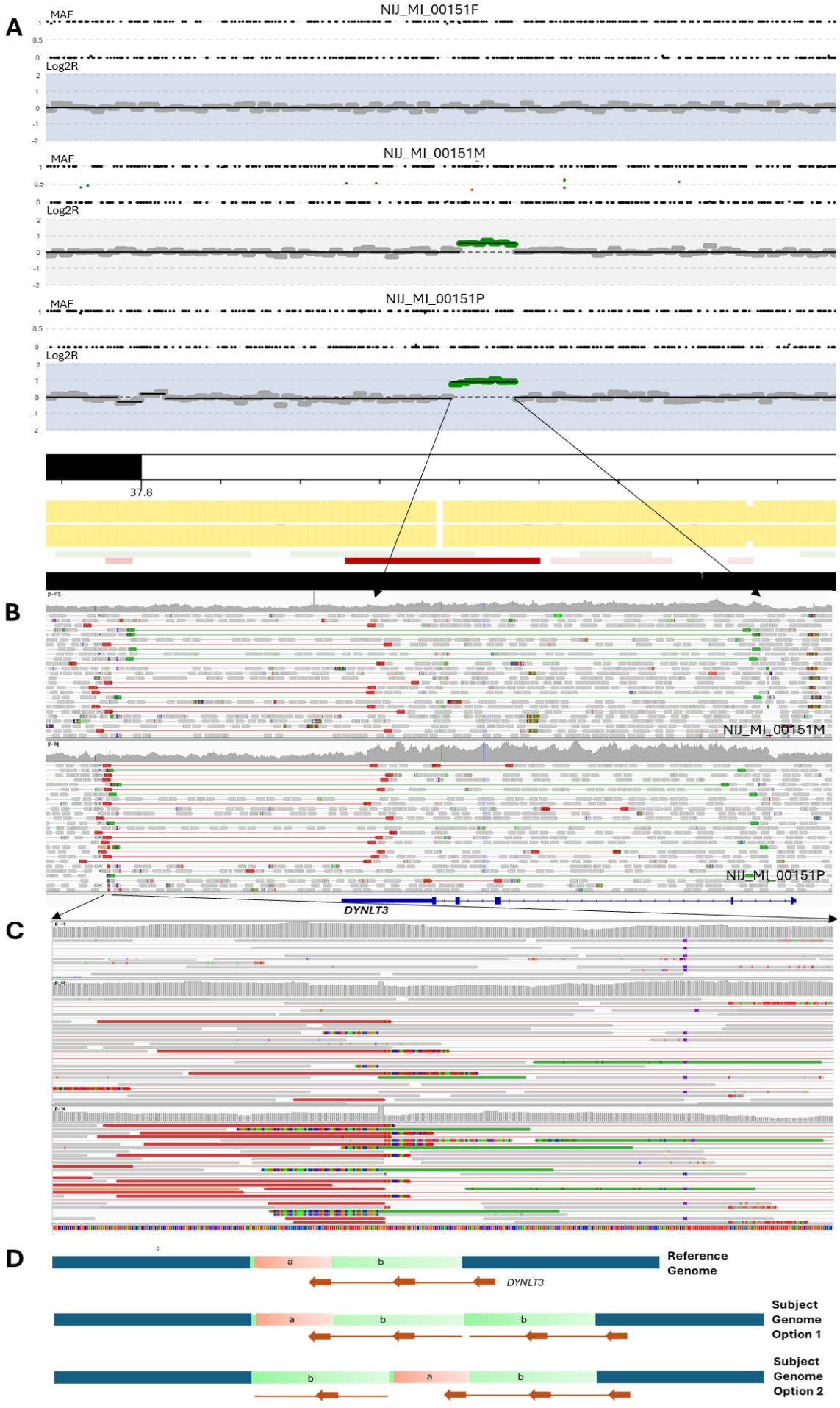

A
NIJ_MI_00151F
MAF
Log2R
NIJ_MI_00151M
MAF
Log2R
NIJ_MI_00151P
MAF
Log2R
37.8
B
NIJ_MI_00151M
NIJ_MI_00151P
DYNLT3
C
D
a
b
Reference Genome
DYNLT3
Subject Genome Option 1
Subject Genome Option 2

Figure 6.12. CNVRobot and IGV plots of the 13kb heterozygous MI complex SV identified in patient NIJ_MI_00151P. A. CNVRobot plot of the 13kb heterozygous MI complex SV. It is identified as a duplication because CNVRobot only considers the read-depth signature. The SV is present in both the mother and the proband. **B.** IGV plot of the heterozygous MI complex SV identified in NIJ_MI_00151P. A sharp increase in read depth and split reads at the breakpoints reveals four breakpoints. There is a tandem duplication with a deletion in it (**C** shows the left breakpoint of the duplication, and it is clearly visible that the deletion begins 3 bp inside the duplicated region.). **D.** Depiction of the consequences of the complex SV on the *DYNLT3* gene. This data does not allow us to determine which of the two copies contains the deleted region, so two possible scenarios are shown. In only one of these scenarios (Option 2), an intact sequence of the gene remains.

The only deletion (6 kb in size) identified on chromosome X in patient NIJ_MI_01662P with azoospermia, previously unreported, removes one copy of the first exon of the *NHS* gene (pLI=1) (Figure 6.13). Additionally, we identified a 29kb deletion in the same patient in close proximity, which was previously unreported. While it does not affect any gene, it may impact the regulation of the *NHS* gene. Although the proximity of these two events might suggest a single complex event (like an inversion deletion), this is unlikely, as our use of read signatures would have enabled us to detect such structural changes. Only intronic deletions have been reported in population databases in this gene. This gene encodes a protein with four conserved nuclear localization signals and is involved in cell differentiation (GO:0030154). Also, the *NHS* gene is associated with X-linked dominant Nance-Horan syndrome (OMIM:302350) and is expressed in all tissues, according to the Protein Atlas Database. Given this inheritance pattern, both the proband and his carrier mother would be expected to show signs of the syndrome. Their clinically unaffected status serves as a key piece of negative evidence, arguing against the deletion being the direct cause of a pathogenic phenotype for this syndrome.

Milunsky et al., (2020) reported a man with NOA and cataract, carrying a deletion of *SCML1-NHS-RAI2* genes. The authors suggested that the *NHS* deletion may explain the cataracts, while the *SCML1* deletion is likely responsible for NOA, because of high expression in testis. Afterwards, Riera-Escamilla et al., 2022, reported additional cases with pathogenic mutations in *SCML1* gene supporting their claim. However, it is possible that either the deletion disrupting the *NHS* gene or the nearby 29 kb deletion could affect the regulation of the *SCML1* gene. However, no potential regulatory element was observed to support this speculation, though testis-specific data in ENCODE is limited. Overall, these findings suggest that while the

deletion of the first exon of the *NHS* gene is less likely to be the primary cause of the patient's phenotype, the possibility of regulatory effects on *SCML1* cannot be ruled out.

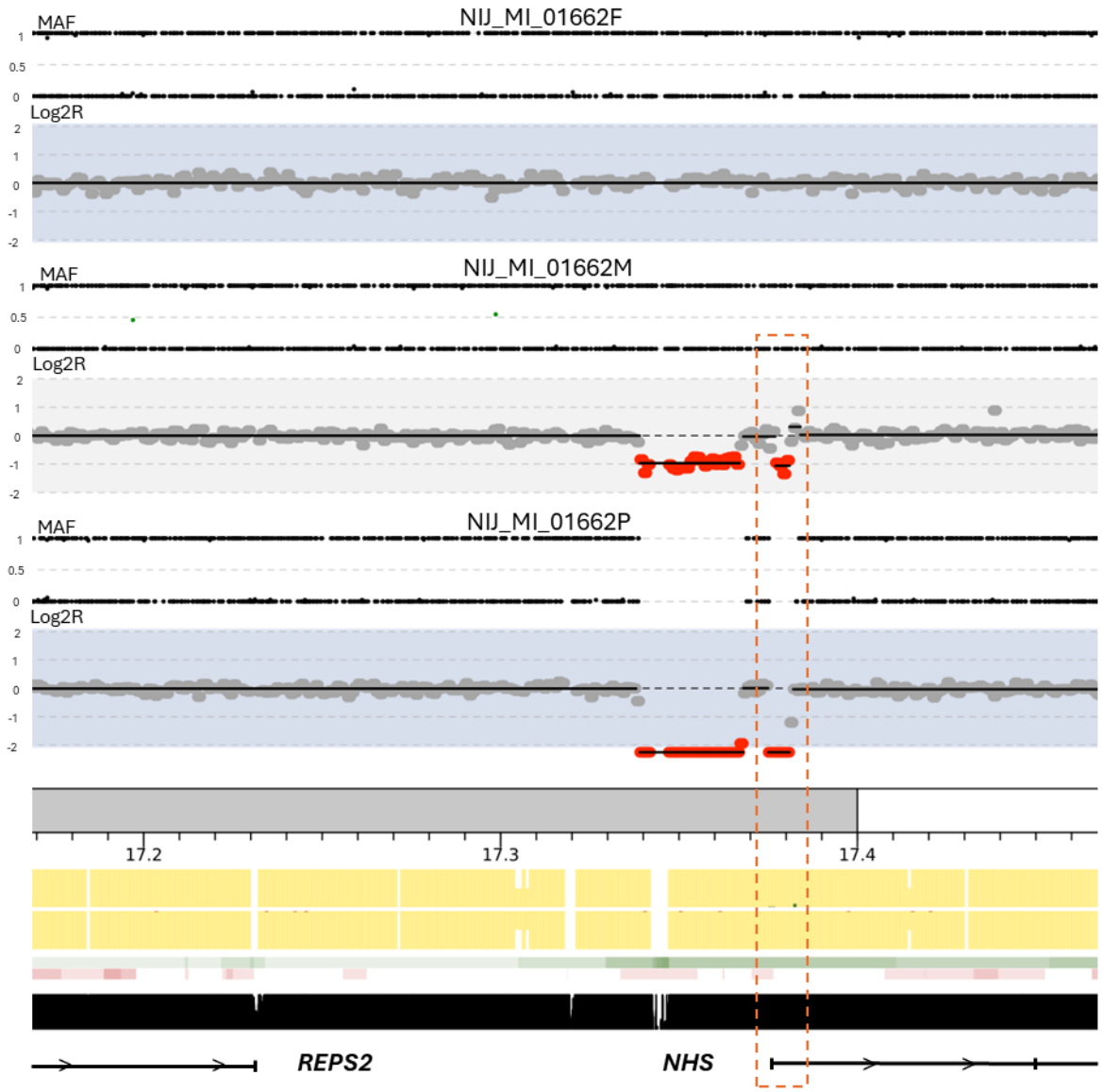


Figure 6.13. CNVRobot plot of the 6kb hemizygous MI deletion detected on chromosome X in patient NIJ_MI_01662P (highlighted with an orange dashed rectangle). The deletion is present in both the mother and the proband, disrupting the only copy of *NHS* gene. The Log2R score of -1 indicates a single copy of the deleted region in the mother and the score of -2 indicates 0 copies in the proband. No SNP was observed for only a part of the deletion in the proband. There is also 29kb deletion nearby the deletion without affecting any gene but may affect the regulation of *NHS* gene.

6.3.4 Replication Study of Candidate Genes

The genes (*UBR2, AKT1, IPO9, FAS, ACTA2, FKBP8, NRK, SYAP1*, and *DYNLT3*), which may be affected by MI SVs and potentially contribute to the observed patient phenotypes, were further investigated in replication cohorts as well as fertile controls (see Chapter 2.5). I

searched for potentially pathogenic or likely pathogenic LoF SNVs in these genes within the Genomics of Male Infertility Group cohort, the MERGE study from Germany, and the fertile control cohort.

One likely pathogenic frameshift SNV in *UBR2*(NM_001363705.2):c.345_346del p.(Ala116SerfsTer2) was identified in singleton NCL_MI_0121P, a patient with azoospermia from Newcastle. One likely pathogenic splice site SNV in *AKT1* (NM_005163.2:c.567+2T>G) was identified in singleton M2315, a patient with azoospermia in MERGE cohort. No pathogenic LoF SNVs were identified in the remaining genes (*IPO9, FAS, ACTA2, FKBP8, NRK, SYAP1,* and *DYNLT3*) or in the fertile control cohort.

6.4 Discussion

In this chapter, I investigated the role of rare MI SVs in a cohort of 216 patient-parent trios presenting with NOA and severe oligozoospermia. I aimed to identify SVs that might contribute to idiopathic male infertility through dominant inheritance patterns.

Our initial comparison between rare MI and PI SVs, after excluding low-confidence calls, common variants, and other confounding factors, revealed 1,855 MI SVs and 1,732 PI SVs. These SVs were categorised based on their size. Statistical analysis using a Pearson's Chi-squared test of independence revealed a significant difference between MI and PI SVs across these categories ($p=0.04138$). However, residual analysis indicated that this difference was primarily driven by an over-representation of <2.5kb duplications in the PI group and >100kb duplications in the MI group. This could be explained either by the increased error susceptibility in the identification of duplications due to technical limitations or by longer SVs in fathers to be detected as multiple smaller SVs, rather than reflecting biological variation.

Given the challenge of prioritising likely pathogenic SVs among numerous inherited variants, I focused on MI SVs that might have minimal impact on female fertility but significant effects on male fertility. Such SVs could be transmitted from mothers to sons without undergoing negative selection. I applied two filtration strategies. First, we identified rare MI SVs encompassing at least one exon of Known and Candidate Male Infertility Genes (KCMIG) with a score >1, a probability of loss-of-function intolerance (pLI) score >0.9, and a DOMINO score >0.5, indicating dominant inheritance potential. Second, I aimed to reveal new candidate genes by selecting rare MI SVs larger than 5kb that encompass at least one exon of genes with a pLI score over 0.9 and are absent in paternal samples.

Using the first strategy, I identified a previously unreported MI tandem duplication in patient NIJ_MI_00347P, which encompasses the last five exons of the *UBR2* gene (pLI = 1). Although the gene structure remains intact, the gene may be affected by the duplication of distal enhancers within the duplicated region and potential changes in the 3D structure. *UBR2* (Ubiquitin Protein Ligase E3 Component N-Recognin 2) gene encodes E3 ubiquitin-protein ligase which is a component of the N-end rule pathway. GO annotations for biological processes show that this gene is involved in male meiosis I (GO:0007141), male meiotic nuclear division (GO:0007140), reciprocal meiotic recombination (GO:0007131), and spermatogenesis (GO:0007283). Houston et al., 2021, concluded that the gene is associated with NOA with limited evidence. The pathology report for patient NIJ_MI_00347P indicated that some germ cells were observed in both histological sections and cytological preparations, and spermatozoa were present in the cell suspension. The overall conclusion from the cytology analysis was SCO/minimal GCA, while the histology conclusion noted severe hypospermatogenesis (HS-severe) combined with tubular hyalinization[14]. The patient underwent TESE, which resulted in the retrieval of non-motile sperm. This was followed by ICSI, but no pregnancy was achieved. The duplication disrupting the likely dominant *UBR2* gene in our patient may potentially have led to impaired meiotic processes and spermatogenesis, resulting in extreme oligozoospermia. Additionally, a likely pathogenic frameshift SNV in *UBR2*(NM_001363705.2):c.345_346del p.(Ala116SerfsTer2) was identified in singleton NCL_MI_0121P, a patient with azoospermia in the replication study. Although segregation analysis could not be performed, this represents the second case harbouring a potentially pathogenic mutation associated with the similar phenotype, with no LoF variants detected in the control cohort. The involvement of *UBR2* in critical meiotic and spermatogenic pathways underscores its importance in male fertility. The genetic disruption could perhaps cause a meiotic arrest or defect in spermiogenesis, contributing to the observed infertility. Given the limited evidence and the sporadic nature of the findings, further studies and larger cohorts are needed to provide further evidence for a causal role of mutations affecting this gene in male infertility. Seeking additional cases through collaborations and gene-matching platforms like GeneMatcher (Sobreira et al., 2015) would be a valuable next step. While acknowledging the difficulty in obtaining patient tissue to investigate the direct impact of the identified duplication, the fundamental role of *UBR2* in meiotic and spermatogenic pathways

[14] There are no germ cells or Sertoli cells in the seminiferous tubules of the testis.

can be investigated by creating an in vitro model. For example, using CRISPR-Cas9 to disrupt *UBR2* function in a relevant cell line would allow for the assessment of key cellular consequences, such as defects in meiotic progression, increased rates of apoptosis, or altered expression of critical spermatogenesis-related genes.

Applying the second strategy, we identified three rare MI deletions. In patient NIJ_MI_00352P, a 697kb deletion on chromosome 14 encompassing 12 genes, including *AKT1* (pLI = 0.98), *CEP170B* (pLI = 1), and *INF2* (pLI = 0.97). Akt (Protein kinase B, PKB) is a serine/threonine kinase that plays a key role in regulating cell survival, insulin signalling, angiogenesis, and tumor formation. The somatic gene mutations has been associated with several cancers, including breast, ovarian, and colorectal cancers (OMIM:164730) and predicted to be dominant by DOMINO. Aitken & Koppers, (2011) reviewed earlier research on DNA damage in spermatozoa, suggesting that apoptosis is the default pathway for these cells. They noted that there seem to be no endogenous chemical triggers for this process, with pro-survival factors being the only elements that prevent spermatozoa from undergoing apoptosis. Later that year, Koppers et al., (2011) identified *AKT1* as a crucial factor for spermatozoa survival in humans, proposing that its phosphorylated state prevents cells from following the default apoptotic pathway (Aitken, 2018; Koppers et al., 2011). Additionally, the *AKT1* gene was shown to be vital for normal spermatogenesis in mice (Kim et al., 2012). The deletion also affected one copy of the high pLI score genes *INF2* (pLI=0.97) and *CEP170B* (pLI=1). However, since the functions of these genes in humans are still unknown, we could not determine the impact of their deletion. Hence, the deletion may cause disruption in *AKT1* gene leading to NOA observed in patient. Additionally, the likely pathogenic splicing site mutation in *AKT1* gene was identified in singleton M2315, a patient with azoospermia attributed to meiotic arrest, in the replication study. Although segregation analysis could not be performed, this represents the second case harbouring a potentially pathogenic mutation associated with the similar phenotype, with no LoF variants detected in the control cohort. Still, further studies are needed to fully understand the exact role of *AKT1* in male infertility. To functionally establish the role of *AKT1* haploinsufficiency in male infertility, a crucial next step would be to use CRISPR-Cas9 to create a heterozygous knockout of *AKT1* in a human germ-cell-like in vitro model. This would allow for a direct assessment of its impact on key processes like apoptosis and meiotic progression, thereby testing the mechanisms suggested by both the deletion and the splicing variant found in the patients.

Additionally, the identification of a deletion affecting *AKT1*, a known proto-oncogene, warrants comment on our study's protocol for managing incidental findings. To handle such findings, our analysis was designed to prospectively exclude genes from the ACMG Secondary Findings (SF) list. However, *AKT1* is not currently included on this list. Consequently, it was not masked during our analysis, which permitted its identification and evaluation as a potential candidate gene for male infertility.

In patient NIJ_MI_00151P, a 6kb deletion removes exons 11-13 of the *IPO9* gene (pLI = 1). *IPO9* is involved in nuclear import and is expressed across various tissues, including the testis. The improper regulation of specific importins was linked to various disorders, including infertility by disrupting miosis shown in mice (Hu et al., 2010; Navarrete-López et al., 2023). Palacios et al., (2021) showed that null mutations in *IPO*9 lead to defects in chromosome segregation and condensation during meiosis in both male and female Drosophila. Even though we might expect to observe phenotypes associated with meiosis such as germ cell aplasia, the patient's histology and cytology report reveals a SCO phenotype. The studies referenced were conducted in mice and Drosophila, and there are no reported mutations in the *IPO9* gene in infertile males. So, it is still possible that this gene could contribute to infertility in this patient.

Among the MI duplications, several encompassed genes with high pLI scores. While duplications are challenging to interpret due to their complex effects on gene dosage and regulation, two duplications are noteworthy. In patient NIJ_MI_01724P, a 136kb tandem duplication encompasses *FAS* (pLI = 1) and *ACTA2* (pLI = 0.93). The *FAS* gene encodes a TNF-receptor superfamily protein that regulates programmed cell death and is associated with malignancies and immune system diseases (OMIM: 134637). No autoimmune disease was reported in the patient. This gene also plays a key role in spermatogenesis by regulating apoptosis through its interaction with the FAS ligand (*FASL*). Sertoli cells in the testes express *FASL*, which binds to the *FAS* receptor on germ cells, triggering apoptosis (Rajender, 2012). This process is essential for maintaining the balance between germ cells and Sertoli cells, ensuring that only a viable number of germ cells are supported, and defective cells are eliminated (M. Wang & Su, 2018; Sharma et al., 2023). Few studies were conducted on the effect of *FAS* gene polymorphisms on male fertility, suggesting that some polymorphisms are associated with altered sperm apoptosis, poor semen quality, and idiopathic azoospermia and oligozoospermia(Ji et al., 2009; W. Wang et al., 2009; Asgari et al., 2017). However, there is

no rare pathogenic SNVs or SVs reported in the *FAS* gene causing male infertility. In our patient with severe OAT, it is likely that the duplication increases *FAS* gene expression (Mottes et al., 2021), which may enhance apoptosis by binding to *FASL*, contributing to the patient's phenotype. Gene and protein expression analysis of the specific genes in the patient is required as a first line test for this hypothesis. The *ACTA2* gene encodes one of six highly conserved actin proteins, which are involved in cell motility, structure, integrity, and intercellular signalling. It has been identified as an RNA marker for human testicular peritubular myoid cells through scRNA sequencing and is expressed from birth onwards (Di Persio & Neuhaus, 2023). Myoid cells are involved in both the mechanical and biochemical regulation of spermatogenesis (MAEKAWA et al., 1996; Welsh et al., 2009). Studies showed that the loss or dysfunction of myoid cells can impair spermatogenesis by affecting the biochemical environment necessary for germ cell development (Zhou et al., 2019). Therefore, it is possible that the disruption of this gene could explain the patient's phenotype by causing reduced sperm motility, impaired spermatogenesis, and defects in sperm maturation. However, it is first necessary to show and prove how this gene is affected by the duplication and, if so, to investigate the downstream effects.

A 12kb duplication on chromosome 19 in patient NIJ_MI_02365P with extreme oligozoospermia disrupts the *FKBP8* gene (pLI=1) (Figure 6.8). Wyrwoll et al., 2022, reported six individuals with bi-allelic loss-of-function mutations in the *FKBP6* gene, which belongs to the same family as the *FKBP8* gene, all of whom had either no sperm or extremely few sperm in the ejaculate. The *FKBP8* gene encodes FK506-binding protein 8, which plays a role in maintaining mitochondrial function by regulating mitochondrial dynamics and inhibiting apoptosis. It is closely related to mitophagy, the selective autophagic process that targets damaged or unnecessary mitochondria for degradation. The *FKBP8* gene initiates mitophagy by interacting with LC3 through its LC3-interacting receptor (LIR) domain, facilitating the clearance of defective mitochondria (Kirat et al., 2023). Mitophagy is essential for maintaining mitochondrial quality and energy homeostasis during spermatogenesis, particularly in the processes where mitochondrial dynamics are critical, such as during spermiogenesis. Any disruption in mitophagy could lead to mitochondrial dysfunction, impairing the energy supply and homeostasis required for spermatogenesis (Rotimi et al., 2024). Consequently, the duplication identified in the patient might contribute to the patient`s phenotype by interfering with the proper maturation of germ cells and the release of mature sperms.

I also identified a validated 659kb MI inversion on chromosome 16 in patient NIJ_MI_01490P, encompassing genes exclusively expressed in the testis: *PRM1, PRM2, PRM3,* and *TNP2*. Protamines (PRMs) are believed to play a crucial role in chromatin condensation, transcriptional repression, preservation of the haploid male genome and sperm formation (Cho et al., 2001). There are two types of protamines, *PRM1* and *PRM2*, encoded by the *PRM1* and *PRM2* genes, respectively, both located on chromosome 16. During sperm development, 85% of histones are replaced by protamines, which help protect DNA from harmful agents. An imbalance in the ratio of histones to protamines has been linked to chromatin deficiencies in sperm, increasing the risk of DNA damage and male infertility. Moreover, maintaining a proper ratio of *PRM1* to *PRM2* (typically between 0.8 and 1.2) is essential for normal sperm function (Hamidian et al., 2020).

Several studies have investigated the significance of the *PRM1* gene in male fertility (Akmal et al., 2016). Polymorphisms in the *PRM1* gene were associated with an increased risk of male infertility (Nemati et al., 2020), and pathogenic heterozygous mutations have been observed in the coding and UTR regions of the gene in patients with OAT (Ravel et al., 2007; Nasirshalal et al., 2020). Additionally, studies with KO mice have shown that the loss of one allele of *PRM1* leads to subfertility and disrupts the proper processing of *PRM2*, highlighting the essential role of *PRM1* in sperm development and function (Merges et al., 2022).

Another gene potentially impacted by the inversion is *TEKT5*, which encodes the Tektin 5 protein. Tektins are crucial components of flagella, and changes in their expression or mutations in Tektins in mice have been linked to defective sperm motility (Cao et al., 2011). Wyrwoll et al., 2022, reported a patient with Sertoli cell-only syndrome (SCO) who had a complete deletion of one allele of the *TEKT5* gene and a potentially causative SNV in the other allele. In the replication study, they also identified other patients with compound heterozygous missense mutations affecting this gene. They argued that, despite the diagnosis of SCO in this patient and different conditions such as maturation arrest or meiotic arrest (MeiA) in others, *TEKT5* could still be a joint underlying cause. This variability in phenotype is similarly observed in patients with pathogenic variants in other well-established male infertility genes, such as *TEX14*. However, it remains uncertain whether the detected *TEKT5* variants are the definitive cause of impaired spermatogenesis in these individuals, and further functional analyses are necessary to clarify the pathogenicity of these missense variants. In summary, the expression of testis-specific genes, either within or near the inversion, is likely

influenced by the inversion. The patient's phenotype aligns with the potential outcomes of disrupted genes as reported in the literature. While RNA analysis was considered to confirm this, we were unable to perform it due to a lack of a suitable sample for RNA extraction. Therefore, the effect of the inversion on downstream processes requires further investigation through other functional studies, and replication studies are needed, as the roles of these genes in male infertility are not yet well established.

All rare SVs on the X chromosome were maternally inherited, as the X chromosome is exclusively transmitted from the mother in males. The potential effect of these variants on the female carriers was also considered, however, given that the mothers in this cohort were fertile, it is likely that any deleterious effect was compensated by the presence of the normal allele on the second X chromosome. Notable findings include a 541kb duplication in patient NIJ_MI_00992P, disrupting the *NRK* gene and encompassing *PWWP3B* and *SERPINA7*. The significance of this variant is complex; while linked to severe developmental phenotypes in the ClinVar and DECIPHER databases, its classification is conflicting and often listed as a VUS. This duplication was previously reported in another oligospermic man (Tüttelmann et al., 2011) , and our study now identifies a second independent case presenting with infertility but lacking the other severe clinical features. Although the recurrence of this rare duplication in two men with a consistent phenotype is significant, the conflicting evidence suggests further studies are needed to clarify its specific role in spermatogenic failure.

In patient NIJ_MI_00625P, a 22kb duplication disrupts *SYAP1* (pLI = 0.94), highly expressed in the testis. *SYAP1* plays a role in signalling pathways related to cell growth and survival. A similar size duplication involving *SYAP1* have been reported Frank Tüttelmann et al., 2011, in patients with Sertoli cell-only syndrome, indicating its possible contribution to the patient's NOA phenotype.

Another noteworthy SV is a complex rearrangement in patient NIJ_MI_00151P involving a duplication and deletion disrupting *DYNLT3*. *DYNLT3* is essential for chromosome alignment and segregation during meiosis. Houston et al., (2021) evaluated the *DYNLT3* gene and classified it as having "no evidence" for a role in infertility. However, the impact of this CS on the gene product may affect meiosis and potentially explain the patient's phenotype, which could contribute to future updates in gene classification. It's also worth noting that a MI deletion of the *IPO9* gene was introduced in this chapter as a potential cause of the phenotype

in the same patient. As *IPO9* is also related to meiosis, it is possible that one or both SVs contribute to the observed phenotype.

Our study underscores the potential impact of rare MI SVs on male infertility, particularly those affecting genes essential for spermatogenesis. However, functional validation is necessary to confirm the pathogenicity of these variants. The interpretation of duplications and complex SVs remains challenging due to the intricacies of gene regulation and dosage sensitivity. Future studies should include functional assays to assess the impact of identified SVs on gene expression and spermatogenic processes. Expanding the sample size and including fertile controls would enhance statistical power and aid in distinguishing pathogenic variants from benign polymorphisms. Utilising model organisms to study the effects of these SVs on fertility and employing advanced genomic techniques, such as long-read sequencing, would provide a more comprehensive understanding of these complex variants.

6.5 Conclusion

In this chapter, I investigated rare MI SVs in 216 patient-parent trios with NOA and severe oligozoospermia. I found a balanced distribution of MI and PI SVs. Several candidate MI variants, including a novel tandem duplication affecting the *UBR2* gene and deletions in *AKT1* and *IPO9*, were identified as potential disruptors of spermatogenesis. MI duplications in *FAS*, *ACTA2*, and *FKBP8* suggest roles in apoptosis regulation and mitochondrial function. Additionally, an MI inversion on chromosome 16 highlights the complexity of SVs in male infertility. Furthermore, the analysis of MI X-linked SVs highlighted several potentially significant findings, such as duplications in genes like *NRK* and *SYAP1*, which could have implications for spermatogenesis and warrant further investigation.

This study underscores the value of examining genetic variations in male infertility cohorts with available parental samples. We hope that insights from this research, along with the exploration of *de novo* variants discussed in the previous chapter, will encourage the continued advancement of genetic research in male infertility.

Chapter 7. Analysis of SVs and cnnLOHs in Idiopathic NOA and Severe Oligozoospermia: A Recessive Inheritance Perspective

7.1 Introduction

While previous chapters examined dominant inheritance patterns, this chapter explores the role of recessive genetic mechanisms in specific forms of male infertility, namely idiopathic NOA and severe oligozoospermia. Recessive forms arise when two copies of a deleterious variant, either homozygous or compound heterozygous, disrupt the function of both alleles of an autosomal gene. This mechanism is distinct from dominant forms, where a single defective allele is sufficient to cause infertility. Consanguinity or founder effects can increase the prevalence of recessive disorders in certain populations (Xiao and Lauschke, 2021). For instance, Inhorn et al., 2009, showed a notable increase in male infertility seen in consanguineous couples.

The investigation of recessive mechanisms underlying infertility in idiopathic NOA and severe oligospermia can provide an understanding of the genetic contributors to diseases. As an example, Olinger et al., 2024, conducted a large-scale analysis of bi-allelic CNVs in Genomics England's 100,000 Genomes Project data. They examined 11,754 parent-child trios and an additional 18,875 non-trios, with a separate control cohort of 15,440 cancer patients used to assess independent deletion frequencies. By focusing on an AR inheritance model, they identified 34 rare deletions that were homozygous in affected individuals and heterozygous in both parents. These bi-allelic deletions were detected in only 52 trios, affecting 37 genes, 8 of which were previously linked to AR disorders. Among the remaining 29 genes with no known disease association, *SLC66A1* emerged as a candidate for AR rod-cone dystrophy in four families. As we shift our focus to the genetic underpinnings of male infertility, most of the identified genes related to reproductive disorders with at least moderate level of evidence follow an AR inheritance pattern (Houston et al., 2021; Stallmeyer et al., 2024). The reason why the majority of identified genes exhibit an AR mode of inheritance could be that men with infertility are less likely to reproduce and pass on pathogenic variants. Additionally, pedigree data are often insufficient to comprehensively investigate AD and X-linked conditions in this context. Most of the research has primarily focused on recessive and X-linked inheritance, as these are easier to study in singleton data. Rare homozygous or hemizygous variants are more straightforward to interpret than rare heterozygous variants, which are often overlooked.

However, with the trio-based studies, the field has been able to expand its focus beyond recessive and X-linked conditions. This shift is similar with intellectual disability research, where initial studies predominantly identified recessive and X-linked causes, but dominant *de novo* variants became the primary focus once trio data became available (Vissers et al., 2016). While our focus has been on the understudied dominant causes of infertility, trio-data are also highly suitable for studying recessive inheritance, particularly because they enable the identification of compound heterozygous variants, which would be challenging to detect in singleton studies.

Researchers have revealed biallelic mutations in a wide range of genes such as *TAF4B, TEX15, SPINK2, NPAS2 and FANCM* through studies of azoospermia-affected individuals in families, many of which involve consanguineous relationships (Ayhan et al., 2014; Ramasamy et al., 2015; Colombo et al., 2017; Kherraf et al., 2017; Kasak et al., 2018). More recently, Tian, Wang, et al., 2023 and S. Li et al., 2024, identified two recessive candidate genes, *ADAD2* and *DNAH3*, where biallelic mutations cause NOA and OAT, respectively. Moreover, biallelic mutations in the *CFAP43, CFAP44,* and *CFAP54* genes were demonstrated to cause multiple morphological abnormalities of the flagella (MMAF) (Tang et al., 2017; Tian, Tu, et al., 2023). Most research has focused on the role of point mutations, and very few studies have examined the role of CNVs in causing recessive male infertility. However, Wyrwoll et al., 2022, specifically analysed CNVs using aCGH and WES, and identified two individuals with homozygous deletions in the *SYCE1* gene and one individual with a compound heterozygous condition, consisting of a heterozygous *SYCE1* deletion and a likely pathogenic *SYCE1* missense variant on the other allele, explaining the genetic cause of their infertility.

Another important aspect to consider is the contribution of copy-neutral loss of heterozygosity (cnnLOH) to genetic disorders (Gilissen et al., 2012). Homozygosity across large genomic regions can result from uniparental disomy (UPD) or occur in cases of consanguinity marriages. These regions can be detected using SNP microarrays and then incorporated into the prioritisation process. However, Becker et al., 2011, demonstrated that exome data alone provides sufficient informative SNPs for accurate homozygosity mapping. Furthermore, CNVRobot, the tool used in our analysis, identifies cnnLOH regions alongside CNVs. These homozygous regions may unmask recessive mutations, thus playing a critical role in the expression of recessive disorders. Understanding cnnLOH regions is crucial to uncover hidden

recessive causes of infertility given the significant incidence of male infertility in consanguineous populations (Inhorn et al., 2009).

The aim of this chapter is to investigate the recessive mechanisms contributing to male infertility within our cohort. This involves screening for both homozygous deletions and compound heterozygous variants, where a deletion in one allele is accompanied by a second possibly pathogenic variant in the other allele. Additionally, the analysis includes identifying and studying cnnLOH regions, which may unmask recessive mutations that contribute to the observed infertility in this cohort.

7.2 Aims

This chapter aims to

- Prioritise and evaluate the likely pathogenicity of rare homozygous and compound heterozygous SVs.
- Evaluate the distributions of cnnLOHs across the genome and determine the mechanism of occurrence
- Identify novel candidate recessive genes in infertile men with idiopathic NOA and severe oligospermia

7.3 Results

In this chapter, I investigated recessive mechanisms contributing to male infertility by examining a cohort of 216 patient-parent trios, all of whom presented with idiopathic NOA or severe oligozoospermia. By focusing on these trios, I aimed to identify and characterize SVs that could underlie recessive forms of male infertility. This approach leverages the family-based nature of the data, increasing the likelihood of uncovering recessive patterns of inheritance that may not be evident in sporadic cases.

As described in Chapter 3, our analysis combined two independent tools, CNVRobot and dysgu-sv, to comprehensively identify SVs within the cohort. Previous work in Chapter 5 focused on detecting *de novo* SVs, followed by an examination of maternally inherited SVs in Chapter 6. Building upon these findings, this chapter delves into recessive inheritance patterns. I aimed to pinpoint SVs inherited from both parental alleles that could contribute to the observed infertility phenotypes.

To refine our analysis and target potentially recessive variants, I first excluded gonosomal SVs. These sex-linked variants were thoroughly assessed in earlier chapters, where their maternal

or paternal modes of inheritance were already established. Additionally, I removed duplications as they pose interpretive challenges because it is often unclear how many copies are present, whether the duplicated segments occur on the same allele or in a bi-allelic configuration, and how varying copy numbers influence the phenotype. Restricting our focus primarily to deletions allowed for a more straightforward interpretation of recessive effects.

For CNVRobot calls, I filtered only those variants with a mean log2ratio less than -1.5, indicating a homozygous deletion. For dysgu-sv calls, I applied filters based on genotype and inheritance predictions, retaining only variants that were homozygous (1/1 denotes homozygosity) indicated as inherited from both parents. Through this filtering process, I identified a total of 307 homozygous deletions.

7.3.1 Biallelic Rare Autosomal Inherited Deletions

From the 307 identified homozygous deletions, a total of 115 were in intragenic and 119 in intergenic regions. To pinpoint these deletions most likely to contribute to male infertility through a recessive mechanism, I employed two filtration strategies, building on the approaches developed in previous chapters.

(i) To investigate known disease genes, I first focused on deletions that potentially disrupt genes already implicated in male infertility. Homozygous deletions affecting at least one exon of KCMIG with a score >1 were retained for further analysis. Heterozygous deletions affecting at least one exon of KCMIG, following a recessive pattern and with a score greater than 1 (to investigate potential compound heterozygosity), were also retained. Applying these criteria, no homozygous deletions remained after filtering. However, I identified seven heterozygous deletions, all affecting at least one exon of a gene with a KCMIG score >1, following a recessive pattern. Among these 7, there were 2 deletions involved with *DPY19L2* gene which is known to cause globozoospermia. Since globozoospermia does not match the azoospermia phenotype of our patients, these variants were excluded from subsequent analyses. For the remaining 5 heterozygous deletions, I sought pathogenic or VUS SNVs on the second allele to establish a compound heterozygous basis. However, no pathogenic SNVs were identified int these affected genes in these patients.

(ii) To uncover novel candidate genes potentially involved in the recessive inheritance of male infertility, I expanded my analysis of rare homozygous deletions to those that encompassed at least one exon of any human gene. This filtration yielded 2 homozygous deletions (Table

7.1). One notable case involved a deletion removing a single exon of the *GBR14* gene in a patient NIJ_MI_02199P who exhibited multiple cnnLOH regions suggesting a high degree of consanguinity (Figure 7.1. A).

Table 7.1. Proband, genomic locations, size, and genes affected by rare homozygous deletions encompassing at least one exon identified in the trio cohort.

Proband	Type	Genomic Location (GRCh38)	Size	Genes
NIJ_MI_02199P	DEL	chr2:164566745-164574418	8kb	*GRB14*
NIJ_MI_00433P	DEL	chr7:143127754-143196814	69kb	*PIP*


A
MAF
NIJ_MI_02199F
Log2R
MAF
NIJ_MI_02199M
Log2R
MAF
NIJ_MI_02199P
Log2R
q24.3
162
163
164
165
166
167
B
MAF
NIJ_MI_02199F
Log2R
MAF
NIJ_MI_02199M
Log2R
MAF
NIJ_MI_02199P
Log2R
164.6
GRB14

Figure 7.1. CNVRobot plot of the loss of heterozygosity (LOH) region identified in patient NIJ_MI_02199P, along with a bi-allelic deletion located within the LOH region. A. CNVRobot plot of a large stretch of loss of heterozygosity (LOH) identified in patient NIJ_MI_02199P. The region is highlighted in light blue in the proband's log2R track. No heterozygous SNP was observed (MAF = 0.5), while a log2R ratio of 0 indicates no change in copy number. **B.** CNVRobot plot of the homozygous deletion detected in patient NIJ_MI_02199P within the LOH region. This deletion is observed in the proband as homozygous and in both parents as heterozygous. In the parents, no heterozygous SNP was observed within the deleted region, and a log2R value of -1 confirms a heterozygous deletion. In the proband, no SNP was observed within the deleted region, and a log2R value of -2 confirms a homozygous deletion. The gene track clearly demonstrates that this deletion removes one exon of the *GRB14* gene on both alleles.

The 8kb homozygous deletion, inherited from both maternal and paternal alleles, removes a single exon of the *GRB14* gene (pLI = 0, DOMINO=0.508 - either recessive or dominant) (Figure 7.1.B). This deletion has been reported in five individuals in a heterozygous state (gnomAD: DEL_CHR2_F7845F9B, AF = 0.00003965). While many LoF mutations, including deletions, have been reported for this gene in gnomAD, all are observed in a heterozygous state, no homozygous LoF cases have been reported. This in-frame deletion removes 21 amino acids from the Ras-associating domain.

The *GRB14* gene encodes a growth factor receptor-binding protein that is part of a small family of adapter proteins known to interact with receptor tyrosine kinases and other signaling molecules. This protein specifically interacts with insulin receptors and insulin-like growth factor receptors, exerting an inhibitory effect on receptor tyrosine kinase signalling, particularly insulin receptor signalling. Interestingly, *GRB14* is highly expressed in the liver and testis according to Protein Atlas Database. However, no direct association with spermatogenesis or male infertility has been established to date. Despite the lack of a clear link to male infertility, protein-protein interaction analysis performed using STRING revealed that *GRB14* interacts with the *FGFR1* gene based on experimental/biochemical data (Figure 7.2). *FGFR1* is a well-established gene implicated in azoospermia (Houston et al., 2021). In replication study, no likely pathogenic/pathogenic LoF SNV found in the *GRB14* gene in the replication and control cohort.

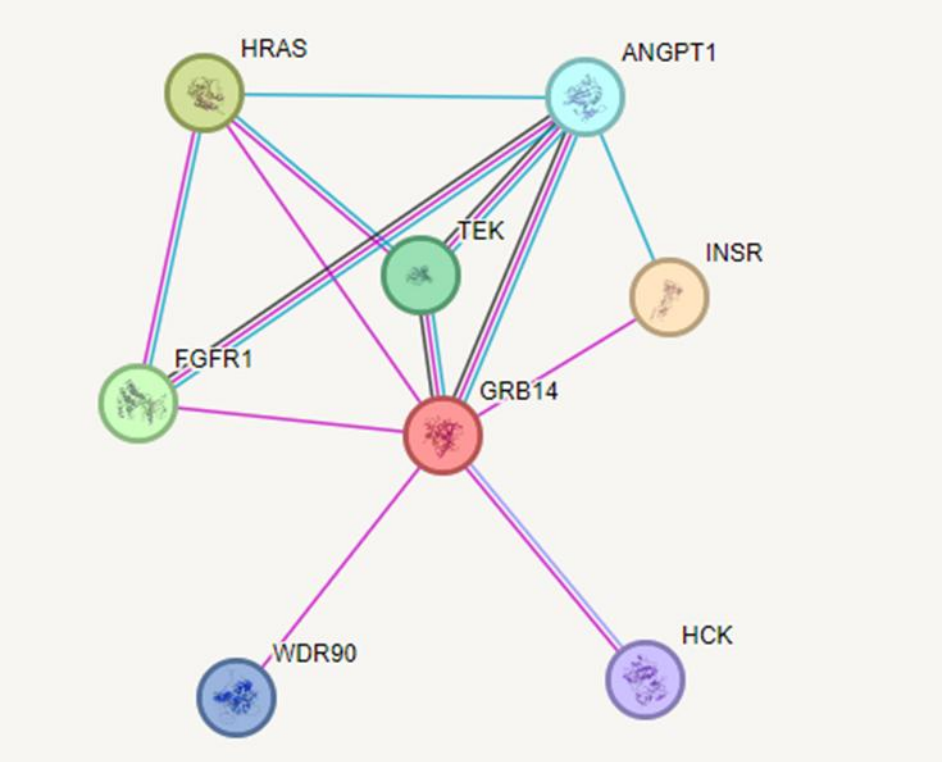


Figure 7.2. STRING analysis for GRB14. Edges represent protein-protein associations. The pink edges represent experimentally validated interactions between the proteins (e.g., between GRB14 and FGFR1). Black and blue ones represent co-expression and information derived from curated databases, respectively. Purple edges indicate protein homology.

The second homozygous deletion, spanning 69kb, was identified in patient NIJ_MI_00433P with azoospermia (Figure 7.3). This deletion removes the entire Prolactin induced protein gene (*PIP*) (pLI = 0.31) on both alleles. It has been reported in 965 individuals in a heterozygous state and in 5 individuals in a homozygous state, 3 of whom are males (GnomAD-SV: DEL_CHR7_5B3B49D4, AF = 0.007653).

The *PIP* gene is a protein-coding gene involved in several biological pathways including MIF-mediated glucocorticoid regulation and ERK signalling. GO annotations for PIP highlight its roles in actin binding and protein binding. According to the Protein Atlas database, *PIP* exhibits selective cytoplasmic expression in salivary glandular cells and breast glands, with low expression in the seminal vesicle. In replication study, no likely pathogenic/pathogenic LoF SNV were found in the *PIP* gene in either infertile patients or the control cohort.

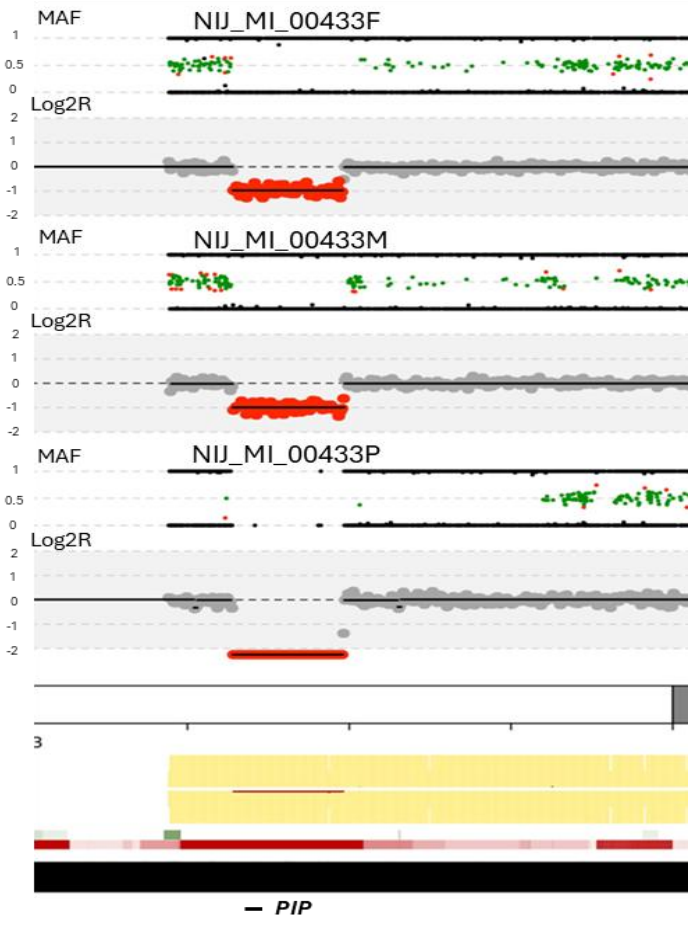


Figure 7.3. CNVRobot plot of the homozygous deletion affecting *PIP* gene detected in patient NIJ_MI_00433P. This deletion is observed in the proband as homozygous and in both parents as heterozygous. In the parents, no heterozygous SNP was observed (MAF = 0.5) within the deleted region, and a log2R value of -1 confirms a heterozygous deletion. In the proband, no SNP was observed within the deleted region, and a log2R value of -2 confirms a homozygous deletion.

To investigate potential compound heterozygosity involving an SV on one allele and a SNV on the other, I focused on high-quality CNVs, those detected by both CNVRobot and dysgu-sv, that encompassed at least one exon of genes with KCMIG scores <2. This analysis identified 18 deletions affecting such genes. My fellow PhD students, Shrooq Alzahrani and Cris Diaz Franco, who are working on SNVs, screened for pathogenic SNVs within them but found none.

7.3.2 Systematic Analysis of Balanced LOHs in the Cohort

In the context of interpreting CNVs in recessive forms of male infertility, it is crucial to examine large spans of LOH regions. A cnnLOH (copy number neutral Loss of Heterozygosity) region is defined as a genomic segment that maintains a normal copy number (i.e., it is neither a duplication nor a deletion) but shows homozygosity throughout the region. Such regions may arise from uniparental disomy (UPD) or be associated with cases of consanguinity. The tool CNVRobot, which I used in this analysis, identifies regions of cnnLOH alongside CNVs, with a size cut-off of 5 Mb (see Figure 7.1.A as an example of cnnLOH).

In total, I detected 173 cnnLOH regions in 87 probands, including 20 in the proband IND_MI_030P and 18 in the proband NIJ_MI_2199P (Figure 7.4). Notably, these two probands were flagged as potentially consanguineous in an independent analysis of genetic relativeness conducted by Dr. Miguel. As reported above, one of these cnnLOH regions in NIJ_MI_2199P harboured a biallelic deletion affecting the *GRB14* gene.

To further investigate potential pathogenic variants, I examined whether any of the cnnLOH regions encompassed genes with a KCMIG score greater than 1. I identified 39 such LOH regions that contained genes scoring above this threshold. Among these, 23 genes are known to have AR inheritance patterns. I then shared these 39 cnnLOH regions with colleagues who work on SNVs, so they could focus on identifying homozygous SNVs in the affected patients in these regions and potentially uncover clinically relevant variants. This work is still ongoing.

Another layer of complexity arises when UPD occurs, as it can affect the expression of imprinted genes. Since imprinted genes have expression patterns that depend on their parental origin, UPD can lead to loss of expression if both alleles originate from the same parent. To explore this, I compared the cnnLOH regions to a curated list of imprinted genes from GeneImprint (Falls et al., 1999). I found that 17 LOH regions encompassed known imprinted genes. Upon further inspection, checking individual SNVs and patterns of

inheritance, no informative SNPs were found confirming UPD. Consequently, no additional investigations were pursued in these cases.

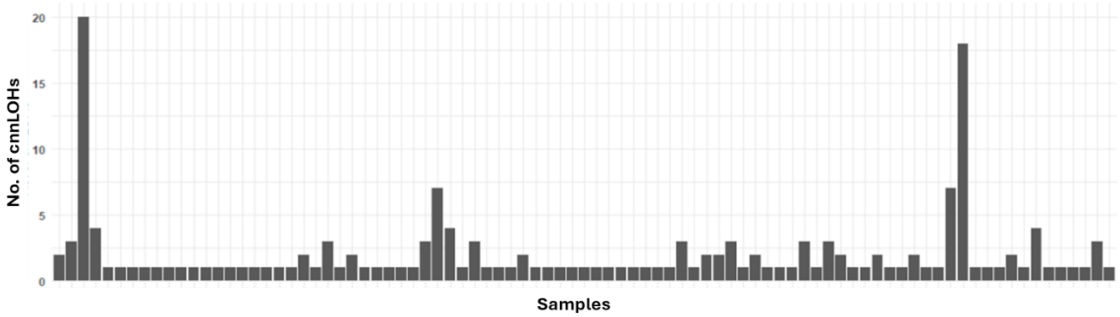


Figure 7.4. Distribution of 173 cnnLOH regions across 87 probands, with notable cases including 20 cnnLOH regions in proband IND_MI_030P and 18 cnnLOH regions in proband NIJ_MI_2199P.

It might be useful here to summarise the genetic findings from the same patient cohort. Across the analyses for *de novo* (Chapter 5), maternally inherited (Chapter 6), and this recessive inheritance (Chapter 7), a total of 34 candidate SVs were identified in 32 unique patients. The majority of these individuals (30 of 32) carried a single candidate variant. However, two patients were found to have multiple candidate SVs identified. Patient NIJ_MI_00352P carried both a *de novo* intergenic SV and a large maternally inherited deletion encompassing the *AKT1* gene, where we think the latter could possibly explain the phenotype. The other patient, NIJ_MI_00151P, was found to have two different maternally inherited variants: an autosomal deletion involving *IPO9* and an X-linked SV involving *DYNLT3*. As both of these genes are involved in meiosis, this case was particularly complex, and we concluded that it is possible that either or both variants contribute to the patient's phenotype.

7.4 Discussion

The primary aim of this chapter was to explore recessive mechanisms underlying severe male infertility by leveraging a large cohort of 216 patient-parent trios. By focusing on trios, I sought to uncover recessive variants, specifically biallelic deletions, that might have been masked in sporadic presentations of azoospermia and severe oligozoospermia. Building on earlier investigations described in Chapters 3, 5, and 6, this approach was designed to identify homozygous or compound heterozygous variants that disrupt genes essential for normal spermatogenesis.

After refining the set of candidate variants by excluding gonosomal SVs and duplications to simplify interpretation, I focused on biallelic deletions. This resulted in a set of 307 homozygous deletions, which is substantially fewer than the number of heterozygous deletions, an outcome consistent with the structural variant reference published by (Collins et al., 2020). Among the 307 homozygous deletions identified, 115 were intragenic. The vast majority of these (113) were entirely intronic, which was an expected finding given that introns make up more than 30% of the genome. From this set, two notable events emerged: one involved the *GRB14* gene and another involved the *PIP* gene. Notably, one of these deletions occurred in a proband which also exhibited extensive cnnLOH regions, suggesting consanguinity and thereby supporting a recessive inheritance pattern.

The first homozygous deletion disrupted a single exon of the *GRB14* gene in a proband (NIJ_MI_2199P) with multiple cnnLOH regions. The *GRB14* gene encodes an adaptor protein that modulates insulin receptor signalling and is expressed in liver and testis, yet no clear role in human spermatogenesis has been established. Although *GRB14* displayed interaction with *FGFR1*, an established azoospermia gene, the absence of a known direct link to testicular function complicates its interpretation. Additionally, functional evidence from mouse models indicates that ablation of *GRB14* had no apparent impact on fertility (Cooney et al., 2004). While informative, these mouse data cannot be directly extrapolated to humans, as species-specific differences in gene function and compensatory pathways may exist. Thus, the *GRB14* finding should be considered preliminary, warranting further investigation into its potential role in human spermatogenesis and male fertility. A comprehensive follow-up study could use CRISPR-Cas9 to assess the broader functional impact of *GRB14* loss in a human germ cell line, while also using co-immunoprecipitation to investigate its specific mechanistic link to spermatogenesis through its interaction with *FGFR1*.

The second homozygous deletion removed the entire *PIP* gene (NIJ_MI_00433P). Although PIP is expressed at low levels in the seminal vesicle, its broader biological roles, ranging from protein binding to modulation of *ERK* signalling, suggest it could be implicated in sperm physiology. Intriguingly, a recent study showed altered levels of *PIP* and *ERK1/2* in sperm from infertile patients compared to healthy controls, as well as improvements in these markers following FSH treatment (Mancini et al., 2023). It has also been proposed that it could influence sperm viscosity (Martinez-Heredia et al., 2008). These findings suggest that PIP may have a subtle, yet significant role in sperm quality and function. Although the complete loss of

PIP in our proband did not immediately provide a definitive link to this azoospermic phenotype. These preliminary insights, coupled with the gene's functions, indicate that PIP could be a candidate for future studies, particularly in light of its potential responsiveness to hormonal modulation. These observations coupled with data on the function of *PIP* gene suggest that this gene could be a candidate for future studies on azoospermia type of infertility, potentially responsiveness to hormonal modulation.

I speculated that one deletion may possibly cause the patient's phenotype by affecting the gene it interacts with, while the other deletion may be influencing the regulation of other genes or hormones. Both cases highlight the potential importance of non-coding causes of infertility, a pattern we also observed in the *de novo* SV analysis. This shows the need to exploring non-coding regions and regulatory mechanisms in future studies of male infertility.

In addition to the two rare biallelic deletions, we also identified heterozygous deletions in KCMIG that might contribute to recessive diseases if a second pathogenic variants present on the other allele. However, no pathogenic SNVs were identified within the 18 genes affected by these deletions.

Moreover, the systematic evaluation of cnnLOH across the cohort underscored the importance of considering large LOH stretches as markers of potential recessive inheritance or consanguinity. We identified 173 cnnLOHs in 87 probands, with particularly extensive LOH in two individuals. Moreover, cnnLOHs may result from UPD, which can affect gene expression and potentially lead to phenotypic consequences. However, in our study, no evidence of UPD emerged to support a direct link between imprinting disturbances.

All together, our findings underscore the complexity of interpreting recessive SVs in male infertility. Although definitive pathogenic roles remain elusive for the specific deletions identified, these results broaden the spectrum of candidate genes and genomic regions that merit further scrutiny. By refining our understanding of recessive inheritance mechanisms and expanding the catalogue of genes involved in male fertility, we move closer to informing more accurate diagnoses and, eventually, personalised therapeutic strategies.

7.5 Conclusion

In conclusion, this chapter highlights the complexity and challenges of identifying recessive structural variants contributing to male infertility. Although I identified intriguing homozygous deletions encompassing *GRB14* and *PIP* genes, as well as several heterozygous deletions in

candidate genes, their direct pathogenic roles in human spermatogenesis remain inconclusive. The presence of large cnnLOH regions and the consideration of UPD further underscore the need for a multifaceted approach to interpret these variants. While no definitive causative variants have been confirmed, the observations presented here expand the repertoire of genomic elements potentially involved in male infertility. Future studies integrating functional assays, advanced genomic analyses, and cross-species comparisons will be critical in translating these preliminary findings into meaningful clinical insights and therapeutic strategies.

Chapter 8. CNVs in a Cohort of Patients with Idiopathic Quantitative Forms of Male Infertility

8.1 Introduction

In addition to the analysed patient-parent trios in previous chapters, I investigated a cohort of 234 patients with idiopathic quantitative forms of male infertility for whom parental DNA samples were unavailable. Most genetic research on male infertility has thus far studied affected individuals only, so this cohort is very representative of a normal study cohort.

Previously, CNV analyses using different technologies, primarily microarray-based comparative genomic hybridisation (aCGH), have revealed new candidate genes in patient-only cohorts. For instance, Lopes et al., 2013 and Lima et al.,2015, reported deletions encompassing the *DMRT1* gene in NOA patients. Moreover, Tüttelmann et al., 2011, and Stouffs et al., 2012, examined the CNV burden on autosomes, comparing patients to controls using aCGH. More recently, Wyrwoll et al., 2022, identified new candidate genes, such as *TEKT5*, discussed in Chapter 6, through the use of both aCGH and WES in a patient cohort. WES has emerged as a powerful tool for uncovering genetic variants including CNVs associated with male infertility (Stallmeyer et al., 2024). While trio-based sequencing facilitates the identification of *de novo* mutations and inheritance patterns, WES studies without parental data can still provide valuable insights into the genetic underpinnings of male infertility (Zhang et al., 2022). In addition, WES has been effective in not only identifying genetic causes of male infertility but also revealing shared genetic roots between male infertility and other diseases (Zhou et al., 2024). On the other hand, since interpreting variants without parental samples is challenging, some researchers focused on the sex chromosomes, as chromosome X is inherited from the mother and chromosome Y from the father (Chianese et al., 2014; Lo Giacco et al., 2014; Yatsenko et al., 2015; Liu et al., 2021; Riera-Escamilla et al., 2022). All these studies showed that CNV analysis in patient-only cohorts can yield valuable insights, even though the absence of inheritance information does limit variant interpretation.

In this chapter, I present the CNV analysis of WES data from 234 male infertility patients with idiopathic quantitative spermatogenic failure. My objectives were to identify rare genetic CNVs that may contribute to the aetiology of male infertility and to expand the spectrum of infertility-associated genes. By leveraging bioinformatic tools and integrating findings from

previous studies, I aimed to enhance our understanding of the genetic basis of male infertility and identify potential targets for future research and clinical intervention.

8.2 Aims

This chapter aims to

- Determine the potential pathogenicity of rare and large CNVs identified in the cohort of 234 patients.
- Reveal novel candidate dominant, recessive and X- or Y-linked male infertility genes.

8.3 Results

The CNV calling from the WES data of the 234 patients identified a total of 4,036 CNVs, with an average of 17 CNVs identified per sample, 6 of which carried more than 100 CNVs. These 6 samples were examined further to assess their validity, and an increased number of deletions were found distributed over all chromosomes. There was not any enrichment on any chromosome or region. It is notable that with increased data noise, there may be an increase in false positive calls for smaller sized CNVs. Therefore, CNVs smaller than 20kb from these 6 samples were excluded from further analysis and caution was applied while interpretating other CNVs from these samples, as many of these may be false positive calls. A total of 3, 577 CNVs remained after exclusion of CNVs smaller than 20kb, with an average of 15 CNVs identified per sample. Low quality CNVs were removed in another round exclusion which resulted in 3533 CNVS with the same average CNVs per sample as mentioned above. Of the total number of remaining CNVs, 1,510 were deletions and 2,023 were duplications (Figure 8.1). 1,607 CNVs were rare (present in <1% of the samples of population databases) comprising 659 deletions and 948 duplications. Common CNVs accounted for 1,926 events, including 851 deletions and 1,075 duplications.

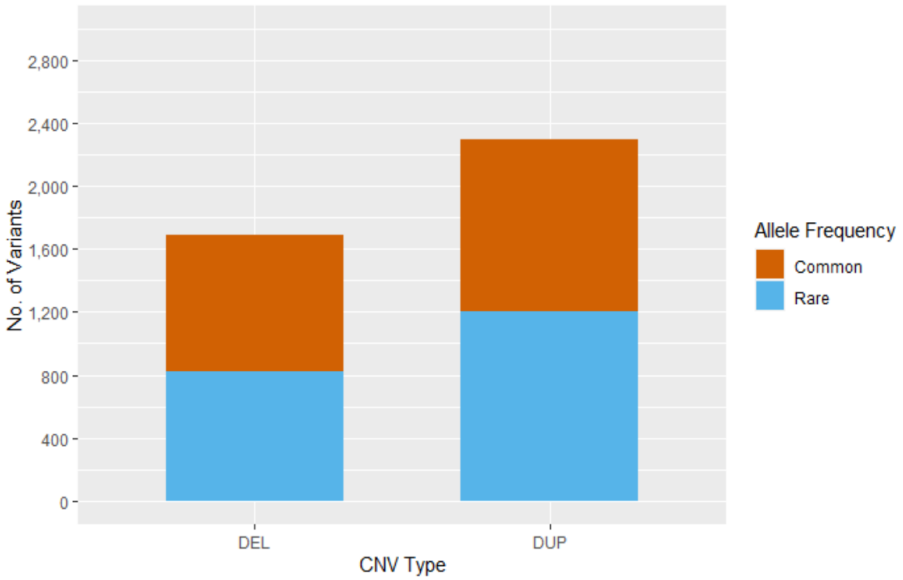


Figure 8.1. Number of CNVs detected in 234 patients based on population allele frequency (rare: <1% in the population databases).

The prioritisation of CNVs was conducted using two complementary approaches. The first approach involved screening for CNVs affecting genes known to be associated with male infertility. This ensured the identification of known pathogenic variations that could explain the observed phenotypes. The second approach focused on revealing novel candidate genes by analysing CNVs in genomic regions not previously implicated in this disease, thereby potentially uncovering new genetic contributors to the condition.

8.3.1 Screening Known Genes

To screen known genes, I applied specific criteria based on inheritance patterns and our predefined Known and Candidate Male Infertility Genes (KCMIG) scores. For autosomal dominant (AD) inheritance model, I selected heterozygous CNVs encompassing KCMIG gene with >1 score. Autosomal recessive (AR) inheritance model was considered for CNVs encompassing genes with an KCMIG score >1 that were either homozygous or compound heterozygous. For the gonosomal inheritance model, CNVs with KCMIG scores >1 were considered. After applying these filtration criteria, CNV plots generated by CNVRobot were visually inspected and false positives were excluded. This filtration resulted in the identification of 4 deletions, 2 of which were located within the pseudo-autosomal region (PAR) on the X or Y chromosome, and three duplications where the breakpoints potentially affect genes (Table 8.1).

Table 8.1. The prioritised CNVs in the analysis aiming to screen known disease genes, detailing proband ID, CNV type, genotype, genomic coordinates, size, encompassed genes, genes with KCMIG >1 and score for these genes.

Proband	Type	Genotype	Genomic Location (GRCh38)	Size (kb)	Genes	Genes KCMIG >1	Inheritance Pattern in Human	Score
NIJ_MI_02282P	DUP	Het	chr15:51214969-51517283	302	*GLDN-CYP19A1-DMXL2*	*CYP19A1*	AD/AR	5
NCL_MI_0045P	DUP	Het	chr2:48580538-49020283	440	*STON1-GTF2A1L - LHCGR-FSHR*	*LHCGR - FSHR*	AD/AR - AR	5 / 3
NCL_MI_0042P	DEL	Het	chr20:59840239-59900385	60	*PHACTR3-SYCP2*	*SYCP2*	AD	3
NIJ_MI_00823P	DUP	Het	chr7:140686889-140834973	148	*BRAF-ADCK2-NDUFB2*	*BRAF*	AD	2
NIJ_MI_00105P	DEL	Het (PAR)	chrX:155773771-156010510	237	*SPRY3-VAMP7-IL9R*	*VAMP7*	AD	2
NIJ_MI_00380P	DEL	Het (PAR)	chrX:155898011-156010510	113	*VAMP7-IL9R*	*VAMP7*	AD	2
NCL_MI_0179P	DEL	Hom	chr10-133453871-133568549	115	*SYCE1-SCART1-CYP2E1-SPRNP1*	*SYCE1*	AR	2

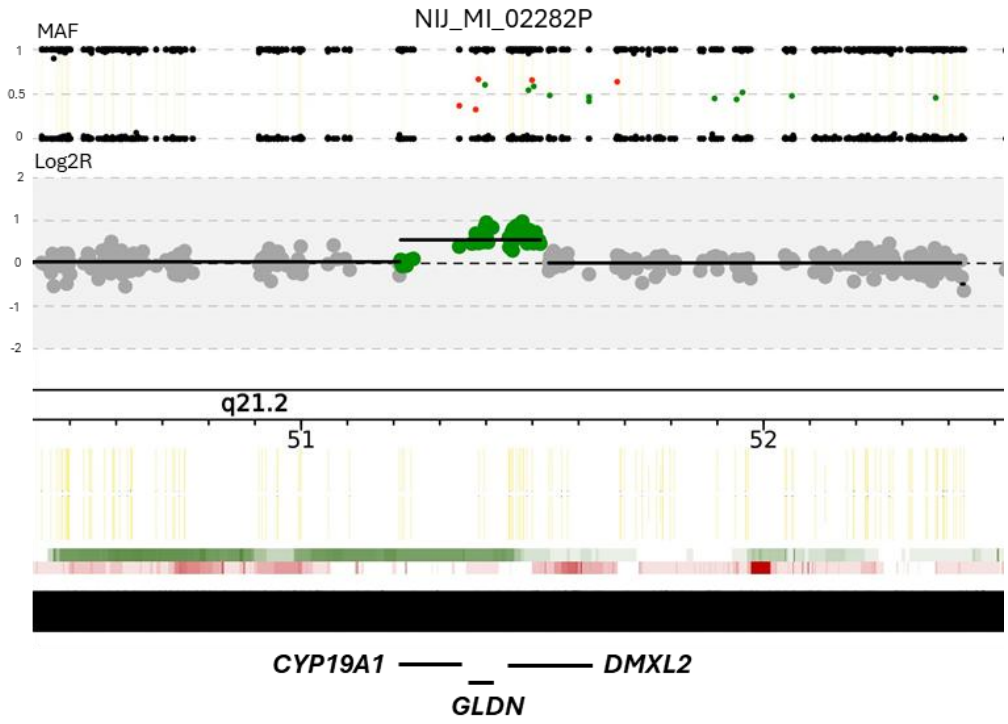


Figure 8.2. CNVRobot plot of the rare 302kb duplication detected in patient NIJ_MI_02282P. The MAF values deviate from the standard positions of 0, 0.5, and 1; instead, intermediate values are observed, indicating additional copy. Also, an expected log2ratio value of approximately 0.58 indicates a heterozygous duplication. The duplication had breakpoints within the coding region of the *CYP19A1* and *DMXL2* genes, and at least one exon lying outside the duplication boundaries.

A 302kb rare heterozygous duplication on chromosome 15 (chr15:51,214,969-51,517,283) was detected in patient NIJ_MI_02282P with azoospermia (Figure 8.2). The similar size duplication has been reported in one individual in GnomAD CNVs v4.1.0 (ID: 174261_DUP, AF= 0.000002154). The duplication in our patient encompasses three genes: *GLDN* (pLI=0, ClinGen Triplosensitivity Score is not available, pTriplo=0.32), *CYP19A1* (pLI=0.02, ClinGen Triplosensitivity Score is not available, pTriplo=0.23), and *DMXL2* (pLI=1, ClinGen Triplosensitivity Score is not available, pTriplo=0.76). One of the breakpoints occurred within the *CYP19A1* gene which is of key interest since it encodes aromatase, an enzyme involved in estrogen biosynthesis and was associated with autosomal dominant aromatase excess syndrome with gynaecomastia (OMI:13300). Also, Houston et al., (2021) concluded that the gene is associated with azoospermia under the category of abnormal development of reproductive organs, with definitive evidence.

In patient NCL_MI_0045P with azoospermia, a 440kb heterozygous duplication on chromosome 2 (chr2:48,580,538-48,729,492) was identified (Figure 8.3). The similar size duplications within the same region was reported in 40 individuals (18 of which are males) in GnomAD CNVs v4.1.0 (ID: 285256_DUP, AF= 0.00008615). The duplication identified in our patient spans multiple genes, fully encompassing *GTF2A1L* (pLI=0, ClinGen Triplosensitivity Score is not available, pTriplo=0.08) and *LHCGR* (pLI=0, ClinGen Triplosensitivity Score is not available, pTriplo=0.11) genes, while partially disrupting *STON1* and *FSHR* genes at the duplication breakpoints. The *LHCGR* and *FSHR* genes encode the luteinizing hormone/choriogonadotropin receptor and the FSH receptor, respectively. *LHCGR* has an AD and AR inheritance pattern with a pLI score of 0, while *FSHR* has an AR inheritance pattern and a pLI score of 0. *LHCGR* was found associated with azoospermia due to Leydig cell dysfunction with hypogonadism (OMIM:238320) (AR), with a definitive level of association, while *FSHR* is associated with azoospermia through hypergonadotropic hypogonadism (AR), with a moderate level of association (Houston et al., 2021). As both genes follow a recessive inheritance pattern, we investigated the presence of any VUS or pathogenic SNVs in *LHCGR* and *FSHR* genes. No VUS or pathogenic SNVs identified in these genes in the patient.

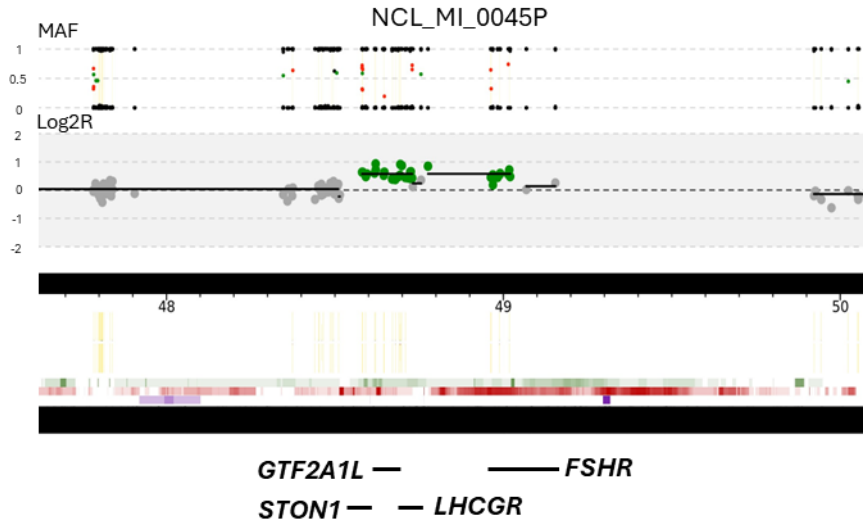


Figure 8.3. CNVRobot plot of the rare 440kb duplication detected in patient NCL_MI_0045P. The MAF values deviate from the standard positions of 0, 0.5, and 1; instead, intermediate values are observed, indicating additional copy. Also, an expected log2ratio value of approximately 0.58 indicates a heterozygous duplication. The duplication had breakpoints within the coding region of the *STON1* and *FSHR* genes and encompass *LHCGR* gene.

A previously unreported 60kb heterozygous deletion on chromosome 20 (chr20:59,840,239-59,900,385) was found in proband NCL_MI_0042P (Figure 8.4). This deletion partially encompasses the genes *PHACTR3* (pLI=1) and *SYCP2* (pLI=1). The *SYCP2* gene's LoF observed/expected (o/e) fraction is 0.25 with a LoF o/e upper bound fraction (LOEUF) of 0.36 (<0.6, gnomAD SV v4.1.0 database). *SYCP2* is a critical component of the synaptonemal complex during meiosis, essential for chromosomal synapsis and recombination (GeneCards).

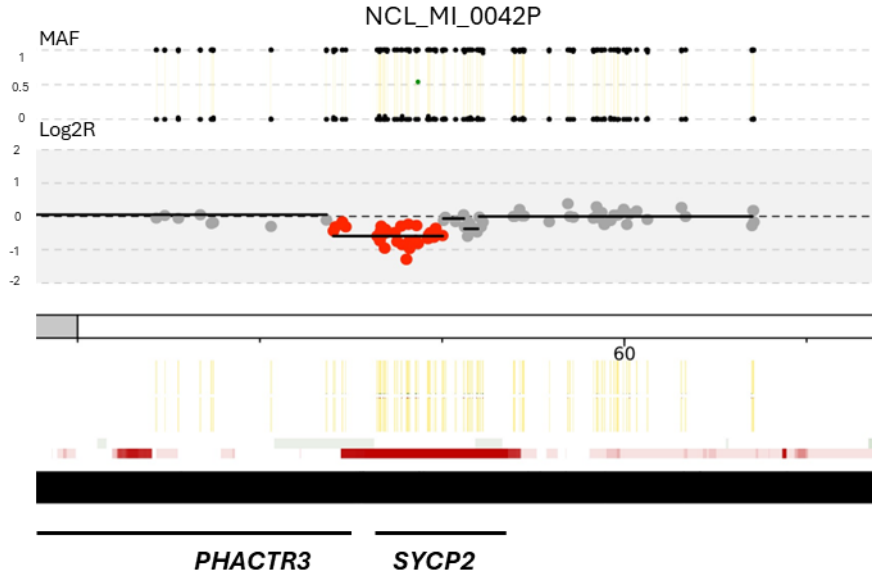


Figure 8.4. CNVRobot plot of the rare 60kb deletion detected in patient NCL_MI_0042P. The log2R value of -1 indicates heterozygous deletion. The deletion removes the part of *SYCP2* gene.

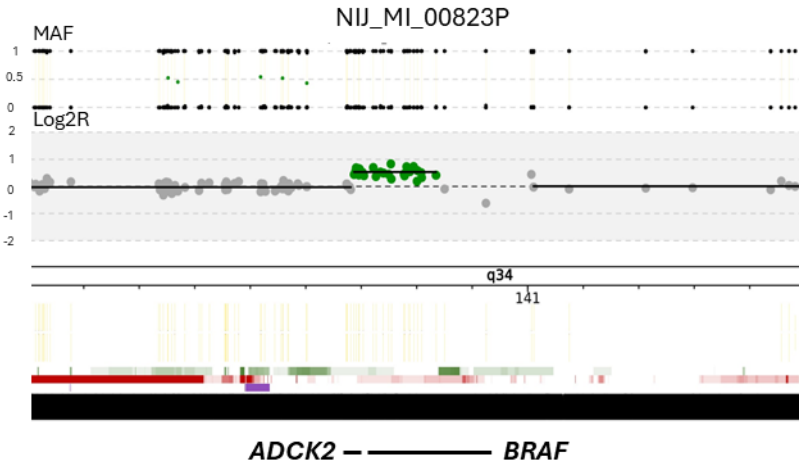


Figure 8.5. CNVRobot plot of the 148kb duplication detected in patient NIJ_MI_00823P. An expected log2ratio value of approximately 0.58 indicates a heterozygous duplication. The duplication contained a breakpoint within the coding region of the *BRAF* gene, with at least one exon lying outside the duplication boundaries.

Patient NIJ_MI_00823P with azoospermia as well as cryptorchidism harboured a 148kb duplication on chromosome 7 (chr7:140,686,889-140,834,973) (Figure 8.5), involving the *BRAF*, *ADCK2*, and *NDUFB2* genes. The *BRAF* gene was disrupted by the one of the breakpoints. The similar size duplication within the same region have been reported in 5 individuals (2 of those are males) in GnomAD SVs v4.1.0 (ID: DUP_CHR7_EAA6B3B0, AF= 0.00003965). *BRAF* has a high pLI score of 1.0 and LoF observed/expected (o/e) fraction is 0.15 with a LoF o/e upper bound fraction (LOEUF) of 0.24 (<0.6, gnomAD SV v4.1.0). The *BRAF* gene is a serine/threonine kinase involved in the MAPK/ERK signalling pathway, which regulates cell growth and differentiation. While somatic mutations in *BRAF* are known in various cancers, germline mutations are associated with cardio-facio-cutaneous syndrome (OMIM:115150) and Noonan syndrome 7 (OMIM:613706). Also, Houston et al., 2021, concluded that the *BRAF* gene is associated with cryptorchidism and azoospermia under the category of abnormal development of reproductive organs with limited evidence.

In patients NIJ_MI_00105P and NIJ_MI_00380P with azoospermia, heterozygous deletions in the pseudoautosomal region (PAR) of the X chromosome were identified. In NIJ_MI_00105P, a 237kb deletion (chrX:155,773,771-156,010,510) encompassing *SPRY3*, *VAMP7*, and *IL9R* was found (Figure 8.6.A). In NIJ_MI_00380P, a 113kb deletion (chrX:155,898,011-156,010,510) affecting *VAMP7* and *IL9R* genes was detected (Figure 8.6.B). The LoF deletion within the same region encompassing *VAMP7*, and *IL9R* genes have been reported in 58 individuals (57 of those are males) in GnomAD SVs v4.1.0 (ID: DEL_CHRX_2F95BBA0, AF= 0.0004600). Additionally, 8 other LoF deletions reported in many individuals in GnomAD SVs v4.1.0. Moreover, the association with male infertility is linked to an increased copy number of the *VAMP7* gene. Specifically, an elevated gene copy number of the vesicle SNARE *VAMP7* disrupts male urogenital development by altering estrogen signalling (Tannour-Louet et al., 2014). These deletions in patients are not considered to be causal for the observed phenotypes.

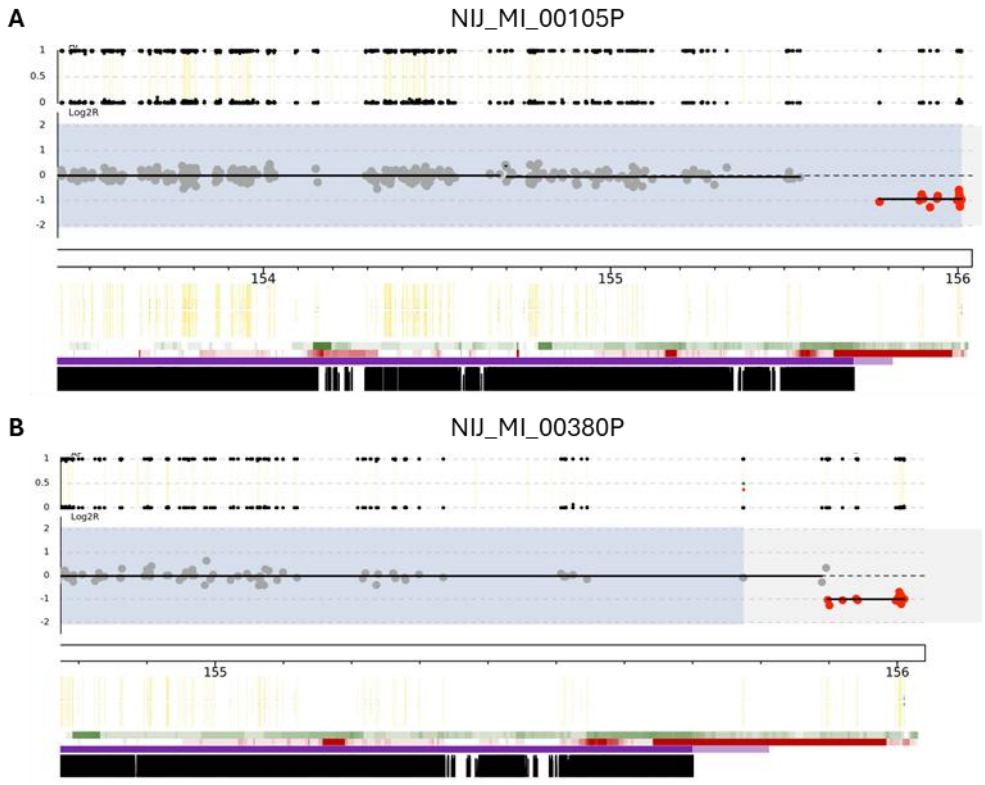


Figure 8.6. A. CNVRobot plot of the 237kb heterozygous deletion detected in patient NIJ_MI_00105P on chromosome X PAR region. B. CNV plots of a 113kb heterozygous deletion identified in patient NIJ_MI_00380P on chromosome X PAR region. The log2R value of -1 indicates heterozygous deletion.

Lastly, a homozygous 115kb deletion on chromosome 10 (chr10:133,453,871-133,568,549) was identified in proband NCL_MI_0179P with azoospermia (Figure 8.7). The region where the deletion occurred is a rearrangement hotspot on chromosome 10. The region was reported as deleted in males and females in GnomAD (ID: DEL_10_115557, AF=0.001521), but all are in heterozygous states. This deletion includes the genes *SYCE1, SCART1, CYP2E1* and *SPRNP1*. The *SYCE1* is an essential component of the synaptonemal complex transverse filaments, crucial for chromosomal synapsis during meiosis (GeneCards).

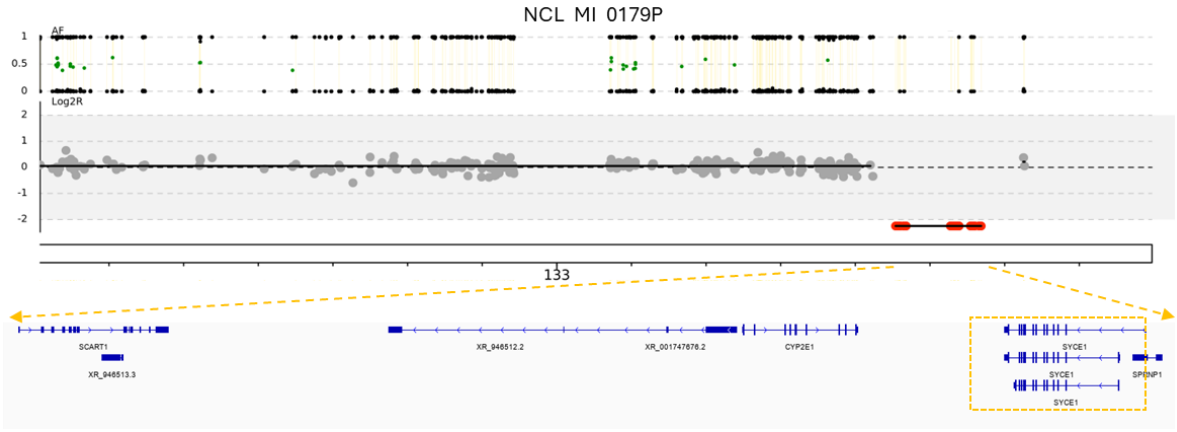


Figure 8.7. CNVRobot plot of the 115kb deletion detected in patient NCL_MI_0179P. The log2R value of -2 indicates homozygous deletion. The homozygous deletion removes the entire *SYCE1* gene in two alleles.

8.3.2 Uncovering Novel Candidate Genes

There were 659 rare deletions and 948 rare duplications. To uncover novel candidate genes, I applied specific filtration criteria based on inheritance patterns, CNV characteristics, and gene constraint scores. For the autosomal dominant (AD) inheritance model, I first selected rare CNVs larger than 5kb, resulting in 609 deletions and 874 duplications. These were further filtered to include only CNVs overlapping genes with a pLI score greater than 0.9, narrowing the numbers to 99 deletions and 140 duplications. Applying an additional filter for a DOMINO score exceeding 0.5 identified 4 deletions and 13 duplications. For the recessive inheritance model, I looked for rare homozygous CNVs, but there was no rare homozygous deletion. Additionally, all rare gonosomal CNVs (2 deletions on chromosome Y) were analysed. Following these filtration steps, CNV plots generated by CNVRobot were visually inspected to exclude false positives (3 deletions and 11 duplications removed). This prioritisation process ultimately identified 5 CNVs, comprising 3 deletions - 2 of which are on chromosome Y- and 2 duplications (Figure 8.8).

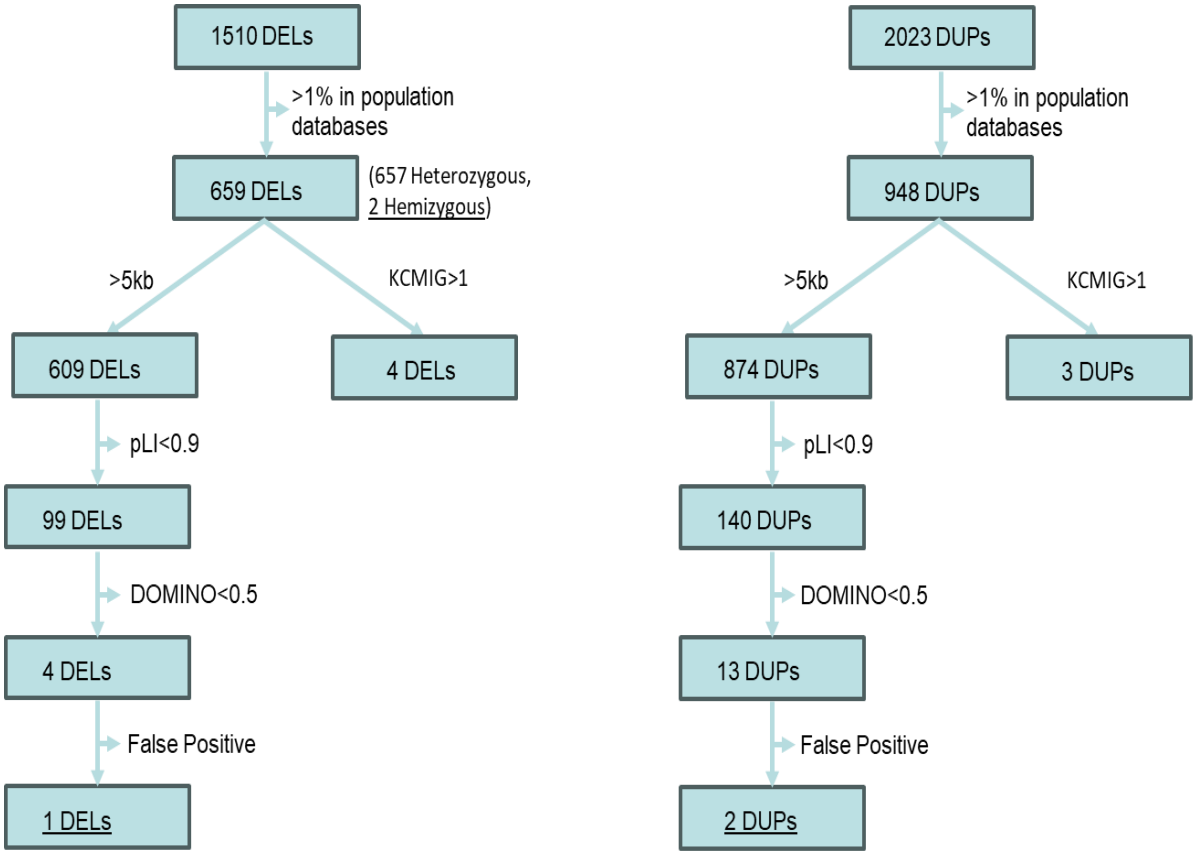


Figure 8.8. **The prioritisation steps for CNVs identified in 234 singletons.** Both DELs and DUPs were prioritised in two main steps: the first based on known and candidate genes, and the second employing an unbiased approach based on CNV size and constraint scores. KCMIG: Known and Candidate Male Infertility Genes, DELs: Deletions, DUPs: Duplications

Table 8.2 The prioritised CNVs in the analysis aiming to reveal novel candidate genes, detailing proband ID, CNV type, genotype, genomic coordinates, size, genes, genes with pLI >0.9 and KCMIG score for encompassed genes if exist.

Proband	Type	Genotype	Genomic Location (GRCh38)	Size (kb)	Genes	Genes with pLI >0.9	KCMIG Score (if exist)
NCL_MI_0008P	DUP	Het	chr18:21015294 -21029036	13	*ROCK1*	*ROCK1*	1
MAN_MI_0010P	DUP	Het	chr20:10408527 -10673650	265	*SLX4IP-MKKS-JAG1*	*JAG1*	1
MAN_MI_0013P	DEL	Het	chr11:1193289-11952966	20	*USP47*	*USP47*	-
NIJ_MI_00507P	DEL	Het	chr20:36770772 -36816408	45	*SOGA1-DSN1*	*SOGA1*	1C
MAN_MI_0007P	DEL	Hem	Y:6867936-7101071	233	*TBL1Y-AMELY*	-	*AMELY* - 1
SHF_MI_0015P	DEL	Hem	Y:18529700-22417700	3888	*HSFY1-TTTY9B-TTTY9A-HSFY2-TTTY14-BCORP1-TXLNGY-KDM5D-TTTY10-EIF1AY-RPS4Y2-PRORY-RBMY2EP-RBMY1B-RBMY1A1-TTTY13-RBMY1D-RBMY1E-PRY2-TTTY6B-RBMY1F-TTTY5-RBMY2FP-RBMY1J*	*KDM5D*	*EIF1AY (1)-HSFY1 (1)-RBMY1A1 (1)-HSFY2 (1B)-KDM5D (1C)-RBMY1E (1C)-RBMY1F (1C)-RBMY2FP (1C)-TTTY5 (1C)*

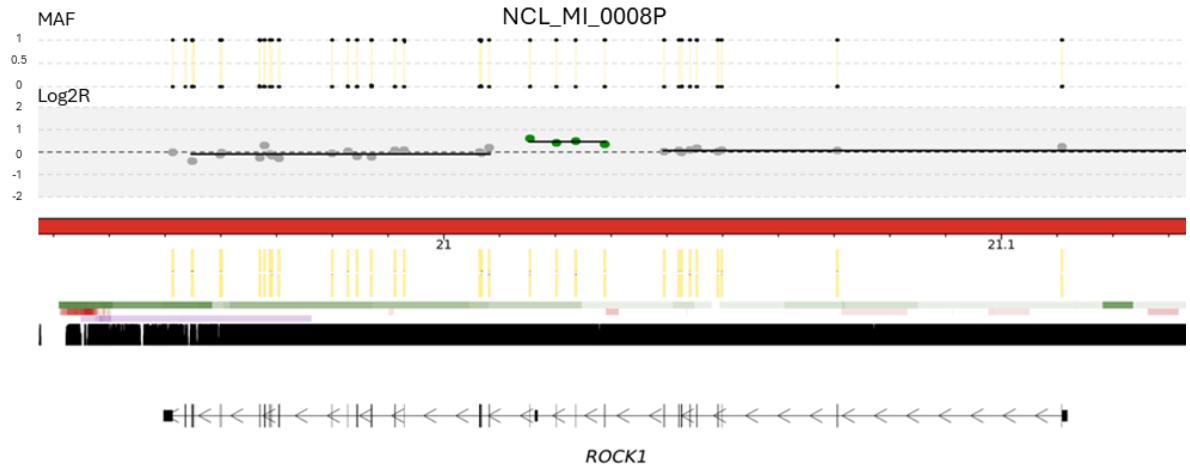


Figure 8.9. CNVRobot plot of the 13kb duplication detected in patient NCL_MI_0008P. An expected log2ratio value of approximately 0.58 indicates a heterozygous duplication. The duplication encompasses 4 exons of the *ROCK1* gene.

In patient NCL_MI_0008P, a previously unreported heterozygous 13kb duplication on chromosome 18 was identified, encompassing 4 exons within the *ROCK1* (Rho-associated, coiled-coil protein kinase 1) gene (pLI=1) (Figure 8.9). The *ROCK1* gene's LoF observed/expected (o/e) fraction is 0.2 with a LoF o/e upper bound fraction (LOEUF) of 0.29 (<0.6, gnomAD SV v4.1.0). No SVs were reported in the population databases in this gene (gnomAD SV v4.1.0). This gene is expressed in all tissues at RNA and protein level according to Protein Atlas Database. The *ROCK1* phosphorylates and activates *SOX9* in Sertoli cells to initiate testes formation (Mizuno et al., 2013). However, the role of the gene in spermatogenesis is unknown and no pathogenic mutations have been reported in patients with any form of male infertility yet.

Patient MAN_MI_0010P harboured a previously unreported heterozygous duplication of a 265kb region on chromosome 20 encompassing the genes *SLX4IP*, *MKKS*, and *JAG1* (Figure 8.10). One of the breakpoints occurred within the *JAG1* gene (pLI=1). No LoF SV reported in the gene and LoF observed/expected (o/e) fraction is 0.14 with a LoF o/e upper bound fraction (LOEUF) of 0.21 (<0.6, gnomAD SV v4.1.0), underscoring potential pathogenicity when dosage is altered. Mutations in *JAG1* are known to cause Alagille syndrome (OMIM:118450), characterised by hepatic, cardiac, and skeletal abnormalities. Furthermore, this gene was shown to be a candidate gene for hypogonadotropic hypogonadism (Quaynor et al., 2016).

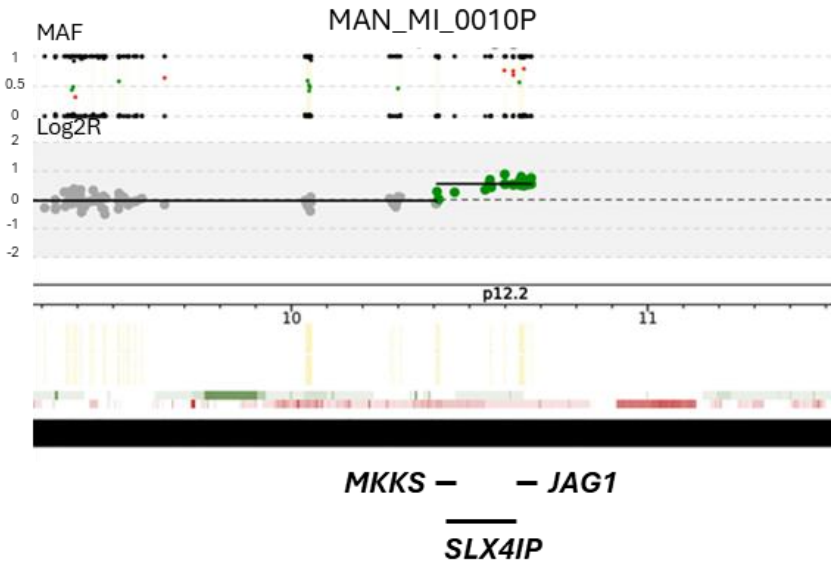


Figure 8.10. CNVRobot plot of the 265kb duplication identified in patient MAN_MI_0010P. An expected log2ratio value of approximately 0.58 indicates a heterozygous duplication. The duplication has breakpoints in the *JAG1* and *MKKS* genes.

A heterozygous deletion of a 20kb region at chromosome 11 was detected in patient MAN_MI_0013P, affecting the *USP47* gene. There was also one intronic *de novo* deletion identified in the trio cohort described in chapter 4. So, findings for this gene was fully described in chapter 4.

In patient NIJ_MI_00507P with azoospermia, I identified a heterozygous deletion of a 46kb region on chromosome 20 partially removing the genes *SOGA1* (exon5-15) (pLI=1) and *DSN1* (exon1-3) (pLI=0) in one allele (Figure 8.11). LoF observed/expected (o/e) fraction is 0.1 with a LoF o/e upper bound fraction (LOEUF) of 0.19 (<0.6, gnomAD SV v2.1). The *SOGA1* gene is implicated in glucose metabolism and autophagy regulation. While its role in human disease is not well characterised, the deletion may have phenotypic consequences due to the gene's predicted loss-of-function intolerance.

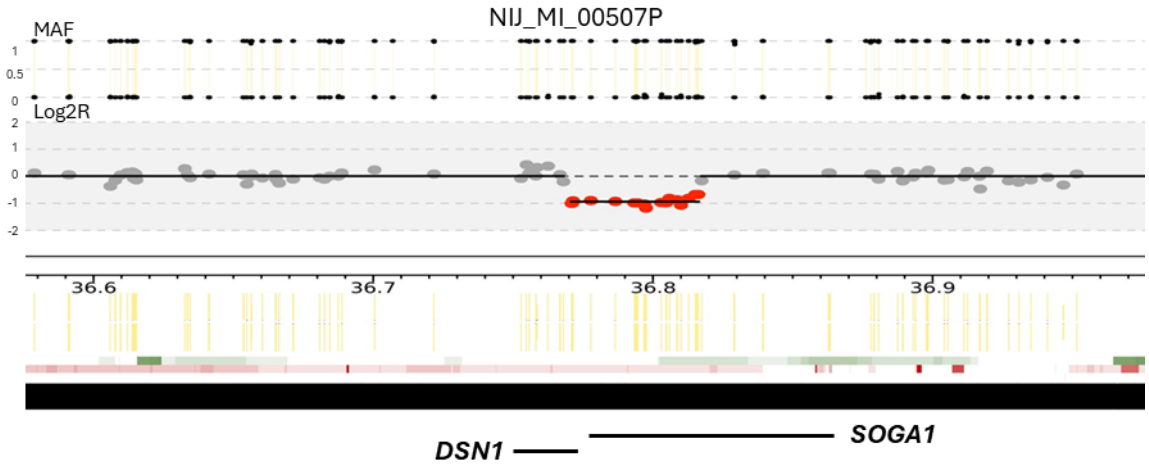


Figure 8.11. CNVRobot plot of the 46kb heterozygous deletion detected in patient NIJ_MI_00507P. The log2R value of -1 indicates heterozygous deletion. The deletion removes the part of *SOGA1* and *DSN1* genes in one allele.

Patient MAN_MI_0007P exhibits a hemizygous deletion of a 233kb region on Y chromosome Y, involving the genes *TBL1Y* and *AMELY* (Figure 8.12). Both gene have a pLI score of 0. *AMELY* encodes a protein involved in tooth enamel formation, but its expression is significantly lower than its X-linked counterpart *(AMELX)* and individuals with *AMELY* deletions often do not show noticeable dental issues due to *AMELX* can compensate (Jobling et al., 2007). The *TBL1Y* gene is a Y-linked homolog of *TBL1X*, it plays a role in cardiac differentiation, and its reduced expression has been linked to impaired heart development. This gene was associated with favourable lipid profiles in specific populations, however, its role in other tissues like the testes remains unclear (Colaco & Modi, 2018).

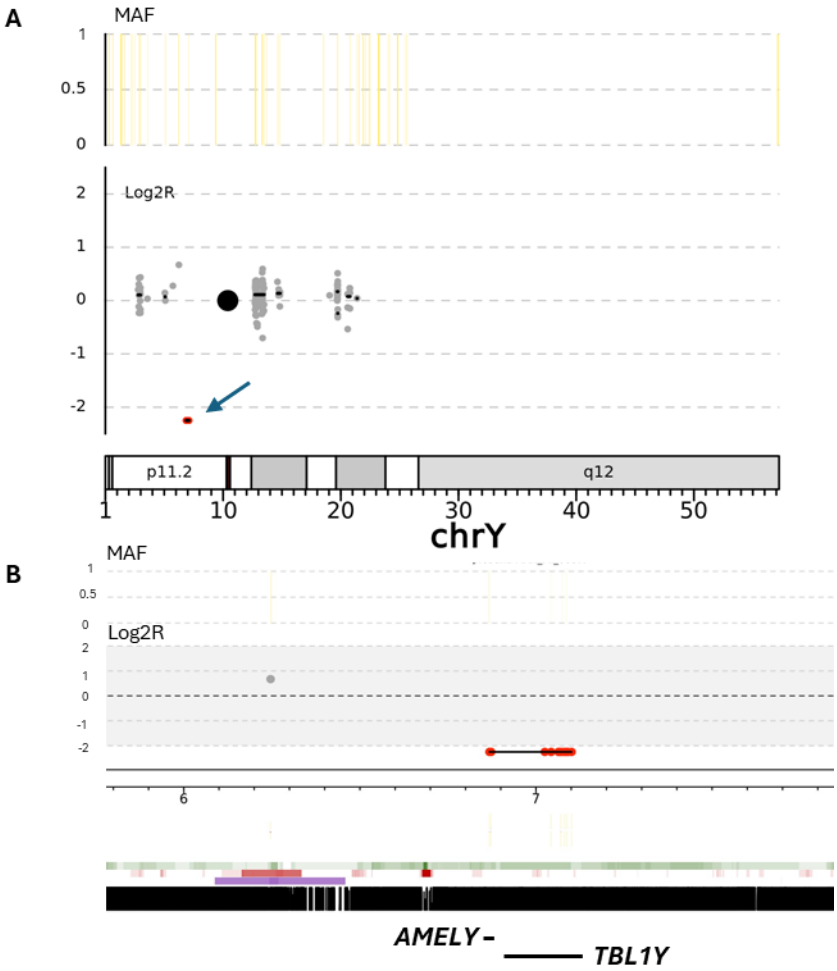


Figure 8.12. A. CNVRobot plot of entire chromosome Y of patient MAN_MI_0007P. The small red dot indicated with arrow represent the deletion. **B.** Zoom in on the hemizygous deletion detected in patient MAN_MI_0007P on chromosome Y.The log2R value of -2 indicates hemizygous deletion. The hemizygous deletion removes the one copy of *AMELY* and *TBLY1* genes.

Although AZF deletions are typically screened and used as exclusion criteria, in patient SHF_MI_0015P, a 3.9Mb hemizygous deletion, representing 65.75% of the AZFb region, was identified on the q arm of the Y chromosome (Figure 8.13). This deletion affects multiple genes, including *HSFY1, HSFY2, KDM5D, EIF1AY* and several members of the *RBMY* gene family.

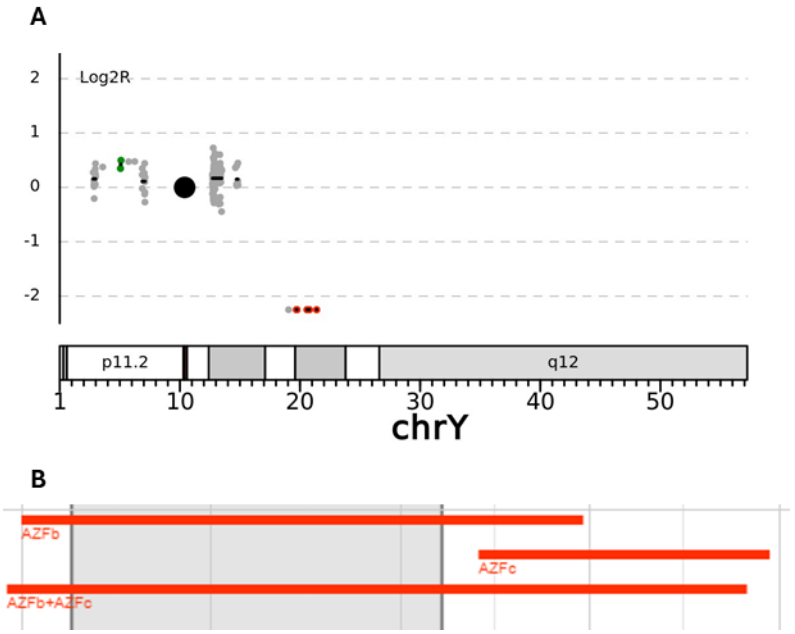


Figure 8.13. A. CNVRobot plot of entire chromosome Y of patient SHF_MI_0015P. The log2R value of -2 within the region marked by red dots (probes) indicates hemizygous deletion. **B.** This track displays AZFb and AZFc deletions on chromosome Y, highlighting the affected part, which constitutes 65.75% of the AZFb region. Possible breakpoints are indicated by grey lines.

It should be noted that several identified CNVs could potentially lead to the formation of novel fusion genes, particularly if the duplications are tandem. This possibility was identified for two duplications spanning the *CYP19A1-DMXL2* and *MKKS-JAG1* gene pairs, and a deletion involving *SOGA1-DSN1*. However, the analysis to confirm such fusion events was not performed. The CNVs in this study were identified using WES data, which lacks the resolution to precisely map breakpoints. Without this information, the exact structure of the rearrangements, including whether a duplication is tandem or has been inserted elsewhere, cannot be determined. Therefore, while the potential for fusion genes is acknowledged, their confirmation would require further investigation with WGS.

8.3.3 Replication Study of Candidate Genes

The genes (*BRAF, SYCP2, SYCE1, ROCK1, JAG1, USP47,* and *SOGA1*), which may be affected by CNVs and potentially contribute to the observed patient phenotypes, were further investigated in replication cohorts as well as fertile controls (see Chapter 2.5). I searched for potentially pathogenic or likely pathogenic LoF SNVs in these genes within the Genomics of Male Infertility Group cohort, the MERGE study from Germany, and the fertile control cohort.

In the *ROCK1* gene, one likely pathogenic splice site mutation *ROCK1*(NM_005406.3):c.175+2T>A was identified in singleton NIJ_MI_2343P with azoospermia from Nijmegen, and another likely pathogenic stop-gain mutation *ROCK1*(NM_005406.3):c.3784C>T p.(Arg1262Ter) was detected in proband NCL_MI_0037P with azoospermia from Newcastle, within our Genomics of Male Infertility Group cohort. However, this mutation was inherited from the father, questioning its impact on fertility. Additionally, in the MERGE cohort, one stop-gain mutation *ROCK1*(NM_005406.3):c.3371G>A p.(Trp1124Ter) was identified in M2834 with cryptozoospermia, and one splice site mutation *ROCK1*(NM_005406.3):c.959+1G>T was detected in M3192 with azoospermia, both in singleton cases. No potentially pathogenic mutations were identified in the control cohort.

In the *SYCP2* gene, two likely pathogenic frameshift and one likely pathogenic splice site mutations were identified in three different singleton cases. These include *SYCP2*(NM_014258.4):c.2759_2763del p.(Thr920ArgfsTer2) in NCL_MI_0076P with OAT from Newcastle, *SYCP2*(NM_014258.4):c.1395_1398del p.(Gln466LeufsTer16) in NIJ_MI_1878P with oligozoospermia from Nijmegen, and *SYCP2*(NM_014258.4):c.-46-1G>A in NIJ_MI_2192P with azoospermia from Nijmegen. Additionally, in the MERGE cohort, four likely pathogenic frameshift variants were identified in four different singletons: *SYCP2*(NM_014258.4):c.2793_2797del p.(Lys932SerfsTer3) in M1581 with cryptozoospermia, *SYCP2*(NM_014258.4):c.2022_2025del p.(Lys674AsnfsTer8) in M1686 with cryptozoospermia, *SYCP2*(NM_014258.4):c.3052dup p.(Ala1018GlyfsTer9) in M2477 with azoospermia, and *SYCP2*(NM_014258.4):c.1395_1398del p.(Gln466LeufsTer16) in M3066 with azoospermia. Furthermore, in the fertile control cohort, two likely pathogenic splice site mutations (*SYCP2*(NM_014258.4):c.514-1G>T and *SYCP2*(NM_014258.4):c.514-2A>T) were identified in two fertile fathers.

Three potentially pathogenic heterozygous SNVs were identified in three patients in the recessive *SYCE1* gene. No pathogenic LoF SNVs were identified in the remaining genes (*BRAF, JAG1, USP47,* and *SOGA1*) in the fertile control cohort.

8.4 Discussion

In this chapter, I analysed CNVs detected in a cohort of 234 patients diagnosed with azoospermia, aiming to identify genetic variants that could explain their infertility. I identified a total of 3533 CNVs (1,510 deletions and 2,023 duplications), with an average of 15 CNVs per sample. Notably, we found more duplications than deletions, which is the opposite of findings seen in WGS. This observation is likely a result of methodological biases inherent to WES rather than a true biological difference. WGS is susceptible to a higher rate of false-positive deletions due to data noise (largely from spurious split and discordant reads), whereas WES, which relies on read depth signatures in targeted regions, can generate false-positive duplications due to uneven read depth profiles. From the total CNVs, 1607 were rare CNVs (present in less than 1% of the population databases), which I prioritised for further analysis. I employed two complementary approaches to interpret these CNVs, a: screening for variants affecting known male infertility genes and b: uncovering novel candidate genes that might contribute to the disease. Through these methods, I identified several CNVs that could potentially explain the patients' phenotypes. Among the CNVs affecting known male infertility genes, I identified deletions and duplications involving *SYCP2, SYCE1, CYP19A1, LHCGR, FSHR*, and *BRAF* genes.

The 302kb rare heterozygous duplication on chromosome 15 detected in patient NIJ_MI_02282P encompasses the genes *GLDN, CYP19A1, and DMXL2* with one of the breakpoints occurring within the *CYP19A1* gene. The *CYP19A1* gene encodes aromatase, a critical enzyme in estrogen biosynthesis. Alterations in this gene was link to aromatase excess syndrome and gynecomastia (OMIM:139300). No clinical signs of aromatase excess syndrome or gynecomastia were reported for the individual in this study. While duplications of similar size are exceedingly rare in population databases (reported in only one individual in gnomAD CNVs v4.1.0, ID: 174261_DUP, allele frequency of 0.000002154 with unspecified gender) the impact of this duplication on gene function remains unclear. If the duplication is tandem and does not disrupt the gene structure, it may not affect *CYP19A1* function. Additionally, without detailed hormonal profiles or phenotypic information regarding the patient's endocrine system, it is challenging to interpret the potential effects of this duplication on estrogen levels

and its contribution to azoospermia. Estrogen plays a crucial role in male reproductive function, imbalances of this hormone can affect the process of spermatogenesis (Schulster et al., 2016). Therefore, further clinical evaluation including hormonal assays and functional studies are necessary to determine whether this duplication has pathogenic relevance in this patient's infertility.

The 60kb heterozygous deletion on chromosome 20 in patient NCL_MI_0042P encompasses *SYCP2* gene, a critical component of the synaptonemal complex essential for chromosomal synapsis during meiosis. The *SYCP2* was moderately associated with male infertility due to severe oligozoospermia (OMIM:258150) (AD) (Houston et al., 2021). Recently three unrelated Chinese patients with oligoasthenozoospermia were presented, all carrying heterozygous *SYCP2* frameshift variants (Li et al., 2024). We also identified seven patients carrying potentially pathogenic SNVs in the *SYCP2* gene in our replication study, which may account for the observed patient phenotypes. These new cases significantly strengthen the clinical validity of the *SYCP2* gene association with male infertility, advancing it from a limited-to-moderate level to a strong causal relationship between heterozygous loss-of-function variants in *SYCP2* and autosomal dominant male infertility.

Similarly, the homozygous 115kb deletion on chromosome 10 in patient NCL_MI_0179P removes the entire *SYCE1* gene, another essential component of the synaptonemal complex. The gene is provisionally related to spermatogenic failure in OMIM (616950). Houston et al., 2021, concluded that *SYCE1* has a limited relationship with NOA. Recently, two novel CNVs within the *SYCE1* gene were reported in two unrelated patients with NOA leading to meiotic arrest (Huang et al., 2022). Similarly, Wyrwoll et al., 2022, determined the cause of infertility in two individuals with homozygous deletions of *SYCE1* and in one individual with a heterozygous deletion combined with a likely pathogenic missense variant on the second allele. Furthermore, Feng et al., 2022, identified a novel homozygous frameshift mutation (c.689_690del; p.F230fs) that altered *SYCE1* expression patterns and blocked spermatogenesis. These new cases significantly strengthen the clinical validity of the association of *SYCE1* gene with male infertility. Thus, the homozygous deletion of *SYCE1* in this patient is likely a causative factor for his infertility.

In patient NIJ_MI_00823P, who presented with azoospermia and cryptorchidism, I identified a 148kb duplication on chromosome 7 involving genes *BRAF*, *ADCK2*, and *NDUFB2*. One of the breakpoints disrupted the *BRAF* gene. The *BRAF* gene encodes a serine/threonine kinase in

the MAPK/ERK signalling pathway, crucial for cell growth and differentiation, and germline mutations in this gene are associated with cardio-facio-cutaneous syndrome and Noonan syndrome 7 (conditions that include cryptorchidism). However, the patient did not present with the other characteristic features of these syndromes, which may be because the syndromes are typically caused by specific point mutations, whereas the functional impact of a duplication could be distinct and lead to a more restricted phenotype. Although similar duplications are rare in population databases and the reproductive health status of those individuals is unknown, the patient's phenotype aligns with features associated with *BRAF* mutations as noted in the Houston et al. (2021) review. This suggests that this duplication could explain his condition. However, if the duplication is tandem and does not disrupt the gene structure, it may not affect *BRAF* function. On the other hand, since the breakpoint occurs within the coding region, there is a potential for altered gene function. Our case could contribute to valuable evidence to the limited data supporting the association between *BRAF* and male infertility. To confirm the pathogenicity of this duplication, a series of further investigations are necessary. Initially, WGS could precisely map the breakpoints, followed by segregation analysis within the family. Submitting the variant to a matchmaking database like GeneMatcher (Sobreira et al., 2015) could identify additional cases, and ultimately, functional studies using CRISPR in cell lines or a model organism would be required to understand its definitive impact on male reproductive health.

As we were pursuing novel candidate genes, I identified duplications and deletions involving *ROCK1, JAG1 and SOGA1* genes and regions of the Y chromosome. The heterozygous duplication of 4 exons in *ROCK1* gene in patient NCL_MI_0008P is noteworthy, as *ROCK1* phosphorylates and activates SOX9, a transcription factor essential for Sertoli cell function and testis development (Mizuno et al., 2013). Additionally, four potentially pathogenic SNVs were identified in four patients within the replication study, one of which was paternally inherited, whereas none were detected in the control cohort. While the role of *ROCK1* in spermatogenesis is not fully understood, its involvement in testis formation suggests that alterations in *ROCK1* dosage could impact male fertility. To investigate this, functional studies could involve cell-based assays to determine if the duplication or identified SNVs in *ROCK1* lead to aberrant phosphorylation and activation of its target, *SOX9*.

In patient MAN_MI_0010P, the duplication involving *JAG1* with a breakpoint within the gene may have phenotypic consequences. *JAG1* encodes a ligand in the Notch signalling pathway

which plays a role in gonadal development and function (Quaynor et al., 2016). While mutations in *JAG1* are known to cause the multi-system disorder Alagille syndrome (Li et al., 1997), no features of this syndrome were reported in the patient, and the gene's potential role in hypogonadotropic hypogonadism suggests that disruptions in *JAG1* could contribute to isolated infertility. To better determine the pathogenicity of this duplication, precisely mapping the breakpoints and performing segregation analysis within the family would be critical initial steps. Furthermore, submitting the variant to a matchmaking database like GeneMatcher (Sobreira et al., 2015) could help identify other individuals with similar *JAG1* disruptions.

The other heterozygous deletion involving *SOGA1* in patient NIJ_MI_00507P may have phenotypic effects. The *SOGA1* gene is implicated in glucose metabolism and autophagy regulation. It was also shown that it is required for chromosome segregation (Ferreira et al., 2021). Given the importance of metabolic processes in spermatogenesis and chromosome segregation, alterations in *SOGA1* could potentially impact fertility. However, the specific role of *SOGA1* in male reproduction remains to be investigated. Further clinical correlation and functional assays are required to elucidate the potential impact of this genetic alteration.

The hemizygous deletion on the Y chromosome in patient SHF_MI_0015P affect genes important for spermatogenesis. The deletion encompassing a significant portion of the AZFb region is likely pathogenic, as deletions of this region are known to cause azoospermia due to meiotic arrest (Krausz & Casamonti, 2017). The genes in the AZFb locus are crucial for sperm development, particularly for the progression of sperms through meiosis into spermiogenesis. Patients with AZFb deletions typically exhibit a testicular phenotype of maturation arrest, often at the spermatocyte stage with a lack of post-meiotic germ cells. The severity of testicular findings correlates with the extent of the AZFb deletion with complete deletions causing more severe spermatogenic block than partial deletions. Although the chance of retrieving mature sperm in cases of AZFb deletions is very low, rare instances of sperm retrieval were reported especially in cases of partial deletions (Colaco and Modi, 2018).

Collectively, our analysis identified several CNVs that potentially contribute to male infertility in our cohort. Deletions involving *SYCP2* and *SYCE1* are strong candidates due to their essential roles in meiosis. Duplications affecting *CYP19A1, LHCGR, FSHR* and *BRAF* genes may influence hormonal regulation and developmental pathways critical for spermatogenesis. The

identification of novel candidate genes such as *ROCK1* and *JAG1* expands our understanding of the genetic landscape of male infertility.

While functional studies are necessary to confirm the pathogenicity of these CNVs and elucidate the mechanisms by which they impact fertility, such investigations were beyond the scope of the current project due to time and monetary constraints. Additionally, larger studies and international collaborations are needed to identify recurrently mutated genes and to validate novel findings. Our participation in the International Male Infertility Genomics Consortium (IMIGC) provides a platform for such collaborative efforts, which are essential for tackling the high genetic heterogeneity of male infertility.

8.5 Conclusion

In this chapter, I analysed CNV profiles of 234 patients diagnosed with azoospermia. Although parental samples were not available to determine the inheritance patterns of the identified CNVs, I successfully identified several rare deletions and duplications affecting both known male infertility genes such as *SYCP2* and *SYCE1* and novel candidate genes like *ROCK1* and *JAG1*. The analysis also uncovered a deletion on the Y chromosome affecting critical regions like AZFb. These findings highlighted the genetic complexity underlying male infertility and confirmed that whole-exome sequencing is an effective method for detecting CNVs in patient-only cohorts. Our study underscored the value of investigating CNVs in singletons and emphasised the need for future functional studies to validate these candidate genes and elucidate their roles in spermatogenesis which could lead to improved genetic counselling and personalised therapeutic strategies.

Chapter 9. General Discussion

9.1 WGS is an Effective Tool to Identify Structural Variations

Traditionally, SVs have been detected using karyotyping and hybridisation-based assays. Although karyotyping offers relatively low resolution, it remains in clinical use primarily due to its low cost (Balachandran & Beck, 2020). However, emerging next-generation sequencing (NGS) technologies and advanced computational algorithms are reshaping SV detection and steadily replacing older methods (Korbel et al., 2007; Bentley et al., 2008; McKernan et al., 2009; Alkan et al., 2011). Multiple studies have demonstrated that WGS provides better diagnostic accuracy compared to chromosomal microarray (CMA), karyotyping, and other targeted assays by detecting diverse SV types at base-pair resolution as well as non-coding SNVs (Meng et al., 2023; van der Sanden et al., 2023; Pagnamenta et al., 2023; Collins & Talkowski, 2025).

WGS methodologies can rely on short-read or long-read platforms, each with distinct advantages and disadvantages (Chaisson et al., 2019). Integrating multiple platforms improves access to previously challenging genomic regions and enables the generation of fully phased genomes (Altemose et al., 2022). For example, an integrated approach yielded a gapless, telomere-to-telomere human genome assembly, illustrating how a more complete reference genome can enhance our understanding of human genetic variation, particularly SVs (Aganezov et al., 2022). Despite these significant advances in variant discovery, supported by population-level and functional genomic resources (Collins et al., 2020), high costs and specialised expertise may impede broader global implementation. Alongside platform innovations, developments in bioinformatic tools have been equally important. New pipelines increasingly use combinatorial methods to improve accuracy and reduce false positives in SV calling (Zhao et al., 2013).

In this study, we chose to combine CNVRobot and dysgu-sv, as this combination allowed for the detection of all SV types with single base pair breakpoint resolution and high sensitivity (see chapter 3). Additionally, to further optimize our pipeline and assess performance of recently developed OGM in SV detection, we studied nine samples from our male infertility cohort using both WGS and OGM. Given OGM's reported low false-positive rate (Dremsek et al., 2021), we used CNVs overlapping between OGM and srWGS as a 'gold standard' to evaluate the performance of two srWGS SV callers, CNVRobot and dysgu-sv. Our analysis

revealed that combining these tools enhanced detection sensitivity for these presumed true CNVs. Additionally, acknowledging the need for long-read sequencing to gain even deeper insights for SVs, our group recently initiated a pilot study in male infertility using long-read WGS for selected patient-parent trios. Our data suggests this approach will provide more sensitive and specific SV detection, thereby advancing our overall understanding of the genetic architecture underlying these conditions.

9.2 Novel Candidate Genes in Azoospermia and Severe Oligozoospermia

In this project, method optimisation for SV detection in WGS data was conducted, followed by analysis of WGS data from 216 patient-parent trios. The study looked to identify pathogenic SVs potentially explaining patients' azoospermia or oligozoospermia. Different inheritance patterns were explored, leveraging the patient-parent trio data, with a primary focus on dominant inheritance. In Chapters 5 and 6, I focused on *de novo* and maternally inherited SVs, followed by recessive inheritance in Chapter 7. In addition to patient-parent trios, CNV analysis was performed on a cohort of 234 patients without parental data.

De novo SV analysis uncovered several candidate genes, including *NXT2, USP47, BHLHE40*, and *MAZ*. The only exonic *de novo* deletion identified occurred on the X chromosome, encompassing the *NXT2* gene. *NXT2* exhibits moderate protein expression levels and high RNA expression levels in the testis. Additionally, the gene is evolutionarily conserved in eutherian mammals, and a previous study has indicated that it plays a non-essential role in mouse fertility (Khan et al., 2018). Our group collaborated with the IMIGC, and Dicke et al. (2024) conducted a study investigating the specific function of *NXT2* in nuclear RNA export within the human testis. This study aimed to explore the molecular interactions of *NXT2* and its role in male fertility. With additional cases harbouring pathogenic LoF mutations and functional work, it has been shown that *NXT2* is essential for normal germ cell development in humans (Dicke et al., 2024). Another candidate, *USP47*, was possibly impacted by a *de novo* intronic deletion of unknown significance. Additionally, a 20 kb heterozygous deletion encompassing eleven exons of *USP47* was identified in a singleton case with oligozoospermia. The *USP47* gene encodes a ubiquitin-specific protease that acts as a regulator of cell growth and genome integrity. The gene has low tissue specificity, according to the Protein Atlas database. *A de novo* mutation in the Y chromosome gene *USP9Y*, which similarly encodes a ubiquitin-specific protease, has been reported in a man with azoospermia (Sun et al., 1999). The identified SVs may explain patients' phenotypes, further validation is required. A minigene splicing assay

could first clarify the impact of the intronic deletion, while identifying additional cases via platforms like GeneMatcher (Sobreira et al., 2015) would be crucial to strengthen the gene-disease association. If more evidence is gathered, creating a knockout mouse model would be useful to definitively establish the role of *USP47* in spermatogenesis. The candidate gene *BHLHE40*, located near the largest *de novo* deletion identified, encodes a transcription factor attributed to play a role in Sertoli cell signalling and Sertoli gene activation (Fice & Robaire, 2023). One heterozygous frameshift variant, classified as likely pathogenic, was identified in the *BHLHE40* gene in the replication cohort, but none in fertile controls. While the presence of two cases with a similar phenotype is significant, further work is needed. A crucial next step could be to submit *BHLHE40* to a matchmaking database such as GeneMatcher (Sobreira et al., 2015) to identify recurrent cases in independent cohorts. To subsequently investigate the gene's function, CRISPR-Cas9 could be used to knock down *BHLHE40* in a Sertoli cell line, followed by RNA-sequencing to identify its downstream gene targets and clarify its role in Sertoli cell signalling. Lastly, a 525 kb *de novo* deletion encompassing 29 genes, including *MAZ*, was identified in one of the German trios (where WES was performed). This deletion has been previously reported and linked to severe phenotypes like intellectual disability but with incomplete penetrance (Auwerx et al., 2024). The lack of severe phenotypes in this patient may reflect incomplete penetrance, though *MAZ* haploinsufficiency could contribute to fertility issues.

Analysis of maternally inherited SVs revealed several interesting genes (*UBR2, AKT1, IPO9, FAS, ACTA2, FKBP8, TEKT5, NRK, SYAP1, and DYNLT3*). Although beyond the scope of this study due to time and funding constraints, both replication studies in larger cohorts and a variety of functional studies tailored to each gene and mutation are needed to elucidate their potential roles. Among these interesting genes, *UBR2* and *AKT1* emerged as particularly interesting, as possibly pathogenic LoF SNVs were identified in the replication study. I identified a previously unreported MI tandem duplication in a patient, encompassing the last five exons of the *UBR2* gene. Although the gene structure remains intact, the duplication of distal enhancers within the duplicated region and potential changes in 3D structure may affect the gene. *UBR2* gene encodes an E3 ubiquitin-protein ligase, a component of the N-end rule pathway. GO annotations for biological processes indicate that this gene is involved in male meiosis I (GO:0007141), male meiotic nuclear division (GO:0007140), reciprocal meiotic recombination (GO:0007131), and spermatogenesis (GO:0007283). Houston et al. (2021) concluded that the

gene is associated with NOA with limited evidence. The involvement of *UBR2* in critical meiotic and spermatogenic pathways underscores its importance in male fertility. Genetic disruption could potentially cause meiotic arrest or defects in spermiogenesis, contributing to the observed infertility in our patients. Further studies and larger cohorts are needed to provide additional evidence for a causal role of mutations affecting this gene in male infertility. Seeking additional cases through collaborations and gene-matching platforms like GeneMatcher (Sobreira et al., 2015) would be a valuable next step. While acknowledging the difficulty in obtaining patient tissue to investigate the direct impact of the identified duplication, the fundamental role of *UBR2* in meiotic and spermatogenic pathways can be investigated by creating an in vitro model. For example, using CRISPR-Cas9 to disrupt *UBR2* function in a relevant cell line would allow for the assessment of key cellular consequences, such as defects in meiotic progression, increased rates of apoptosis, or altered expression of critical spermatogenesis-related genes. A 697 kb deletion on chromosome 14, encompassing 12 genes including *AKT1*, was also identified. *AKT1* is a serine/threonine kinase that plays a key role in regulating cell survival, insulin signalling, angiogenesis, and tumour formation. The gene has been associated with several cancers, including breast, ovarian, and colorectal cancers (OMIM:164730). Koppers et al. ,2011, identified *AKT1* as a crucial factor for spermatozoa survival in humans, proposing that its phosphorylated state prevents cells from following the default apoptotic pathway (Aitken, 2018; Koppers et al., 2011). Additionally, the *AKT1* gene was shown to be vital for normal spermatogenesis in mice (Kim et al., 2012). Hence, the deletion may disrupt *AKT1*, leading to the NOA observed in the patient. Furthermore, a likely pathogenic splice site mutation in the *AKT1* gene was identified in a patient with azoospermia attributed to meiotic arrest in the replication study. These findings underline the need for further studies to clarify the role of *AKT1* in human spermatogenesis.

From the analysis in Chapter 7, where we investigated recessive inheritance, the *GRB14* gene emerged as a candidate A homozygous deletion disrupting a single exon of *GRB14* was identified in a proband from a consanguineous family displaying multiple cnnLOH regions. The *GRB14* gene encodes an adaptor protein that modulates insulin receptor signalling and is expressed in the liver and testis, yet no clear role in human spermatogenesis has been established. Although *GRB14* displayed interaction with *FGFR1*, an established azoospermia gene, the absence of a known direct link to testicular function complicates its interpretation. The *GRB14* finding should be considered preliminary, warranting further investigation into its

potential role in human spermatogenesis and male fertility. A comprehensive follow-up study could use CRISPR-Cas9 to assess the broader functional impact of *GRB14* loss in a human germ cell line, while also using co-immunoprecipitation to investigate its specific mechanistic link to spermatogenesis through its interaction with *FGFR1*.

In the patient-only cohort, where WES was performed on 234 singletons, the candidate genes *BRAF, SYCP2, SYCE1, ROCK1, JAG1,* and *SOGA1* were identified. Four possibly pathogenic SNVs were identified in the *ROCK1* gene in the replication study, one of which was inherited from the father, while the origins of the others remain unknown. No potentially pathogenic mutations were identified in the *ROCK1* gene in the control cohort but one predicted pathogenic mutation was paternally inherited and can be considered as control. In the *SYCP2* gene, seven possibly pathogenic SNVs were identified in the replication study, with two potentially pathogenic SNVs detected in two fertile fathers. *SYCP2* and *SYCE1* genes were previously associated with male infertility, moderately and limitedly, respectively. With additional cases from the literature, the homozygous deletion encompassing *SYCE1* as well as the heterozygous deletion encompassing *SYCP2* identified in two patients provide a clear diagnosis for these patients. These new cases significantly strengthen the clinical validity of the association of these genes with male infertility. The remaining candidate genes require a multi-step validation approach. The first step is to precisely map the breakpoints of any CNVs and perform segregation analysis to determine their inheritance. Subsequently, replication in larger cohorts, facilitated by international collaborators or platforms like GeneMatcher (Sobreira et al., 2015), is needed to strengthen the gene-disease validity. Finally, tailored functional studies are necessary to elucidate the underlying pathogenic mechanisms.

It also must be noted that one of the limitations of the current study, and of most genetic studies in this field, is the exclusive reliance on DNA isolated from blood. This approach cannot detect somatic variants or mosaicism confined to the germline, where a pathogenic variant could be present in the testicular tissue but absent in peripheral blood. Therefore, a promising direction for future research is to analyse DNA from testicular tissue itself. A practical and ethically sound source for this material would be the residual testicular tissue often discarded after a TESE procedure. Analysing this tissue would allow for the detection of somatic variants, potentially increasing the diagnostic yield and uncovering a new class of genetic causes for male infertility.

9.3 NGS Should be Implemented into Male Infertility Diagnostic and Research

The advent of NGS in male infertility assessment initially focused on targeted gene panels, in which hundreds of known or suspected disease-related genes could be analysed in parallel (Xavier et al., 2021). Xavier et al., 2021, highlighted that in 2019, nearly two-thirds (62%) of genomic research focused on male infertility incorporated NGS-based approaches. Since the cost of NGS technology continues to decline WES has become increasingly common in male infertility research (Fakhro et al., 2018; Oud et al., 2020; Stallmeyer et al., 2024). Specifically, two major cohorts, established through the Genetics of Male Infertility Initiative (GEMINI) and the Male Reproductive Genomics (MERGE) studies, have significantly contributed to the identification of novel candidate genes associated with male infertility using WES approaches (Wyrwoll et al., 2020; Salas-Huetos et al., 2021; Hardy et al., 2021). The GEMINI cohort comprises around 1,200 azoospermic patients, while the MERGE study includes around 1,000 infertile men. The MERGE cohort was also used in this study as a replication cohort, highlighting again the importance of large-scale cohorts and collaborative efforts in advancing research. Despite the valuable contribution of these datasets, their primary focus has been on SNV analysis, and to date, no report has been published on CNVs using these extensive cohorts. In Chapter 8, where we performed WES on singletons, we identified a homozygous deletion encompassing the well-established recessive male infertility gene *SYCE1*, as well as a large deletion involving the AZFb region of the Y chromosome (in a patient from the centre where AZF screening is not diagnostically performed). Both findings provide clear diagnoses, demonstrating the relevance of WES-based CNV studies in male infertility diagnostics. The implementation of WES in routine diagnostics for male infertility still lags behind other medical fields. One major factor contributing to this delay is the limited number of validated gene-disease associations. Many candidate genes currently lack robust evidence, and identifying variants within these genes could lead to false-positive results or inconclusive findings (Houston et al., 2021).

In response to these challenges, collaborative efforts have increased among researchers worldwide, most notably through the IMIGC (http://www.imigc.org/). The IMIGC has initiated large-scale studies and fostered data-sharing, enabling more rigorous validation of gene-disease relationships. By consolidating both genetic and clinical data, these collaborations help refine the set of diagnostically relevant genes, improving the reliability of results and informing clinical decision-making (Oud et al., 2021; Nagirnaja et al., 2022; Riera-Escamilla et

al., 2022). Furthermore, standardised phenotyping protocols and guidelines for the use of genomic techniques in clinical settings have been proposed, enhancing consistency and reproducibility across research groups and diagnostic laboratories (Wyrwoll et al., 2024). An example of the value of collaboration was shown by Stallmeyer et al., 2024. They systematically reviewed how WES has contributed to identifying novel disease genes linked to isolated (non-syndromic) male infertility. Their comprehensive literature search highlighted the complexity of conditions such as azoospermia and oligozoospermia (reduced sperm count) and asthenozoospermia and teratozoospermia (impaired sperm motility and/or morphology), with well over 100 candidate genes described for each disorder. By applying the standardised evaluation criteria of the ClinGen Gene Curation Working Group, they identified 70 genes with at least moderate evidence of contributing to male infertility. They suggested that including these validated genes in clinical exome sequencing will likely increase diagnostic yield and facilitate more accurate counselling for affected individuals.

Beyond WES, WGS offers the added advantage of detecting non-coding variants and providing better characterization of SVs. Both *de novo* SVs and homozygous deletions identified in this thesis may point to interesting non-coding regions potentially involved in male infertility. These regions warrant further exploration, as they could play a significant role in the disease. Unfortunately, large-scale WGS studies have not yet been undertaken by our collaborators, limiting our ability to investigate the replication of these non-coding SVs in this thesis. This study is the first to apply WGS to male infertility patient-parent trios, with a detailed analysis of SVs. A significant portion of the cohort analysed in this thesis had also been previously examined using WES (Oud et al., 2021). Although a systematic comparison was not conducted, I was able to detect balanced SVs, SVs in non-coding regions, and determine breakpoints at single-base resolution capabilities that are not possible with WES. Exemplifying, as discussed in Chapter 5, a *de novo* deletion encompassing the *NXT2* gene was initially identified as 7 kb in size in WES (Oud et al., 2021) but was found to be 42 kb after srWGS in this study.

Nevertheless, due to the lower costs and the current limited knowledge of non-coding variation, WES remains the more commonly used technique for disease-gene identification and genetic diagnosis in male infertility. It is highly recommended that WES be adopted as the first-line genetic test (Fakhro et al., 2018; Wyrwoll et al., 2023). As sequencing expenses continue to decrease and our understanding of non-coding regions grows, WGS is expected to become more widely integrated into basic research and clinical diagnostics.

In summary, it is clear that applying diagnostic exome or genome sequencing to individuals with severe, unexplained male infertility could offer a significant improvement in curating gene-disease associations and diagnostic yield, potentially adding 5-10% above current clinical standards (Stallmeyer et al., 2024). Expanding the use of NGS within routine practice also require ongoing collaboration and rigorous gene curation. These steps will ultimately pave the way for enhanced patient management, personalised treatment options, and improved outcomes in the field of male infertility.

9.4 Future Directions of Male Infertility Diagnostics

This section centres on two fundamental questions: why is genetic diagnosis important, and how can it be improved? Understanding the underlying genetic causes of male infertility is important, as it not only clarifies the aetiology of the condition but also guides clinical management. For example, genetic testing such as AZF deletion screening on the Y chromosome has already demonstrated its value in clinical decision-making. Complete deletions in the AZFa, AZFb, or AZFc regions serve as strong indicators of the likelihood of successful sperm retrieval via TESE. Specifically, men with a complete AZFc deletion have up to a 50% chance of sperm recovery, whereas those with complete deletions of the AZFa, AZFb, or AZFbc regions are unlikely to benefit from TESE (Krausz & Riera-Escamilla, 2018). Moreover, the expanding research in male infertility genomics is uncovering new genes that, when affected by pathogenic mutations, can serve as predictors for outcomes in ART. For instance, mutations in the *AURKC* gene suggest that the sperm are unlikely to result in a viable pregnancy even with ART interventions (Wyrwoll et al., 2023), while homozygous deletions in the *CATSPER2* gene indicate that ICSI may be necessary. Additionally, pathogenic variants in genes like *DNAH1*, which are associated with MMAF, underscore the need for specialised ART procedures such as oocyte activation (Stallmeyer et al., 2024).

Another critical aspect of genetic diagnosis is its role in assessing the risk of transmitting infertility to offspring via ART (Veltman & Tüttelmann, 2024). The risk depends on factors such as the inheritance pattern, genotype, sex, and the specific chromosome carrying the mutation. For instance, in cases where the infertility is due to dominant genetic causes, there is a 50% chance that the damaging mutation will be passed on to the offspring. A *de novo* variant arising in the germline can, therefore, become an inherited mutation when ART is used, potentially leading to infertility in male offspring or causing female carriers to potentially pass the mutation to their children in the future.

Furthermore, identifying the genetic basis of male infertility offers insights into potential health risks beyond reproductive challenges. Several studies have linked infertility with an elevated risk of developing other conditions, including hypertension, ischemic heart disease, diabetes mellitus, autoimmune disorders, and various cancers (Barratt et al., 2021; Kimmins et al., 2024). For example, a 2022 study analysing WES data from a cohort of 836 NOA patients revealed that approximately 1 in 28 carried a medically actionable secondary finding, most notably variants in cancer-associated genes (Kasak & Laan, 2021). Similarly, candidate genes such as *AKT1* and *BRAF*, identified during this research, are well-known in cancer biology. Stratifying patients based on genotype and phenotype can thus provide critical insights into specific comorbidities linked to particular male infertility phenotypes, enabling better clinical follow-up and targeted treatment strategies.

Along with these advantages, uncovering the genetic causes of infertility not only enhances clinical management but also deepens our understanding of human spermatogenesis and overall reproductive biology. Elucidating the functions of candidate genes in humans will contribute to a more comprehensive understanding of the mechanisms governing reproduction and may open avenues for novel therapeutic interventions.

Improving genetic diagnosis and treatment in male infertility involves addressing several complex challenges. Firstly, current research has largely focused on recessive and X-linked causes of infertility (Houston et al., 2021). However, the contribution of dominant genetic causes, whether maternally inherited or, more commonly, *de novo* mutations, remains less explored. Addressing these dominant causes requires a shift toward patient-parent trio-based studies, which include genetic data from the affected individual and both parents. Our group has taken an important initial step in this direction by recruiting around 400 patient-parent trios, providing valuable insights despite the sample size being modest for a comprehensive analysis (Oud et al., 2021). Moving forward, the field of male infertility genetics would greatly benefit from larger cohort studies that incorporate WGS. This could provide a more comprehensive understanding of genetic causes, particularly non-coding regions and *de novo* mutations, which are often overlooked in smaller studies.

Additional challenges arise from the limited understanding of non-coding regions and complex genomic areas. Even when non-coding variants are identified, predicting the functional consequences of SVs is difficult due to incomplete knowledge of regulatory regions across different tissues. Additionally, rigorous functional validation studies are essential to confirm

the roles of newly identified genes in male reproductive biology. In this context, multi-omic approaches, as highlighted in a recent review, hold promise for better interpretation of variants and elucidating the underlying molecular mechanisms (Wagner et al., 2023).

Regarding advancements in treatment, gene therapy has shown promise in preclinical models. In one study with infertile mice, lentiviral gene transfer and CRISPR-Cas9-based correction of mutation in the *Tex11* gene have successfully restored spermatogenesis. (Y.-H. Wang et al., 2021). Despite these promising results, gene therapy remains in its infancy for clinical application in male infertility due to high costs, limited human studies, and the current availability of effective ART methods. In the future, if specific subgroups of patients are identified for whom ART is not a viable option, gene therapy might emerge as a beneficial alternative, particularly for those with co-morbid conditions that further compromise fertility.

In summary, advancing genetic diagnostics in male infertility holds the promise of more precise clinical interventions, better prediction of ART outcomes, improved patient counselling regarding hereditary risks, and enhanced understanding of associated systemic health conditions. These developments will ultimately lead to more personalised and effective management strategies for male infertility and related health issues.

9.5 Concluding Remarks

Infertility affects approximately 10-15% of couples worldwide, with male factors contributing to nearly half of these cases. Male infertility is a complex and genetically heterogeneous disorder. To date, only a limited number of causative genes have been identified, most of which follow a recessive inheritance pattern. In this study, we present the largest cohort to date of patients with isolated forms of male infertility (specifically NOA and severe oligozoospermia) where WGS has been performed. By utilising patient-parent trio-based WGS analysis, we were able to systematically investigate different modes of inheritance and explore SVs. Our analysis identified several novel candidate genes potentially involved in male infertility, in addition to providing definitive genetic diagnoses for a subset of patients. These findings highlight the utility of comprehensive genomic approaches in uncovering the underlying genetic architecture of male infertility. Furthermore, our results support the integration of NGS technologies into clinical practice for the diagnostic evaluation of male infertility. I propose that WES may serve as an initial diagnostic tool, with WGS reserved for cases where WES is inconclusive and where resources and expertise are available.

Finally, this study underscores the need for larger cohorts, collaborative efforts, as well as functional studies, to validate candidate genes and further elucidate the molecular mechanisms of male infertility. With a deeper understanding of the genetic basis of male infertility, we can improve diagnostic accuracy, genetic counselling, and personalised clinical management for affected individuals.

Bibliography

Abel, H.J. & Duncavage, E.J. (2013) 'Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches', *Cancer Genetics*, 206(12), pp. 432-440.

Abel, H.J., Larson, D.E., Regier, A.A., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., Buyske, S., Matise, T.C., Muzny, D.M., Zody, M.C., Lander, E.S., Dutcher, S.K., Stitziel, N.O. & Hall, I.M. (2020) 'Mapping and characterization of structural variation in 17,795 human genomes', *Nature*, 583(7814), pp. 83-89.

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

Adamson, G.D., Dyer, S., Zegers-Hochschild, F., Chambers, G., De Mouzon, J., Ishihara, O., Kupka, M., Baker, V., Banker, M., Elgindy, E., Fu, B. & Jwa, S.C. (2024) 'O-122 ICMART Preliminary World Report 2020', *Human Reproduction*, 39(Supplement_1).

Aganezov, S., Yan, S.M., Soto, D.C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D.J., Shafin, K., Shumate, A., Xiao, C., Wagner, J., McDaniel, J., Olson, N.D., Sauria, M.E.G., Vollger, M.R., Rhie, A., Meredith, M., Martin, S., Lee, J., et al. (2022) 'A complete reference genome improves analysis of human genetic variation', *Science*, 376(6588).

Akmal, M., Aulanni'am, A., Widodo, M.A., Sumitro, S.B., Purnomo, B.B. & Widodo (2016) 'The important role of protamine in spermatogenesis and quality of sperm: A mini review', *Asian Pacific Journal of Reproduction*, 5(5), pp. 357-360.

Alkan, C., Coe, B.P. & Eichler, E.E. (2011) 'Genome structural variation discovery and genotyping', *Nature Reviews Genetics*, 12(5).

Altemose, N., Logsdon, G.A., Bzikadze, A. V., Sidhwani, P., Langley, S.A., Caldas, G. V., Hoyt, S.J., Uralsky, L., Ryabov, F.D., Shew, C.J., Sauria, M.E.G., Borchers, M., Gershman, A., Mikheenko, A., Shepelev, V.A., Dvorkina, T., Kunyavskaya, O., Vollger, M.R., Rhie, A., et al. (2022) 'Complete genomic and epigenetic maps of human centromeres', *Science*, 376(6588).

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

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. & Gouil, Q. (2020) 'Opportunities and challenges in long-read sequencing data analysis', *Genome Biology*, 21(1).

Amemiya, H.M., Kundaje, A. & Boyle, A.P. (2019) 'The ENCODE Blacklist: Identification of Problematic Regions of the Genome', *Scientific Reports*, 9(1), p. 9354.

Amor, H. & Hammadeh, M.E. (2022) 'A Systematic Review of the Impact of Mitochondrial Variations on Male Infertility', *Genes*, 13(7), p. 1182.

Andrews, S. (2010) *FastQC: A Quality Control Tool for High Throughput Sequence Data*. [Online]

Ariel, M., Cedar, H. & McCarrey, J. (1994) 'Developmental changes in methylation of spermatogenesis-specific genes include reprogramming in the epididymis', *Nature Genetics*, 7(1).

Asgari, R., Mansouri, K., Bakhtiari, M., Bidmeshkipour, A., Yari, K., Shaveisi-Zadeh, F. & Vaisi-Raygani, A. (2017) 'Association of FAS-670A/G and FASL-844C/T polymorphisms with idiopathic azoospermia in Western Iran', *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 218pp. 55-59.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., Gibbs, R.A., Green, E.D., Hurles, M.E., Knoppers, B.M., Korbel, J.O., Lander, E.S., Lee, C., et al. (2015) 'A global reference for human genetic variation', *Nature*, 526(7571).

Auwerx, C., Kutalik, Z. & Reymond, A. (2024) 'The pleiotropic spectrum of proximal 16p11.2 CNVs', *The American Journal of Human Genetics*, 111(11), pp. 2309-2346.

Ayhan, Ö., Balkan, M., Guven, A., Hazan, R., Atar, M., Tok, A. & Tolun, A. (2014) 'Truncating mutations in *TAF4B* and *ZMYND15* causing recessive azoospermia', *Journal of Medical Genetics*, 51(4), pp. 239-244.

Balachandran, P. & Beck, C.R. (2020) 'Structural variant identification and characterization', *Chromosome Research*, 28(1).

Barratt, C.L.R., De Jonge, C.J., Anderson, R.A., Eisenberg, M.L., Garrido, N., Rautakallio Hokkanen, S., Krausz, C., Kimmins, S., O'Bryan, M.K., Pacey, A.A., Tüttelmann, F. & Veltman, J.A. (2021) 'A global approach to addressing the policy, research and social challenges of male reproductive health', *Human Reproduction Open*, 2021(1).

Barseghyan, H., Tang, W., Wang, R.T., Almalvez, M., Segura, E., Bramble, M.S., Lipson, A., Douine, E.D., Lee, H., Délot, E.C., Nelson, S.F. & Vilain, E. (2017) 'Next-generation mapping: a novel

approach for detection of pathogenic structural variants with a potential utility in clinical diagnosis', *Genome Medicine*, 9(1).

Becker, J., Semler, O., Gilissen, C., Li, Y., Bolz, H.J., Giunta, C., Bergmann, C., Rohrbach, M., Koerber, F., Zimmermann, K., de Vries, P., Wirth, B., Schoenau, E., Wollnik, B., Veltman, J.A., Hoischen, A. & Netzer, C. (2011) 'Exome Sequencing Identifies Truncating Mutations in Human SERPINF1 in Autosomal-Recessive Osteogenesis Imperfecta', *The American Journal of Human Genetics*, 88(3), pp. 362-371.

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.-L. & Abel, L. (2015) 'Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants', *Proceedings of the National Academy of Sciences*, 112(17).

Belling, K., Russo, F., Jensen, A.B., Dalgaard, M.D., Westergaard, D., Rajpert-De Meyts, E., Skakkebæk, N.E., Juul, A. & Brunak, S. (2017) 'Klinefelter syndrome comorbidities linked to increased X chromosome gene dosage and altered protein interactome activity', *Human Molecular Genetics*, 26(7), pp. 1219-1229.

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

Belyeu, J.R., Brand, H., Wang, H., Zhao, X., Pedersen, B.S., Feusier, J., Gupta, M., Nicholas, T.J., Brown, J., Baird, L., Devlin, B., Sanders, S.J., Jorde, L.B., Talkowski, M.E. & Quinlan, A.R. (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*, 108(4), pp. 597-607.

van Belzen, I.A.E.M., Schönhuth, A., Kemmeren, P. & Hehir-Kwa, J.Y. (2021) 'Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology', *npj Precision Oncology*, 5(1), p. 15.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., et al. (2008) 'Accurate

whole human genome sequencing using reversible terminator chemistry', *Nature*, 456(7218).

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

Bonduelle, M. (2002) 'Prenatal testing in ICSI pregnancies: incidence of chromosomal anomalies in 1586 karyotypes and relation to sperm parameters', *Human Reproduction*, 17(10).

Bourc'his, D. & Bestor, T.H. (2004) 'Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L', *Nature*, 431(7004).

Brugh, V.M. & Lipshultz, L.I. (2004) 'Male factor infertility', *Medical Clinics of North America*, 88(2).

Cao, W., Ijiri, T.W., Huang, A.P. & Gerton, G.L. (2011) 'Characterization of a Novel Tektin Member, TEKT5, in Mouse Sperm', *Journal of Andrology*, 32(1), pp. 55-69.

Carvalho, C.M.B., Zhang, F. & Lupski, J.R. (2011) 'Structural variation of the human genome: mechanisms, assays, and role in male infertility', *Systems Biology in Reproductive Medicine*, 57(1-2).

Castro, C.P., Diehl, A.G. & Boyle, A.P. (2023) 'Challenges in screening for de novo noncoding variants contributing to genetically complex phenotypes', *Human Genetics and Genomics Advances*, 4(3), p. 100210.

Catford, S.R., Halliday, J., Lewis, S., O'Bryan, M.K., Handelsman, D.J., Hart, R.J., McBain, J., Rombauts, L., Amor, D.J., Saffery, R. & McLachlan, R.I. (2022) 'Reproductive function in men conceived with in vitro fertilization and intracytoplasmic sperm injection', *Fertility and Sterility*, 117(4), pp. 727-737.

Centers for Disease Control and Prevention (2022) *2020 Assisted Reproductive Technology Fertility Clinic and National Summary Report*.

Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., Fan, X., Wen, J., Handsaker, R.E., Fairley, S., Kronenberg, Z.N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A.M., et al. (2019) 'Multi-platform discovery of haplotype-resolved structural variation in human genomes', *Nature Communications*, 10(1), p. 1784.

Chandley, A.C., Christie, S., Fletcher, J., Frackiewicz, A. & Jacobs, P.A. (1972) 'Translocation heterozygosity and associated subfertility in man', *Cytogenetic and Genome Research*, 11(6), pp. 516-533.

Chandley, A.C., Edmond, P., Christie, S., Gowans, L., Fletcher, J., Frackiewicz, A. & Newton, M. (1975) 'Cytogenetics and infertility in man*', *Annals of Human Genetics*, 39(2), pp. 231-254.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S. & Saunders, C.T. (2016) 'Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications', *Bioinformatics*, 32(8).

Cherry, N. (2001) 'Occupational exposure to solvents and male infertility', *Occupational and Environmental Medicine*, 58(10),.

Chianese, C., Gunning, A.C., Giachini, C., Daguin, F., Balercia, G., Ars, E., Giacco, D. Lo, Ruiz-Castañé, E., Forti, G. & Krausz, C. (2014) 'X Chromosome-Linked CNVs in Male Infertility: Discovery of Overall Duplication Load and Recurrent, Patient-Specific Gains with Potential Clinical Relevance', *PLoS ONE*, 9(6), p. e97746.

Cho, C., Willis, W.D., Goulding, E.H., Jung-Ha, H., Choi, Y.-C., Hecht, N.B. & Eddy, E.M. (2001) 'Haploinsufficiency of protamine-1 or -2 causes infertility in mice', *Nature Genetics*, 28(1), pp. 82-86.

Cleal, K. & Baird, D.M. (2021) 'Dysgu: efficient structural variant calling using short or long reads', *bioRxiv*, p. 2021.05.28.446147.

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

Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A. V., Lowther, C., Gauthier, L.D., Wang, H., Watts, N.A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C.W., Huang, Y., Brookings, T., et al. (2020) 'A structural variation reference for medical and population genetics', *Nature*, 581(7809), pp. 444-451.

Collins, R.L., Glessner, J.T., Porcu, E., Lepamets, M., Brandon, R., Lauricella, C., Han, L., Morley, T., Niestroj, L.-M., Ulirsch, J., Everett, S., Howrigan, D.P., Boone, P.M., Fu, J., Karczewski, K.J., Kellaris, G., Lowther, C., Lucente, D., Mohajeri, K., et al. (2022) 'A cross-disorder dosage sensitivity map of the human genome', *Cell*, 185(16), pp. 3041-3055.e25.

Collins, R.L. & Talkowski, M.E. (2025) ‘Diversity and consequences of structural variation in the human genome’, *Nature Reviews Genetics*.

Colombo, R., Pontoglio, A. & Bini, M. (2017) ‘Two Novel <i>TEX15</i> Mutations in a Family with Nonobstructive Azoospermia’, *Gynecologic and Obstetric Investigation*, 82(3), pp. 283-286.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C.H., Kristiansson, K., MacArthur, D.G., MacDonald, J.R., Onyiah, I., Pang, A.W.C., Robson, S., et al. (2010) ‘Origins and functional impact of copy number variation in the human genome’, *Nature*, 464(7289).

Cooney, G.J., Lyons, R.J., Crew, A.J., Jensen, T.E., Molero, J.C., Mitchell, C.J., Biden, T.J., Ormandy, C.J., James, D.E. & Daly, R.J. (2004) ‘Improved glucose homeostasis and enhanced insulin signalling in Grb14-deficient mice’, *The EMBO Journal*, 23(3), pp. 582-593.

Cosenza, M.R., Rodriguez-Martin, B. & Korbel, J.O. (2022) ‘Structural Variation in Cancer: Role, Prevalence, and Mechanisms’, *Annual Review of Genomics and Human Genetics*, 23(1), pp. 123-152.

Cozzolino, D.J. (2001) ‘Varicocele as a progressive lesion: positive effect of varicocele repair’, *Human Reproduction Update*, 7(1).

Dam, A.H.D.M., Koscinski, I., Kremer, J.A.M., Moutou, C., Jaeger, A.-S., Oudakker, A.R., Tournaye, H., Charlet, N., Lagier-Tourenne, C., van Bokhoven, H. & Viville, S. (2007) ‘Homozygous Mutation in SPATA16 Is Associated with Male Infertility in Human Globozoospermia’, *The American Journal of Human Genetics*, 81(4), pp. 813-820.

Delgado-Bermúdez, A., Yeste, M., Bonet, S. & Pinart, E. (2024) ‘Physiological role of potassium channels in mammalian germ cell differentiation, maturation, and capacitation’, *Andrology*,

Dennis, J., Walker, L., Tyrer, J., Michailidou, K. & Easton, D.F. (2021) ‘Detecting rare copy number variants from Illumina genotyping arrays with the CamCNV pipeline: Segmentation of *z* -scores improves detection and reliability’, *Genetic Epidemiology*, 45(3), pp. 237-248.

Dicke, A.-K., Ahmedani, A., Ma, L., van der Heijden, G.W., Koser, S.A., Krallmann, C., Kalyon, O., Xavier, M.J., Veltman, J.A., Kliesch, S., Neuhaus, N., Kotaja, N., Tüttelmann, F. & Stallmeyer,

B. (2024) ‘NXT2 is the key player for nuclear RNA export in the human testis and critical for spermatogenesis’, *medRxiv*, p. 2024.08.01.24310552.

Ding, X., Cao, L., Zheng, Y., Zhou, X., He, X. & Xu, S. (2021) ‘Insights Into the Evolution of Spermatogenesis-Related Ubiquitin-Proteasome System Genes in Abdominal Testicular Laurasiatherians’, *Genes*,

Dixon, J.R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V.T., Yardımcı, G.G., Chakraborty, A., Bann, D. v., Wang, Y., Clark, R., Zhang, L., Yang, H., Liu, T., Iyyanki, S., An, L., Pool, C., Sasaki, T., Rivera-Mulia, J.C., et al. (2018) ‘Integrative detection and analysis of structural variation in cancer genomes’, *Nature Genetics*, 50(10).

Dong, Y., Pan, Y., Wang, R., Zhang, Z., Xi, Q. & Liu, R.-Z. (2015) ‘Copy number variations in spermatogenic failure patients with chromosomal abnormalities and unexplained azoospermia’, *Genetics and Molecular Research*, 14(4).

Dowsing, A.T., Yong, E., Clark, M., McLachlan, R.I., de Kretser, D.M. & Trounson, A.O. (1999) ‘Linkage between male infertility and trinucleotide repeat expansion in the androgen-receptor gene’, *The Lancet*, 354(9179).

Dremsek, P., Schwarz, T., Weil, B., Malashka, A., Laccone, F. & Neesen, J. (2021) ‘Optical Genome Mapping in Routine Human Genetic Diagnostics—Its Advantages and Limitations’, *Genes*, 12(12), p. 1958.

Dym, M. & Fawcett, D.W. (1971) ‘Further Observations on the Numbers of Spermatogonia, Spermatocytes, and Spermatids Connected by Intercellular Bridges in the Mammalian Testis1’, *Biology of Reproduction*, 4(2), pp. 195-215.

Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., et al. (2021) ‘Haplotype-resolved diverse human genomes and integrated analysis of structural variation’, *Science*, 372(6537).

Elia, J., Glessner, J.T., Wang, K., Takahashi, N., Shtir, C.J., Hadley, D., Sleiman, P.M.A., Zhang, H., Kim, C.E., Robison, R., Lyon, G.J., Flory, J.H., Bradfield, J.P., Imielinski, M., Hou, C., Frackelton, E.C., Chiavacci, R.M., Sakurai, T., Rabin, C., et al. (2012) ‘Genome-wide copy number variation study associates metabotropic glutamate receptor gene networks with attention deficit hyperactivity disorder’, *Nature Genetics*, 44(1).

Ellingford, J.M., Ahn, J.W., Bagnall, R.D., Baralle, D., Barton, S., Campbell, C., Downes, K., Ellard, S., Duff-Farrier, C., FitzPatrick, D.R., Greally, J.M., Ingles, J., Krishnan, N., Lord, J., Martin, H.C., Newman, W.G., O'Donnell-Luria, A., Ramsden, S.C., Rehm, H.L., et al. (2022) 'Recommendations for clinical interpretation of variants found in non-coding regions of the genome', *Genome Medicine*, 14(1), p. 73.

Emanuele MA & Emanuele NV (1998) 'Alcohol's effects on male reproduction. ', *Alcohol Health Res World.* , 22(3), pp. 195-201.

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

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

Ewels, P., Magnusson, M., Lundin, S. & Käller, M. (2016) 'MultiQC: summarize analysis results for multiple tools and samples in a single report', *Bioinformatics*, 32(19).

Fakhro, K.A., Elbardisi, H., Arafa, M., Robay, A., Rodriguez-Flores, J.L., Mezey, J.G., Crystal, R.G., Al-Shakaki, A., Syed, N., Abi Khalil, C., Malek, J.A., Al-Ansari, A. & Al Said, S. (2018) 'Point-of-care whole-exome sequencing of idiopathic male infertility', *Genetics in Medicine*, 20(11), pp. 1365-1373.

Falls, J.G., Pulford, D.J., Wylie, A.A. & Jirtle, R.L. (1999) 'Genomic Imprinting: Implications for Human Disease', *The American Journal of Pathology*, 154(3), pp. 635-647.

Fazal, S., Danzi, M.C., Cintra, V.P., Bis-Brewer, D.M., Dolzhenko, E., Eberle, M.A. & Zuchner, S. (2020) 'Large scale in silico characterization of repeat expansion variation in human genomes', *Scientific Data*, 7(1).

Feng, K., Ge, H., Chen, H., Cui, C., Zhang, S., Zhang, C., Meng, L., Guo, H. & Zhang, L. (2022) 'Novel exon mutation in *SYCE1* gene is associated with non-obstructive azoospermia', *Journal of Cellular and Molecular Medicine*, 26(4), pp. 1245-1252.

Ferlin, A. (2004) 'Androgen receptor gene CAG and GGC repeat lengths in idiopathic male infertility', *Molecular Human Reproduction*, 10(6).

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

Ferreira, L.T., Logarinho, E., Macedo, J.C., Maia, A.R.R. & Maiato, H. (2021) 'SOGA1 and SOGA2/MTCL1 are CLASP-interacting proteins required for faithful chromosome segregation in human cells', *Chromosome Research*, 29(2), pp. 159-173.

Feuk, L., Carson, A.R. & Scherer, S.W. (2006) 'Structural variation in the human genome', *Nature Reviews Genetics*, 7(2), pp. 85-97.

Fice, H.E. & Robaire, B. (2023) 'Aging affects gene expression in spermatids of Brown Norway rats', *Experimental Gerontology*, 173p. 112086.

Firke, S. (2021) *janitor: Simple Tools for Examining and Cleaning Dirty Data*.

Firth, H. V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Vooren, S. Van, Moreau, Y., Pettett, R.M. & Carter, N.P. (2009) 'DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources', *The American Journal of Human Genetics*, 84(4).

Fu, Y., Timp, W., & Sedlazeck, F. J. (2025). Computational analysis of DNA methylation from long-read sequencing. Nature Reviews Genetics, 26(9), 620–634.

Gel, B. & Serra, E. (2017) 'karyoploteR : an R / Bioconductor package to plot customizable genomes displaying arbitrary data', *Bioinformatics*, 33(19), pp. 3088-3090.

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

Gershoni, M., Hauser, R., Yogev, L., Lehavi, O., Azem, F., Yavetz, H., Pietrokovski, S. & Kleiman, S.E. (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.

Lo Giacco, D, Chianese, C., Ars, E., Ruiz-Castañé, E., Forti, G. & Krausz, C. (2014) 'Recurrent X chromosome-linked deletions: discovery of new genetic factors in male infertility', *Journal of Medical Genetics*, 51(5), pp. 340-344.

Lo Giacco, Deborah, Chianese, C., Sánchez-Curbelo, J., Bassas, L., Ruiz, P., Rajmil, O., Sarquella, J., Vives, A., Ruiz-Castañé, E., Oliva, R., Ars, E. & Krausz, C. (2014) 'Clinical relevance of Y-linked

CNV screening in male infertility: new insights based on the 8-year experience of a diagnostic genetic laboratory', *European Journal of Human Genetics*, 22(6), pp. 754-761.

Gilissen, C., Hoischen, A., Brunner, H.G. & Veltman, J.A. (2012) 'Disease gene identification strategies for exome sequencing', *European Journal of Human Genetics*, 20(5), pp. 490-497.

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

Girirajan, S. & Eichler, E.E. (2011) 'De Novo CNVs in Bipolar Disorder: Recurrent Themes or New Directions?', *Neuron*, 72(6).

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

Guan, P. & Sung, W.-K. (2016) 'Structural variation detection using next-generation sequencing data', *Methods*, 102.

Guo, S. (2023) 'Transcriptome Studies Reveal the N6-Methyladenosine Differences in Testis of Yaks at Juvenile and Sexual Maturity Stages', *Animals*.

Hamidian, S., Talebi, A.R., Fesahat, F., Bayat, M., Mirjalili, A.M., Ashrafzadeh, H.R., Rajabi, M., Montazeri, F. & Babaei, S. (2020) 'The effect of vitamin C on the gene expression profile of sperm protamines in the male partners of couples with recurrent pregnancy loss: A randomized clinical trial', *Clinical and Experimental Reproductive Medicine*, 47(1), pp. 68-76.

Harbuz, R., Zouari, R., Pierre, V., Ben Khelifa, M., Kharouf, M., Coutton, C., Merdassi, G., Abada, F., Escoffier, J., Nikas, Y., Vialard, F., Koscinski, I., Triki, C., Sermondade, N., Schweitzer, T., Zhioua, A., Zhioua, F., Latrous, H., Halouani, L., et al. (2011) 'A Recurrent Deletion of DPY19L2 Causes Infertility in Man by Blocking Sperm Head Elongation and Acrosome Formation', *The American Journal of Human Genetics*, 88(3), pp. 351-361.

Hardy, J., Pollock, N., Gingrich, T., Sweet, P., Ramesh, A., Kuong, J., Basar, A., Jiang, H., Hwang, K., Vukina, J., Jaffe, T., Olszewska, M., Kurpisz, M. & Yatsenko, A.N. (2022) 'Genomic testing for copy number and single nucleotide variants in spermatogenic failure', *Journal of Assisted Reproduction and Genetics*, 39(9), pp. 2103-2114.

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

Hastings, P.J., Lupski, J.R., Rosenberg, S.M. & Ira, G. (2009) 'Mechanisms of change in gene copy number', *Nature Reviews Genetics*, 10(8).

Hehir-Kwa, J.Y., Rodriguez-Santiago, B., Vissers, L.E., de Leeuw, N., Pfundt, R., Buitelaar, J.K., Perez-Jurado, L.A. & Veltman, J.A. (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.

Herati, A.S., Butler, P.R. & Lamb, D.J. (2017) 'The Genetic Basis of Male Infertility', in *The Sperm Cell*. [Online]. Cambridge University Press. pp. 208-229.

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

Ho, S.S., Urban, A.E. & Mills, R.E. (2020) 'Structural variation in the sequencing era', *Nature Reviews Genetics*, 21(3).

Holt, G.S., Batty, L.E., Alobaidi, B.K.S., Smith, H.E., Oud, M.S., Ramos, L., Xavier, M.J. & Veltman, J.A. (2022) 'Phasing of de novo mutations using a scaled-up multiple amplicon long-read sequencing approach', *Human Mutation*, 43(11), pp. 1545-1556.

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

Hu, J., Wang, F., Yuan, Y., Zhu, X., Wang, Y., Zhang, Y., Kou, Z., Wang, S. & Gao, S. (2010) 'Novel Importin-α Family Member Kpna7 Is Required for Normal Fertility and Fecundity in the Mouse*', *Journal of Biological Chemistry*, 285(43), pp. 33113-33122.

Hubbard, L., Rambhatla, A., & Colpi, G. M. (2025). Differentiation between nonobstructive azoospermia and obstructive azoospermia: then and now. Asian Journal of Andrology, 27(3), 298–306.

Huang, N., Wen, Y., Guo, X., Li, Z., Dai, J., Ni, B., Yu, J., Lin, Y., Zhou, W., Yao, B., Jiang, Y., Sha, J., Conrad, D.F. & Hu, Z. (2015) 'A Screen for Genomic Disorders of Infertility Identifies MAST2 Duplications Associated with Nonobstructive Azoospermia in Humans1', *Biology of Reproduction*, 93(3).

Huang, W., Li, L., Myers, J.R. & Marth, G.T. (2012) 'ART: a next-generation sequencing read simulator', *Bioinformatics*, 28(4), pp. 593-594.

Huang, X., Wang, H.-L., Qi, S.-T., Wang, Z.-B., Tong, J.-S., Zhang, Q.-H., Ouyang, Y.-C., Hou, Y., Schatten, H., Qi, Z.-Q. & Sun, Q.-Y. (2011) 'DYNLT3 Is Required for Chromosome Alignment During Mouse Oocyte Meiotic Maturation', *Reproductive Sciences*, 18(10), pp. 983-989.

Huang, Y., Tian, R., Xu, J., Ji, Z., Zhang, Y., Zhao, L., Yang, C., Li, P., Zhi, E., Bai, H., Han, S., Luo, J., Zhao, J., Zhang, J., Zhou, Z., Li, Z. & Yao, C. (2022) 'Novel copy number variations within SYCE1 caused meiotic arrest and non-obstructive azoospermia', *BMC Medical Genomics*, 15(1), p. 137.

Ignatieva, E. V., Osadchuk, A. V., Kleshchev, M.A., Bogomolov, A.G. & Osadchuk, L. V. (2021) 'A Catalog of Human Genes Associated With Pathozoospermia and Functional Characteristics of These Genes', *Frontiers in Genetics*, 12.

Inhorn, M.C., Kobeissi, L., Nassar, Z., Lakkis, D. & Fakih, M.H. (2009) 'Consanguinity and family clustering of male factor infertility in Lebanon', *Fertility and Sterility*, 91(4), pp. 1104-1109.

Jarow, J.P. (2001) 'Effects of varicocele on male fertility', *Human Reproduction Update*, 7(1).

Jarow, J.P. (2003) 'Endocrine causes of male infertility', *Urologic Clinics of North America*, 30(1).

Ji, G., Gu, A., Hu, F., Wang, S., Liang, J., Xia, Y., Lu, C., Song, L., Fu, G. & Wang, X. (2009) 'Polymorphisms in cell death pathway genes are associated with altered sperm apoptosis and poor semen quality', *Human Reproduction*, 24(10), pp. 2439-2446.

Ji, Z., Yao, C., Yang, C., Huang, C., Zhao, L., Han, X., Zhu, Z., Zhi, E., Liu, N., Zhou, Z., & Li, Z. (2021). Novel Hemizygous Mutations of TEX11 Cause Meiotic Arrest and Non-obstructive Azoospermia in Chinese Han Population. Frontiers in Genetics, 12.

Jin, S.-K. & Yang, W.-X. (2017) 'Factors and pathways involved in capacitation: how are they regulated?', *Oncotarget*, 8(2), pp. 3600-3627.

Jobling, M.A., Lo, I.C.C., Turner, D.J., Bowden, G.R., Lee, A.C., Xue, Y., Carvalho-Silva, D., Hurles, M.E., Adams, S.M., Chang, Y.M., Kraaijenbrink, T., Henke, J., Guanti, G., McKeown, B., van Oorschot, R.A.H., Mitchell, R.J., de Knijff, P., Tyler-Smith, C. & Parkin, E.J. (2007) 'Structural variation on the short arm of the human Y chromosome: recurrent multigene deletions encompassing Amelogenin Y', *Human Molecular Genetics*, 16(3), pp. 307-316.

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

Joshi, M. & Rajender, S. (2020) 'Long non-coding RNAs (lncRNAs) in spermatogenesis and male infertility', *Reproductive Biology and Endocrinology*, 18(1), p. 103.

Jung, H., Yang, T.-P., Walker, S., Danecek, P., Salinas, I.G., Neville, M.D.C., Firth, H., Scally, A., Hurles, M., Campbell, P. & Rahbari, R. (2024) 'Deciphering the role of germline complex de novo structural variations in rare disorders', *bioRxiv*, p. 2024.04.03.587925.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., Gauthier, L.D., Brand, H., Solomonson, M., Watts, N.A., Rhodes, D., Singer-Berk, M., England, E.M., Seaby, E.G., Kosmicki, J.A., et al. (2020) 'The mutational constraint spectrum quantified from variation in 141,456 humans', *Nature*, 581(7809), pp. 434-443.

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

Kasak, L., Punab, M., Nagirnaja, L., Grigorova, M., Minajeva, A., Lopes, A.M., Punab, A.M., Aston, K.I., Carvalho, F., Laasik, E., Smith, L.B., Conrad, D.F. & Laan, M. (2018) 'Bi-allelic Recessive

Loss-of-Function Variants in FANCM Cause Non-obstructive Azoospermia', *The American Journal of Human Genetics*, 103(2), pp. 200-212.

Kassambara, A. (2020) *ggpubr: 'ggplot2' Based Publication Ready Plots*.

Khan, M., Jabeen, N., Khan, T., Hussain, H.M.J., Ali, A., Khan, R., Jiang, L., Li, T., Tao, Q., Zhang, X., Yin, H., Yu, C., Jiang, X. & Shi, Q. (2018) 'The evolutionarily conserved genes: Tex37, Ccdc73, Prss55 and Nxt2 are dispensable for fertility in mice', *Scientific Reports*, 8(1), p. 4975.

Khayat, M.M., Sahraeian, S.M.E., Zarate, S., Carroll, A., Hong, H., Pan, B., Shi, L., Gibbs, R.A., Mohiyuddin, M., Zheng, Y. & Sedlazeck, F.J. (2021) 'Hidden biases in germline structural variant detection', *Genome Biology*, 22(1), p. 347.

Kherraf, Z., Christou-Kent, M., Karaouzene, T., Amiri-Yekta, A., Martinez, G., Vargas, A.S., Lambert, E., Borel, C., Dorphin, B., Aknin-Seifer, I., Mitchell, M.J., Metzler-Guillemain, C., Escoffier, J., Nef, S., Grepillat, M., Thierry-Mieg, N., Satre, V., Bailly, M., Boitrelle, F., et al. (2017) '<scp>SPINK</scp> 2 deficiency causes infertility by inducing sperm defects in heterozygotes and azoospermia in homozygotes', *EMBO Molecular Medicine*, 9(8), pp. 1132-1149.

Kimmins, S., Anderson, R.A., Barratt, C.L.R., Behre, H.M., Catford, S.R., De Jonge, C.J., Delbes, G., Eisenberg, M.L., Garrido, N., Houston, B.J., Jørgensen, N., Krausz, C., Lismer, A., McLachlan, R.I., Minhas, S., Moss, T., Pacey, A., Priskorn, L., Schlatt, S., et al. (2024) 'Frequency, morbidity and equity — the case for increased research on male fertility', *Nature Reviews Urology*, 21(2), pp. 102-124.

Kirat, D., Alahwany, A.M., Arisha, A.H., Abdelkhalek, A. & Miyasho, T. (2023) 'Role of Macroautophagy in Mammalian Male Reproductive Physiology', *Cells*, 12(9), p. 1322.

Kirov, G., Pocklington, A.J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., Moran, J., Chambert, K., Toncheva, D., Georgieva, L., Grozeva, D., Fjodorova, M., Wollerton, R., Rees, E., Nikolov, I., van de Lagemaat, L.N., Bayés, À., Fernandez, E., Olason, P.I., et al. (2012) 'De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia', *Molecular Psychiatry*, 17(2).

Kloosterman, W.P., Francioli, L.C., Hormozdiari, F., Marschall, T., Hehir-Kwa, J.Y., Abdellaoui, A., Lameijer, E.-W., Moed, M.H., Koval, V., Renkens, I., van Roosmalen, M.J., Arp, P., Karssen, L.C., Coe, B.P., Handsaker, R.E., Suchiman, E.D., Cuppen, E., Thung, D.T., McVey, M., et al.

(2015) ‘Characteristics of de novo structural changes in the human genome’, *Genome Research*, 25(6), pp. 792-801.

Komsta, L. (2022) *outliers: Tests for Outliers*.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., Taillon, B.E., Chen, Z., Tanzer, A., Saunders, A.C.E., Chi, J., Yang, F., Carter, N.P., Hurles, M.E., Weissman, S.M., et al. (2007) ‘Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome’, *Science*, 318(5849).

Kort, H.I. (2006) ‘Impact of Body Mass Index Values on Sperm Quantity and Quality’, *Journal of Andrology*, 27(3).

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M. & Kamatani, Y. (2019) ‘Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing’, *Genome Biology*, 20(1).

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

Krausz, C. & Casamonti, E. (2017) ‘Spermatogenic failure and the Y chromosome’, *Human Genetics*, 136(5), pp. 637-655.

Krausz, C., Escamilla, A.R. & Chianese, C. (2015) ‘Genetics of male infertility: from research to clinic’, *REPRODUCTION*, 150(5), pp. R159-R174.

Krausz, C., Navarro-Costa, P., Wilke, M. & Tüttelmann, F. (2024) ‘EAA/EMQN best practice guidelines for molecular diagnosis of Y-chromosomal microdeletions: State of the art 2023’, *Andrology*, 12(3), pp. 487-504.

Krausz, C. & Riera-Escamilla, A. (2018) ‘Genetics of male infertility’, *Nature Reviews Urology*, 15(6).

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software, 82(13).

Kuzniar, A., Maassen, J., Verhoeven, S., Santuari, L., Shneider, C., Kloosterman, W.P. & de Ridder, J. (2020) ‘sv-callers: a highly portable parallel workflow for structural variant detection in whole-genome sequence data’, *PeerJ*, 8.

Kyrgiafini, M.-A., Sarafidou, T. & Mamuris, Z. (2022) 'The Role of Long Noncoding RNAs on Male Infertility: A Systematic Review and In Silico Analysis', *Biology*, 11(10), p. 1510.

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

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. & Carey, V. (2013) 'Software for Computing and Annotating Genomic Ranges', *PLoS Computational Biology*, 9(8).

Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. (2014) 'LUMPY: a probabilistic framework for structural variant discovery', *Genome Biology*, 15(6).

Lee, W., Zamudio-Ochoa, A., Buchel, G., Podlesniy, P., Marti Gutierrez, N., Puigròs, M., Calderon, A., Tang, H.-Y., Li, L., Mikhalchenko, A., Koski, A., Trullas, R., Mitalipov, S., & Temiakov, D. (2023). Molecular basis for maternal inheritance of human mitochondrial DNA. Nature Genetics, 55(10), 1632–1639.

Levy-Sakin, M., Pastor, S., Mostovoy, Y., Li, L., Leung, A.K.Y., McCaffrey, J., Young, E., Lam, E.T., Hastie, A.R., Wong, K.H.Y., Chung, C.Y.L., Ma, W., Sibert, J., Rajagopalan, R., Jin, N., Chow, E.Y.C., Chu, C., Poon, A., Lin, C., et al. (2019) 'Genome maps across 26 human populations reveal population-specific patterns of structural variation', *Nature Communications*, 10(1).

Li, H. & Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14).

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16).

Li, J., Schilit, S.L.P., Liang, S., Qin, N., Teng, X. & Zhang, J. (2024) 'Novel Loss-of-Function SYCP2 Variants in Infertile Males Upgrade the Gene-Disease Clinical Validity Classification for SYCP2 and Male Infertility to Strong', *Genes*, 15(8), p. 1092.

Li, S., Zhang, Z., Xie, L., Zhao, Y., Chen, H., Zhang, S., Cai, Y., Ren, B., Liu, W., Tang, S. & Sha, Y. (2024) 'Novel bi-allelic DNAH3 variants cause oligoasthenoteratozoospermia', *Frontiers in Endocrinology*, 15.

Li, Y.R., Glessner, J.T., Coe, B.P., Li, J., Mohebnasab, M., Chang, X., Connolly, J., Kao, C., Wei, Z., Bradfield, J., Kim, C., Hou, C., Khan, M., Mentch, F., Qiu, H., Bakay, M., Cardinale, C., Lemma, M., Abrams, D., et al. (2020) 'Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations', *Nature Communications*, 11(1), p. 255.

de Ligt, J., Willemsen, M.H., van Bon, B.W.M., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., del Rosario, M., Hoischen, A., Scheffer, H., de Vries, B.B.A., Brunner, H.G., Veltman, J.A. & Vissers, L.E.L.M. (2012) 'Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability', *New England Journal of Medicine*, 367(20), pp. 1921-1929.

Lillepea, K., Juchnewitsch, A.-G., Kasak, L., Valkna, A., Dutta, A., Pomm, K., Poolamets, O., Nagirnaja, L., Tamp, E., Mahyari, E., Vihljajev, V., Tjagur, S., Papadimitriou, S., Riera-Escamilla, A., Versbraegen, N., Farnetani, G., Castillo-Madeen, H., Sütt, M., Kübarsepp, V., et al. (2024) 'Toward clinical exomes in diagnostics and management of male infertility', *The American Journal of Human Genetics*, 111(5), pp. 877-895.

Lima, A.C., Carvalho, F., Gonçalves, J., Fernandes, S., Marques, P.I., Sousa, M., Barros, A., Seixas, S., Amorim, A., Conrad, D.F. & Lopes, A.M. (2015) 'Rare double sex and mab-3-related transcription factor 1 regulatory variants in severe spermatogenic failure', *Andrology*, 3(5), pp. 825-833.

Lin, K., Smit, S., Bonnema, G., Sanchez-Perez, G. & de Ridder, D. (2015) 'Making the difference: integrating structural variation detection tools', *Briefings in Bioinformatics*, 16(5).

Linardopoulou, E. v., Williams, E.M., Fan, Y., Friedman, C., Young, J.M. & Trask, B.J. (2005) 'Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication', *Nature*, 437(7055).

Liu, Y., Wang, G., Zhang, F. & Dai, L. (2021) 'An NGS-based approach to identify Y-chromosome variation in non-obstructive azoospermia', *Andrologia*, 53(10).

Liu, Y., Zhang, M., Sun, J., Chang, W., Sun, M., Zhang, S. & Wu, J. (2020) 'Comparison of multiple algorithms to reliably detect structural variants in pears', *BMC Genomics*, 21(1).

Lopes, A.M., Aston, K.I., Thompson, E., Carvalho, F., Gonçalves, J., Huang, N., Matthiesen, R., Noordam, M.J., Quintela, I., Ramu, A., Seabra, C., Wilfert, A.B., Dai, J., Downie, J.M., Fernandes, S., Guo, X., Sha, J., Amorim, A., Barros, A., et al. (2013) 'Human Spermatogenic

Failure Purges Deleterious Mutation Load from the Autosomes and Both Sex Chromosomes, including the Gene DMRT1', *PLoS Genetics*, 9(3), p. e1003349.

Luo, T., Chen, H., Zou, Q., Wang, T., Cheng, Y., Wang, H., Wang, F., Jin, Z., Chen, Y., Weng, S. & Zeng, X. (2019) 'A novel copy number variation in *CATSPER2* causes idiopathic male infertility with normal semen parameters', *Human Reproduction*, 34(3).

Maekawa, M., Kamimura, K. & Nagano, T. (1996) 'Peritubular Myoid Cells in the Testis: Their Structure and Function.', *Archives of Histology and Cytology*, 59(1), pp. 1-13.

Mai, Z., Yang, D., Wang, D., Zhang, J., Zhou, Q., Han, B. & Sun, Z. (2024) 'A narrative review of mitochondrial dysfunction and male infertility', *Translational Andrology and Urology*, 13(9), pp. 2134-2145.

Mancini, F., Di Nicuolo, F., Teveroni, E., Vergani, E., Bianchetti, G., Bruno, C., Grande, G., Iavarone, F., Maulucci, G., De Spirito, M., Urbani, A., Pontecorvi, A. & Milardi, D. (2023) 'Combined evaluation of prolactin-induced peptide (PIP) and extracellular signal-regulated kinase (ERK) as new sperm biomarkers of FSH treatment efficacy in normogonadotropic idiopathic infertile men', *Journal of Endocrinological Investigation*, 47(2), pp. 455-468.

Mantere, T., Neveling, K., Pebrel-Richard, C., Benoist, M., van der Zande, G., Kater-Baats, E., Baatout, I., van Beek, R., Yammine, T., Oorsprong, M., Hsoumi, F., Olde-Weghuis, D., Majdali, W., Vermeulen, S., Pauper, M., Lebbar, A., Stevens-Kroef, M., Sanlaville, D., Dupont, J.M., et al. (2021) 'Optical genome mapping enables constitutional chromosomal aberration detection', *The American Journal of Human Genetics*, 108(8), pp. 1409-1422.

Martinez-Heredia, J., de Mateo, S., Vidal-Taboada, J.M., Ballesca, J.L. & Oliva, R. (2008) 'Identification of proteomic differences in asthenozoospermic sperm samples', *Human Reproduction*, 23(4), pp. 783-791.

Marques, C.J., Costa, P., Vaz, B., Carvalho, F., Fernandes, S., Barros, A. & Sousa, M. (2008) 'Abnormal methylation of imprinted genes in human sperm is associated with oligozoospermia', *MHR: Basic science of reproductive medicine*, 14(2), pp. 67-74.

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

Massaia, A. & Xue, Y. (2017) 'Human Y chromosome copy number variation in the next generation sequencing era and beyond', *Human Genetics*, 136(5), pp. 591-603.

McCarroll, S.A., Huett, A., Kuballa, P., Chilewski, S.D., Landry, A., Goyette, P., Zody, M.C., Hall, J.L., Brant, S.R., Cho, J.H., Duerr, R.H., Silverberg, M.S., Taylor, K.D., Rioux, J.D., Altshuler, D., Daly, M.J. & Xavier, R.J. (2008) 'Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease', *Nature Genetics*, 40(9).

McDonald-McGinn, D.M. & Sullivan, K.E. (2011) 'Chromosome 22q11.2 Deletion Syndrome (DiGeorge Syndrome/Velocardiofacial Syndrome)', *Medicine*, 90(1), pp. 1-18.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., Zhang, Z., Ranade, S.S., Dimalanta, E.T., Hyland, F.C., Sokolsky, T.D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C.L., et al. (2009) 'Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding', *Genome Research*, 19(9).

Medvedev, P., Stanciu, M. & Brudno, M. (2009) 'Computational methods for discovering structural variation with next-generation sequencing', *Nature Methods*, 6(S11).

Meng, X., Wang, M., Luo, M., Sun, L., Yan, Q. & Liu, Y. (2023) 'Systematic evaluation of multiple NGS platforms for structural variants detection', *Journal of Biological Chemistry*, 299(12), p. 105436.

Merges, G.E., Meier, J., Schneider, S., Kruse, A., Fröbius, A.C., Kirfel, G., Steger, K., Arévalo, L. & Schorle, H. (2022) 'Loss of *Prm1* leads to defective chromatin protamination, impaired PRM2 processing, reduced sperm motility and subfertility in male mice', *Development*, 149(12).

Merkel, D. (2014) 'Docker: lightweight linux containers for consistent development and deployment', *Linux journal*, 2014(239), pp. 2-2.

Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., Schneider, V.A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., et al. (2020) 'Telomere-to-telomere assembly of a complete human X chromosome', *Nature*, 585(7823), .

Milunsky, A., Milunsky, J.M., Dong, W., Hovhannisyan, H. & Oates, R.D. (2020) 'A contiguous microdeletion syndrome at Xp23.13 with non-obstructive azoospermia and congenital cataracts', *Journal of Assisted Reproduction and Genetics*, 37(2), pp. 471-475.

Minhas, S., Bettocchi, C., Boeri, L., Capogrosso, P., Carvalho, J., Cilesiz, N.C., Cocci, A., Corona, G., Dimitropoulos, K., Gül, M., Hatzichristodoulou, G., Jones, T.H., Kadioglu, A., Martínez Salamanca, J.I., Milenkovic, U., Modgil, V., Russo, G.I., Serefoglu, E.C., Tharakan, T., et al. (2021) 'European Association of Urology Guidelines on Male Sexual and Reproductive Health: 2021 Update on Male Infertility', *European Urology*, 80(5), pp. 603-620.

Mizuno, K., Kojima, Y., Kamisawa, H., Moritoki, Y., Nishio, H., Kohri, K. & Hayashi, Y. (2013) 'Gene Expression Profile During Testicular Development in Patients With SRY-negative 46,XX Testicular Disorder of Sex Development', *Urology*, 82(6), pp. 1453.e1-1453.e7.

Moreno, R.D., Urriola-Muñoz, P. & Lagos-Cabré, R. (2011) 'The emerging role of matrix metalloproteases of the ADAM family in male germ cell apoptosis', *Spermatogenesis*, 1(3), pp. 195-208.

Mosaad, Y.M., Shahin, D., Elkholy, A.A.-M., Mosbah, A. & Badawy, W. (2012) 'CAG repeat length in androgen receptor gene and male infertility in Egyptian patients', *Andrologia*, 44(1).

Mottes, F., Villa, C., Osella, M. & Caselle, M. (2021) 'The impact of whole genome duplications on the human gene regulatory networks', *PLOS Computational Biology*, 17(12), p. e1009638.

Mueller, J.L., Mahadevaiah, S.K., Park, P.J., Warburton, P.E., Page, D.C. & Turner, J.M.A. (2008) 'The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression', *Nature Genetics*, 40(6), pp. 794-799.

Mueller, J.L., Skaletsky, H., Brown, L.G., Zaghlul, S., Rock, S., Graves, T., Auger, K., Warren, W.C., Wilson, R.K. & Page, D.C. (2013) 'Independent specialization of the human and mouse X chromosomes for the male germ line', *Nature Genetics*, 45(9), pp. 1083-1087.

Nagirnaja, L., Aston, K.I. & Conrad, D.F. (2018) 'Genetic intersection of male infertility and cancer', *Fertility and Sterility*, 109(1).

Nagirnaja, L., Lopes, A.M., Charng, W.-L., Miller, B., Stakaitis, R., Golubickaite, I., Stendahl, A., Luan, T., Friedrich, C., Mahyari, E., Fadial, E., Kasak, L., Vigh-Conrad, K., Oud, M.S., Xavier,

M.J., Cheers, S.R., James, E.R., Guo, J., Jenkins, T.G., et al. (2022) 'Diverse monogenic subforms of human spermatogenic failure', *Nature Communications*, 13(1), p. 7953.

Nakatochi, M., Kushima, I. & Ozaki, N. (2021) 'Implications of germline copy-number variations in psychiatric disorders: review of large-scale genetic studies', *Journal of Human Genetics*, 66(1), pp. 25-37.

Nasirshalal, M., Tahmasebi-Birgani, M., Dadfar, M., Nikbakht, R., Saberi, A. & Ghandil, P. (2020) 'Identification of the *PRM1* gene mutations in oligoasthenoteratozoospermic men', *Andrologia*, 52(11).

Navarrete-López, P., Maroto, M., Pericuesta, E., Fernández-González, R., Lombó, M., Ramos-Ibeas, P. & Gutiérrez-Adán, A. (2023) 'Loss of the importin Kpna2 causes infertility in male mice by disrupting the translocation of testis-specific transcription factors', *iScience*, 26(7), p. 107134.

Nemati, H., Sadeghi, M., Nazeri, M. & Mohammadi, M. (2020) 'Evaluation of the association between polymorphisms of PRM1 and PRM2 and the risk of male infertility: a systematic review, meta-analysis, and meta-regression', *Scientific Reports*, 10(1), p. 17228.

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

Njogu, M.M., Ricketts, P.-G. & Klaus, A. V (2010) 'Spermatogenic Cyst and Organ Culture in Drosophila Pseudoobscura', *Cell and Tissue Research*.

O'Hara, L. & Smith, L.B. (2015) 'Androgen receptor roles in spermatogenesis and infertility', *Best Practice & Research Clinical Endocrinology & Metabolism*, 29(4), pp. 595-605.

Ohno, S., Zankov, D.P., Ding, W.-G., Itoh, H., Makiyama, T., Doi, T., Shizuta, S., Hattori, T., Miyamoto, A., Naiki, N., Hancox, J.C., Matsuura, H. & Horie, M. (2011) '*KCNE5* (*KCNE1L*) Variants Are Novel Modulators of Brugada Syndrome and Idiopathic Ventricular Fibrillation', *Circulation: Arrhythmia and Electrophysiology*, 4(3), pp. 352-361.

Okonofua, F.E., Ntoimo, L.F.C., Omonkhua, A., Ayodeji, O., Olafusi, C., Unuabonah, E. & Ohenhen, V. (2022) 'Causes and Risk Factors for Male Infertility: A Scoping Review of Published Studies', *International Journal of General Medicine*, Volume 15pp. 5985-5997.

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

Olinger, E., Wilson, I.J., Orr, S., Barroso-Gil, M., Neatu, R., Ambrose, J.C., Arumugam, P., Bevers, R., Bleda, M., Boardman-Pretty, F., Boustred, C.R., Brittain, H., Caulfield, M.J., Chan, G.C., Elgar, G., Fowler, T., Giess, A., Hamblin, A., Henderson, S., et al. (2024) 'Copy-number analysis from genome sequencing data of 11,754 rare-disease parent-child trios: A model for identifying autosomal recessive human gene knockouts including a novel gene for autosomal recessive retinopathy', *Genetics in Medicine Open*, 2p. 101834.

Oud, M.S., Okutman, Ö., Hendricks, L.A.J., de Vries, P.F., Houston, B.J., Vissers, L.E.L.M., O'Bryan, M.K., Ramos, L., Chemes, H.E., Viville, S. & Veltman, J.A. (2020) 'Exome sequencing reveals novel causes as well as new candidate genes for human globozoospermia', *Human Reproduction*, 35(1), pp. 240-252.

Oud, M.S., Ramos, L., O'Bryan, M.K., McLachlan, R.I., Okutman, Ö., Viville, S., Vries, P.F., Smeets, D.F.C.M., Lugtenberg, D., Hehir-Kwa, J.Y., Gilissen, C., de Vorst, M., Vissers, L.E.L.M., Hoischen, A., Meijerink, A.M., Fleischer, K., Veltman, J.A. & Noordam, M.J. (2017) 'Validation and application of a novel integrated genetic screening method to a cohort of 1,112 men with idiopathic azoospermia or severe oligozoospermia', *Human Mutation*, 38(11).

Oud, M.S., Smits, R.M., Smith, H.E., Mastrorosa, F.K., Holt, G.S., Houston, B.J., de Vries, P.F., Alobaidi, B.K.S., Batty, L.E., Ismail, H., Greenwood, J., Sheth, H., Mikulasova, A., Astuti, G.D.N., Gilissen, C., McEleny, K., Turner, H., Coxhead, J., Cockell, S., et al. (2021) 'A de novo paradigm for male infertility', *bioRxiv*, p. 2021.02.27.433155.

Pagnamenta, A.T., Camps, C., Giacopuzzi, E., Taylor, J.M., Hashim, M., Calpena, E., Kaisaki, P.J., Hashimoto, A., Yu, J., Sanders, E., Schwessinger, R., Hughes, J.R., Lunter, G., Dreau, H., Ferla, M., Lange, L., Kesim, Y., Ragoussis, V., Vavoulis, D. V., et al. (2023) 'Structural and non-coding variants increase the diagnostic yield of clinical whole genome sequencing for rare diseases', *Genome Medicine*, 15(1), p. 94.

Palacios, V., Kimble, G.C., Tootle, T.L. & Buszczak, M. (2021) 'Importin-9 regulates chromosome segregation and packaging in *Drosophila* germ cells', *Journal of Cell Science*, 134(7).

PALERMO, G. (1992) ‘Pregnancies after intracytoplasmic injection of single spermatozoon into an oocyte’, *The Lancet*, 340(8810), pp. 17-18.

Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J., Rafiq, M.A., Conrad, D.F., Park, H., Hurles, M.E., Lee, C., Venter, J.C., Kirkness, E.F., Levy, S., Feuk, L. & Scherer, S.W. (2010) ‘Towards a comprehensive structural variation map of an individual human genome’, *Genome Biology*, 11(5).

Paparella, A., L’Abbate, A., Palmisano, D., Chirico, G., Porubsky, D., Catacchio, C.R., Ventura, M., Eichler, E.E., Maggiolini, F.A.M. & Antonacci, F. (2023) ‘Structural Variation Evolution at the 15q11-q13 Disease-Associated Locus’, *International Journal of Molecular Sciences*, 24(21), p. 15818.

Patsalis, P.C., Sismani, C., Quintana-Murci, L., Taleb-Bekkouche, F., Krausz, C. & McElreavey, K. (2002) ‘Effects of transmission of Y chromosome AZFc deletions’, *The Lancet*, 360(9341).

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

Pei, Y., Tanguy, M., Giess, A., Dixit, A., Wilson, L.C., Gibbons, R.J., Twigg, S.R.F., Elgar, G. & Wilkie, A.O.M. (2024) ‘A Comparison of Structural Variant Calling from Short-Read and Nanopore-Based Whole-Genome Sequencing Using Optical Genome Mapping as a Benchmark’, *Genes*, 15(7), p. 925.

Perez, G., Barber, G.P., Benet-Pages, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, C.M., Nassar, L.R., Raney, B.J., Speir, M.L., van Baren, M.J., Vaske, C.J., Haussler, D., Kent, W.J. & Haeussler, M. (2024) ‘The UCSC Genome Browser database: 2025 update’, *Nucleic Acids Research*.

Di Persio, S. & Neuhaus, N. (2023) ‘Human spermatogonial stem cells and their niche in male (in)fertility: novel concepts from single-cell RNA-sequencing’, *Human Reproduction*, 38(1), pp. 1-13.

Picton, H.M., Wyns, C., Anderson, R.A., Goossens, E., Jahnukainen, K., Kliesch, S., Mitchell, R.T., Pennings, G., Rives, N., Tournaye, H., van Pelt, A.M.M., Eichenlaub-Ritter, U. & Schlatt, S. (2015) ‘A European perspective on testicular tissue cryopreservation for fertility preservation in prepubertal and adolescent boys’, *Human Reproduction*, 30(11), pp. 2463-2475.

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A.C., Thiruvahindrapuram, B., MacDonald, J.R., Mills, R., Prasad, A., Noonan, K., Gribble, S., Prigmore, E., Donahoe, P.K., Smith, R.S., Park, J.H., Hurles, M.E., Carter, N.P., et al. (2011) 'Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants', *Nature Biotechnology*, 29(6), pp. 512-520.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., Almeida, J., Bacchelli, E., Bader, G.D., Bailey, A.J., Baird, G., Battaglia, A., Berney, T., Bolshakova, N., Bölte, S., et al. (2010) 'Functional impact of global rare copy number variation in autism spectrum disorders', *Nature*, 466(7304).

Pollard, M.O., Gurdasani, D., Mentzer, A.J., Porter, T. & Sandhu, M.S. (2018) 'Long reads: their purpose and place', *Human Molecular Genetics*, 27(R2).

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

Punjani, N., Kang, C., Lamb, D.J. & Schlegel, P.N. (2021) 'Current updates and future perspectives in the evaluation of azoospermia: A systematic review', *Arab Journal of Urology*, 19(3), pp. 206-214.

Qin, J., Sheng, X., Wang, H., Liang, D., Tan, H. & Xia, J. (2015) 'Assisted reproductive technology and risk of congenital malformations: a meta-analysis based on cohort studies', *Archives of Gynecology and Obstetrics*, 292(4), pp. 777-798.

Quaynor, S.D., Bosley, M.E., Duckworth, C.G., Porter, K.R., Kim, S.-H., Kim, H.-G., Chorich, L.P., Sullivan, M.E., Choi, J.-H., Cameron, R.S. & Layman, L.C. (2016) 'Targeted next generation sequencing approach identifies eighteen new candidate genes in normosmic hypogonadotropic hypogonadism and Kallmann syndrome', *Molecular and Cellular Endocrinology*, 437pp. 86-96.

Quinodoz, M., Royer-Bertrand, B., Cisarova, K., Di Gioia, S.A., Superti-Furga, A. & Rivolta, C. (2017) 'DOMINO: Using Machine Learning to Predict Genes Associated with Dominant Disorders', *The American Journal of Human Genetics*, 101(4), pp. 623-629.

R Core Team (2021) *R: A Language and Environment for Statistical Computing*.

Rajender, S. (2012) 'Apoptosis spermatogenesis and male infertility', *Frontiers in Bioscience*, E4(2), p. 415.

Ramasamy, R., Bakırcıoğlu, M.E., Cengiz, C., Karaca, E., Scovell, J., Jhangiani, S.N., Akdemir, Z.C., Bainbridge, M., Yu, Y., Huff, C., Gibbs, R.A., Lupski, J.R. & Lamb, D.J. (2015) 'Whole-exome sequencing identifies novel homozygous mutation in NPAS2 in family with nonobstructive azoospermia', *Fertility and Sterility*, 104(2), pp. 286-291.

Rato, L., Alves, M.G., Socorro, S., Duarte, A.I., Cavaco, J.E. & Oliveira, P.F. (2012) 'Metabolic regulation is important for spermatogenesis', *Nature Reviews Urology*, 9(6), pp. 330-338.

Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V. & Korbel, J.O. (2012) 'DELLY: structural variant discovery by integrated paired-end and split-read analysis', *Bioinformatics*, 28(18).

Ravel, C., Chantot-Bastaraud, S., El Houate, B., Berthaut, I., Verstraete, L., De Larouziere, V., Lourenço, D., Dumaine, A., Antoine, J.M., Mandelbaum, J., Siffroi, J.P. & McElreavey, K. (2007) 'Mutations in the protamine 1 gene associated with male infertility', *MHR: Basic science of reproductive medicine*, 13(7), pp. 461-464.

Redin, C., Brand, H., Collins, R.L., Kammin, T., Mitchell, E., Hodge, J.C., Hanscom, C., Pillalamarri, V., Seabra, C.M., Abbott, M.-A., Abdul-Rahman, O.A., Aberg, E., Adley, R., Alcaraz-Estrada, S.L., Alkuraya, F.S., An, Y., Anderson, M.-A., Antolik, C., Anyane-Yeboa, K., et al. (2017) 'The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies', *Nature Genetics*, 49(1), pp. 36-45.

Riemondy, K.A., Sheridan, R.M., Gillen, A., Yu, Y., Bennett, C.G. & Hesselberth, J.R. (2017) 'valr: Reproducible genome interval analysis in R', *F1000Research*, 6p. 1025.

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

Riggs, E.R., Andersen, E.F., Cherry, A.M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D.I., South, S.T., Thorland, E.C., Pineda-Alvarez, D., Aradhya, S. & Martin, C.L. (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).

Ritz, A., Bashir, A., Sindi, S., Hsu, D., Hajirasouliha, I. & Raphael, B.J. (2014) 'Characterization of structural variants with single molecule and hybrid sequencing approaches', *Bioinformatics*, 30(24), .

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. & Mesirov, J.P. (2011) 'Integrative genomics viewer', *Nature Biotechnology*, 29(1).

Rogers, M.J. (2021) 'Y chromosome copy number variation and its effects on fertility and other health factors: a review', *Translational Andrology and Urology*, 10(3), pp. 1373-1382.

Rotimi, D.E., Iyobhebhe, M., Oluwayemi, E.T., Evbuomwan, I.O., Asaleye, R.M., Ojo, O.A. & Adeyemi, O.S. (2024) 'Mitophagy and spermatogenesis: Role and mechanisms', *Biochemistry and Biophysics Reports*, 38p. 101698.

Rozen, S.G., Marszalek, J.D., Irenze, K., Skaletsky, H., Brown, L.G., Oates, R.D., Silber, S.J., Ardlie, K. & Page, D.C. (2012) 'AZFc Deletions and Spermatogenic Failure: A Population-Based Survey of 20,000 Y Chromosomes', *The American Journal of Human Genetics*, 91(5), pp. 890-896.

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

van der Sanden, B.P.G.H., Schobers, G., Corominas Galbany, J., Koolen, D.A., Sinnema, M., van Reeuwijk, J., Stumpel, C.T.R.M., Kleefstra, T., de Vries, B.B.A., Ruiterkamp-Versteeg, M., Leijsten, N., Kwint, M., Derks, R., Swinkels, H., den Ouden, A., Pfundt, R., Rinne, T., de Leeuw, N., Stegmann, A.P., et al. (2023) 'The performance of genome sequencing as a first-tier test for neurodevelopmental disorders', *European Journal of Human Genetics*, 31(1), pp. 81-88.

Sarwal, V., Niehus, S., Ayyala, R., Chang, S., Lu, A., Darci-Maher, N., Littman, R., Chhugani, K., Soylev, A., Comarova, Z., Wesel, E., Castellanos, J., Chikka, R., Distler, M.G., Eskin, E., Flint, J. & Mangul, S. (2020) 'A comprehensive benchmarking of WGS-based structural variant callers', *bioRxiv*, p. 2020.04.16.045120.

Savara, J., Novosád, T., Gajdoš, P. & Kriegová, E. (2021) 'Comparison of structural variants detected by optical mapping with long-read next-generation sequencing', *Bioinformatics*.

Sazci, A., Ercelen, N., Ergul, E. & Akpinar, G. (2005) 'Male factor infertility associated with a familial translocation t(1;13)(q24;q10)', *Fertility and Sterility*, 83(5), pp. 1548.e19-1548.e21.

Schrauwen, I., Rajendran, Y., Acharya, A., Öhman, S., Arvio, M., Paetau, R., Siren, A., Avela, K., Granvik, J., Leal, S.M., Määttä, T., Kokkonen, H. & Järvelä, I. (2024) 'Optical genome mapping unveils hidden structural variants in neurodevelopmental disorders', *Scientific Reports*, 14(1), p. 11239.

Schulster, M., Bernie, A. & Ramasamy, R. (2016) 'The role of estradiol in male reproductive function', *Asian Journal of Andrology*, 18(3), p. 435.

Schuppe, H.-C., Pilatz, A., Hossain, H., Diemer, T., Wagenlehner, F. & Weidner, W. (2017) 'Urogenital Infection as a Risk Factor for Male Infertility', *Deutsches Aerzteblatt Online*.

Scully, R., Panday, A., Elango, R. & Willis, N.A. (2019) 'DNA double-strand break repair-pathway choice in somatic mammalian cells', *Nature Reviews Molecular Cell Biology*, 20(11).

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.-H., Hicks, J., Spence, S.J., Lee, A.T., Puura, K., Lehtimäki, T., et al. (2007) 'Strong Association of De Novo Copy Number Mutations with Autism', *Science*, 316(5823), pp. 445-449.

Selvaraj, S., R Dixon, J., Bansal, V. & Ren, B. (2013) 'Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing', *Nature Biotechnology*, 31(12).

Sethi, R., Becker, J., Graaf, J. de, Löwer, M., Suchan, M., Sahin, U. & Weber, D. (2020) 'Integrative analysis of structural variations using short-reads and linked-reads yields highly specific and sensitive predictions', *PLOS Computational Biology*, 16(11).

Shaini Joseph & Smita D Mahale (2021) 'Male Infertility Knowledgebase: decoding the genetic and disease landscape', *Database*.

Sharma, P., Kaushal, N., Saleth, L.R., Ghavami, S., Dhingra, S. & Kaur, P. (2023) 'Oxidative stress-induced apoptosis and autophagy: Balancing the contrary forces in spermatogenesis', *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1869(6), p. 166742.

Sharma Rakesh and Agarwal, A. (2011) 'Spermatogenesis: An Overview', in Ashok Zini Armand and Agarwal (ed.) *Sperm Chromatin: Biological and Clinical Applications in Male Infertility and Assisted Reproduction*. [Online]. New York, NY: Springer New York. pp. 19-44.

Shen, Y., Yan, Y., Liu, Y., Zhang, S., Yang, D., Zhang, P., Li, L., Wang, Y., Ma, Y., Tao, D. & Yang, Y. (2013) 'A significant effect of the TSPY1 copy number on spermatogenesis efficiency and the phenotypic expression of the gr/gr deletion', *Human Molecular Genetics*, 22(8).

Shieh, J.T., Penon-Portmann, M., Wong, K.H.Y., Levy-Sakin, M., Verghese, M., Slavotinek, A., Gallagher, R.C., Mendelsohn, B.A., Tenney, J., Beleford, D., Perry, H., Chow, S.K., Sharo, A.G., Brenner, S.E., Qi, Z., Yu, J., Klein, O.D., Martin, D., Kwok, P.-Y., et al. (2021) 'Application of full-genome analysis to diagnose rare monogenic disorders', *npj Genomic Medicine*, 6(1), p. 77.

Siasi, E., Aleyasin, A., Mowla, S.J. & Sahebkashaf, H. (2011) 'Study of GT-repeat expansion in Heme oxygenase-1 gene promoter as genetic cause of male infertility', *Journal of Assisted Reproduction and Genetics*, 28(8).

Singh, B.K. & Kambayashi, T. (2016) 'The Immunomodulatory Functions of Diacylglycerol Kinase ζ', *Frontiers in Cell and Developmental Biology*, 4.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., Chinwalla, A., Delehaunty, A., Delehaunty, K., Du, H., Fewell, G., Fulton, L., Fulton, R., Graves, T., Hou, S.-F., 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.

Skidmore, Z.L., Wagner, A.H., Lesurf, R., Campbell, K.M., Kunisaki, J., Griffith, O.L. & Griffith, M. (2016) 'GenVisR: Genomic Visualizations in R', *Bioinformatics*, 32p. 3014.

Slatko, B.E., Gardner, A.F. & Ausubel, F.M. (2018) 'Overview of Next-Generation Sequencing Technologies', *Current Protocols in Molecular Biology*, 122(1).

Smith, K.D., Steinberger, E., Steinberger, A. & Perloff, W.H. (1965) 'A Familial Centric Chromosome Fragment', *Cytogenetic and Genome Research*, 4(4-5), pp. 219-226.

Sobreira, N., Schiettecatte, F., Valle, D., & Hamosh, A. (2015). GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. Human Mutation, 36(10), 928–930.

Spielmann, M., Lupiáñez, D.G. & Mundlos, S. (2018) 'Structural variation in the 3D genome', *Nature Reviews Genetics*, 19(7), pp. 453-467.

Srikanth, K., Park, J.-E., Lim, D., Cha, J., Cho, S.-R., Cho, I.-C. & Park, W. (2020) 'A Comparison between Hi-C and 10X Genomics Linked Read Sequencing for Whole Genome Phasing in Hanwoo Cattle', *Genes*, 11(3).

Stallmeyer, B., Dicke, A. & Tüttelmann, F. (2024) 'How exome sequencing improves the diagnostics and management of men with non-syndromic infertility', *Andrology*.

Stankiewicz, P. & Lupski, J.R. (2010) 'Structural Variation in the Human Genome and its Role in Disease', *Annual Review of Medicine*, 61(1).

Staňková, H., Hastie, A.R., Chan, S., Vrána, J., Tulpová, Z., Kubaláková, M., Visendi, P., Hayashi, S., Luo, M., Batley, J., Edwards, D., Doležel, J. & Šimková, H. (2016) 'BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes', *Plant Biotechnology Journal*, 14(7).

Steptoe, P.C. & Edwards, R.G. (1978) 'BIRTH AFTER THE REIMPLANTATION OF A HUMAN EMBRYO', *The Lancet*, 312(8085), p. 366.

Stouffs, K., Vandermaelen, D., Massart, A., Menten, B., Vergult, S., Tournaye, H. & Lissens, W. (2012) 'Array comparative genomic hybridization in male infertility', *Human Reproduction*, 27(3), pp. 921-929.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M.K., Malhotra, A., Stütz, A.M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., et al. (2015a) 'An integrated map of structural variation in 2,504 human genomes', *Nature*, 526(7571).

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M.K., Malhotra, A., Stütz, A.M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., et al. (2015b) 'An integrated map of structural variation in 2,504 human genomes', *Nature*, 526(7571), pp. 75-81.

Sun, F., Turek, P., Greene, C., Ko, E., Rademaker, A. & Martin, R.H. (2007) 'Abnormal progression through meiosis in men with nonobstructive azoospermia', *Fertility and Sterility*, 87(3), pp. 565-571.

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., Bork, P., Jensen, L.J. & von Mering, C. (2023) 'The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest', *Nucleic Acids Research*, 51(D1), pp. D638-D646.

Tang, S., Wang, X., Li, W., Yang, X., Li, Z., Liu, W., Li, C., Zhu, Z., Wang, L., Wang, Jiaxiong, Zhang, L., Sun, X., Zhi, E., Wang, H., Li, H., Jin, L., Luo, Y., Wang, Jian, Yang, S., et al. (2017) 'Biallelic Mutations in CFAP43 and CFAP44 Cause Male Infertility with Multiple Morphological Abnormalities of the Sperm Flagella', *The American Journal of Human Genetics*, 100(6), pp. 854-864.

Tannour-Louet, M., Han, S., Louet, J.-F., Zhang, B., Romero, K., Addai, J., Sahin, A., Cheung, S.W. & Lamb, D.J. (2014) 'Increased gene copy number of VAMP7 disrupts human male urogenital development through altered estrogen action', *Nature Medicine*, 20(7), pp. 715-724.

Tian, S., Tu, C., He, X., Meng, L., Wang, J., Tang, S., Gao, Y., Liu, C., Wu, H., Zhou, Y., Lv, M., Lin, G., Jin, L., Cao, Y., Tang, D., Zhang, F. & Tan, Y.-Q. (2023) 'Biallelic mutations in *CFAP54* cause male infertility with severe MMAF and NOA', *Journal of Medical Genetics*, 60(8), pp. 827-834.

Tian, S., Wang, Z., Liu, L., Zhou, Y., Lv, Y., Tang, D., Wang, J., Jiang, J., Wu, H., Tang, S., Wang, G., Geng, H., Tao, F., Liu, H., He, X., Zhang, F., Li, J., Jin, L., Huang, T., et al. (2023) 'A homozygous frameshift mutation in ADAD2 causes male infertility with spermatogenic impairments', *Journal of Genetics and Genomics*, 50(4), pp. 284-288.

Toragall, M.M., Satapathy, S.K., Kadadevaru, G.G. & Hiremath, M.B. (2018) 'Assessment of Hormone Levels in Men With Fertility Problems Attending a Tertiary Infertility Center in North Karnataka', *International Journal of Scientific Research in Biological Sciences*.

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

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

Tüttelmann, Frank, Simoni, M., Kliesch, S., Ledig, S., Dworniczak, B., Wieacker, P. & Röpke, A. (2011) 'Copy Number Variants in Patients with Severe Oligozoospermia and Sertoli-Cell-Only Syndrome', *PLoS ONE*, 6(4).

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

Tzschach, A., Ramel, C., Kron, A., Seipel, B., Wüster, C., Cordes, U., Liehr, T., Hoeltzenbein, M., Menzel, C., Ropers, H. -H., Ullmann, R., Kalscheuer, V., Decker, J. & Steinberger, D. (2009) 'Hypergonadotropic hypogonadism in a patient with inv ins (2;4)', *International Journal of Andrology*, 32(3), pp. 226-230.

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

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

Vahedi Raad, M., Firouzabadi, A.M., Tofighi Niaki, M., Henkel, R. & Fesahat, F. (2024) 'The impact of mitochondrial impairments on sperm function and male fertility: a systematic review', *Reproductive Biology and Endocrinology*, 22(1), p. 83.

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

Veltman, J.A. & Tüttelmann, F. (2024) 'Why geneticists should care about male infertility', *Nature Reviews Genetics*, 25(12), pp. 823-824.

Vertika, S., Singh, K.K. & Rajender, S. (2020) 'Mitochondria, spermatogenesis, and male infertility - An update', *Mitochondrion*, 54pp. 26-40.

Vincent, M., DAUDIN, M., DE MAS, P., MASSAT, G., MIEUSSET, R., PONTONNIER, F., CALVAS, P., BUJAN, L. & BOURROUILLOU, G. (2002) 'Cytogenetic Investigations of Infertile Men With Low Sperm Counts: A 25-Year Experience', *Journal of Andrology*, 23(1), pp. 18-22.

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

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

Vogt, P. (1996) 'Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11', *Human Molecular Genetics*, 5(7), pp. 933-943.

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

Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., Nord, A.S., Kusenda, M., Malhotra, D., Bhandari, A., Stray, S.M., Rippey, C.F., Roccanova, P., Makarov, V., Lakshmi, B., Findling, R.L., Sikich, L., Stromberg, T., Merriman, B., et al. (2008) 'Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia', *Science*, 320(5875).

Wang, C., Lv, H., Ling, X., Li, H., Diao, F., Dai, J., Du, J., Chen, T., Xi, Q., Zhao, Y., Zhou, K., Xu, B., Han, X., Liu, X., Peng, M., Chen, C., Tao, S., Huang, L., Liu, C., et al. (2021) 'Association of assisted reproductive technology, germline de novo mutations and congenital heart defects in a prospective birth cohort study', *Cell Research*, 31(8).

Wang, G., Li, Y., Xu, S., Ma, S., Yan, R., Zhang, R., Jia, G.-X., Ai, D. & Yang, Q.-E. (2019) 'Gene Expression Dynamics During the Gonocyte to Spermatogonia Transition and Spermatogenesis in the Domestic Yak', *Journal of Animal Science and Biotechnology*.

Wang, J., Tang, C., Wang, Q., Su, J., Ni, T., Yang, W., Wang, Yongsheng, Chen, W., Liu, X., Wang, S., Zhang, J., Song, H., Zhu, J. & Wang, Yuan (2017) 'NRF1 Coordinates With DNA Methylation to Regulate Spermatogenesis', *The Faseb Journal*.

Wang, M. & Su, P. (2018) 'The role of the Fas/FasL signaling pathway in environmental toxicant-induced testicular cell apoptosis: An update', *Systems Biology in Reproductive Medicine*, 64(2), pp. 93-102.

Wang, W., Lu, N., Xia, Y., Gu, A., Wu, B., Liang, J., Zhang, W., Wang, Z., Su, J. & Wang, X. (2009) 'FAS and FASLG polymorphisms and susceptibility to idiopathic azoospermia or severe oligozoospermia', *Reproductive BioMedicine Online*, 18(1), pp. 141-147.

Wang, Y.-H., Yan, M., Zhang, X., Liu, X.-Y., Ding, Y.-F., Lai, C.-P., Tong, M.-H. & Li, J.-S. (2021) 'Rescue of male infertility through correcting a genetic mutation causing meiotic arrest in spermatogonial stem cells', *Asian Journal of Andrology*, 23(6), pp. 590-599.

Weckselblatt, B. & Rudd, M.K. (2015) 'Human Structural Variation: Mechanisms of Chromosome Rearrangements', *Trends in Genetics*, 31(10).

Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J.O. (2013) 'Phenotypic impact of genomic structural variation: insights from and for human disease', *Nature Reviews Genetics*, 14(2), pp. 125-138.

Welsh, M., Saunders, P.T.K., Atanassova, N., Sharpe, R.M. & Smith, L.B. (2009) 'Androgen action *via* testicular peritubular myoid cells is essential for male fertility', *The FASEB Journal*, 23(12), pp. 4218-4230.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., et al. (2019) 'Welcome to the Tidyverse', *Journal of Open Source Software*, 4(43), p. 1686.

Wickham, H., François, R., Henry, L. & Müller, K. (2021) *dplyr: A Grammar of Data Manipulation*.

Wong, L.-P., Ong, R.T.-H., Poh, W.-T., Liu, X., Chen, P., Li, R., Lam, K.K.-Y., Pillai, N.E., Sim, K.-S., Xu, H., Sim, N.-L., Teo, S.-M., Foo, J.-N., Tan, L.W.-L., Lim, Y., Koo, S.-H., Gan, L.S.-H., Cheng, C.-Y., Wee, S., et al. (2013) 'Deep Whole-Genome Sequencing of 100 Southeast Asian Malays', *The American Journal of Human Genetics*, 92(1), pp. 52-66.

World Health Organization (2021) *WHO laboratory manual for the examination and processing of human semen, sixth edition.* Geneva: World Health Organization.

Wu, S., Wan, R., Wang, G., Zhang, Y. & Yang, Q. (2023) 'Comparative proteomic analysis identifies differentially expressed proteins associated with meiotic arrest in cattle-yak hybrids', *PROTEOMICS*, 23(12).

Wyrwoll, M.J., van der Heijden, G.W., Krausz, C., Aston, K.I., Kliesch, S., McLachlan, R., Ramos, L., Conrad, D.F., O'Bryan, M.K., Veltman, J.A. & Tüttelmann, F. (2024) 'Improved phenotypic

classification of male infertility to promote discovery of genetic causes', *Nature Reviews Urology*, 21(2), pp. 91-101.

Wyrwoll, M.J., Köckerling, N., Vockel, M., Dicke, A.-K., Rotte, N., Pohl, E., Emich, J., Wöste, M., Ruckert, C., Wabschke, R., Seggewiss, J., Ledig, S., Tewes, A.-C., Stratis, Y., Cremers, J.F., Wistuba, J., Krallmann, C., Kliesch, S., Röpke, A., et al. (2023) 'Genetic Architecture of Azoospermia—Time to Advance the Standard of Care', *European Urology*, 83(5), pp. 452-462.

Wyrwoll, M.J., Temel, Ş.G., Nagirnaja, L., Oud, M.S., Lopes, A.M., van der Heijden, G.W., Heald, J.S., Rotte, N., Wistuba, J., Wöste, M., Ledig, S., Krenz, H., Smits, R.M., Carvalho, F., Gonçalves, J., Fietz, D., Türkgenç, B., Ergören, M.C., Çetinkaya, M., et al. (2020) 'Bi-allelic Mutations in M1AP Are a Frequent Cause of Meiotic Arrest and Severely Impaired Spermatogenesis Leading to Male Infertility', *The American Journal of Human Genetics*, 107(2), pp. 342-351.

Wyrwoll, M.J., Gaasbeek, C.M., Golubickaite, I., Stakaitis, R., Oud, M.S., Nagirnaja, L., Dion, C., Sindi, E.B., Leitch, H.G., Jayasena, C.N., Sironen, A., Dicke, A.-K., Rotte, N., Stallmeyer, B., Kliesch, S., Grangeiro, C.H.P., Araujo, T.F., Lasko, P., D'Hauwers, K., et al. (2022) 'The piRNA-pathway factor FKBP6 is essential for spermatogenesis but dispensable for control of meiotic LINE-1 expression in humans', *The American Journal of Human Genetics*, 109(10), pp. 1850-1866.

Wyrwoll, M.J., Wabschke, R., Röpke, A., Wöste, M., Ruckert, C., Perrey, S., Rotte, N., Hardy, J., Astica, L., Lupiáñez, D.G., Wistuba, J., Westernströer, B., Schlatt, S., Berman, A.J., Müller, A.M., Kliesch, S., Yatsenko, A.N., Tüttelmann, F. & Friedrich, C. (2022) 'Analysis of copy number variation in men with non-obstructive azoospermia', *Andrology*, 10(8), pp. 1593-1604.

Xavier, M.J., Salas-Huetos, A., Oud, M.S., Aston, K.I. & Veltman, J.A. (2021) 'Disease gene discovery in male infertility: past, present and future', *Human Genetics*, 140(1).

Xiao, F., Lan, A., Lin, Z., Song, J., Zhang, Y., Li, J., Gu, K., Lv, B., Zhao, D., Zeng, S., Zhang, R., Zhao, W., Pan, Z., Deng, X. & Yang, X. (2016) 'Impact of CAG repeat length in the androgen receptor gene on male infertility - a meta-analysis', *Reproductive BioMedicine Online*, 33(1), pp. 39-49.

Xiao, Q. & Lauschke, V.M. (2021) 'The prevalence, genetic complexity and population-specific founder effects of human autosomal recessive disorders', *npj Genomic Medicine*, 6(1), p. 41.

Xin, X., Xu, P., Wang, N., Jiang, Y., Zhang, J., Li, S., Zhu, Y., Zhang, C., Zhang, L., Huang, H., Feng, L. & Wang, S. (2023) 'Copy number variations (CNVs) and karyotyping analysis in males with azoospermia and oligospermia', *BMC Medical Genomics*, 16(1), p. 213.

Xu, F., Guo, G., Zhu, F., Tan, X. & Fan, L. (2021) 'Protein deep profile and model predictions for identifying the causal genes of male infertility based on deep learning', *Information Fusion*, 75pp. 70-89.

Yao, R.A., Akinrinade, O., Chaix, M. & Mital, S. (2020) 'Quality of whole genome sequencing from blood versus saliva derived DNA in cardiac patients', *BMC Medical Genomics*, 13(1), p. 11.

Yatsenko, A.N., Georgiadis, A.P., Röpke, A., Berman, A.J., Jaffe, T., Olszewska, M., Westernströer, B., Sanfilippo, J., Kurpisz, M., Rajkovic, A., Yatsenko, S.A., Kliesch, S., Schlatt, S. & Tüttelmann, F. (2015) 'X-Linked *TEX11* Mutations, Meiotic Arrest, and Azoospermia in Infertile Men', *New England Journal of Medicine*, 372(22), pp. 2097-2107.

Yuan, Y., Chung, C.Y.-L. & Chan, T.-F. (2020) 'Advances in optical mapping for genomic research', *Computational and Structural Biotechnology Journal*, 18.

Yue, J.-X. & Liti, G. (2019) 'simuG: a general-purpose genome simulator', *Bioinformatics*, 35(21), pp. 4442-4444.

Zare, F., Dow, M., Monteleone, N., Hosny, A. & Nabavi, S. (2017) 'An evaluation of copy number variation detection tools for cancer using whole exome sequencing data', *BMC Bioinformatics*, 18(1).

Zarrei, M., MacDonald, J.R., Merico, D. & Scherer, S.W. (2015) 'A copy number variation map of the human genome', *Nature Reviews Genetics*, 16(3), pp. 172-183.

Zegers-Hochschild, F., Adamson, G.D., de Mouzon, J., Ishihara, O., Mansour, R., Nygren, K., Sullivan, E. & van der Poel, S. (2009) 'The International Committee for Monitoring Assisted Reproductive Technology (ICMART) and the World Health Organization (WHO) Revised Glossary on ART Terminology, 2009', *Human Reproduction*, 24(11).

Zenteno-Ruiz, J.C., Kofman-Alfaro, S. & Méndez, J.P. (2001) '46,XX Sex Reversal', *Archives of Medical Research*, 32(6), pp. 559-566.

Zhang, H., Li, W., Jiang, Y., Li, J., Chen, M., Wang, R., Zhao, J., Peng, Z., Huang, H. & Liu, R. (2022) 'Whole Exome Sequencing Identifies Genes Associated With Non-Obstructive Azoospermia', *Frontiers in Genetics*, 13.

Zhang, Y.-S., Dai, R.-L., Wang, R.-X., Zhang, H.-G., Chen, S. & Liu, R.-Z. (2013) 'Analysis of Y Chromosome Microdeletion in 1738 Infertile Men From Northeastern China', *Urology*, 82(3).

Zhang, S., He, Y., Huang, Y., Zhang, Y., Du, Y., Zhang, T., Sun, Y., & Lu, Y. (2025). The accuracy and real resolution of karyotyping technique in detecting chromosomal aberrations identified by molecular genetic methods. Molecular Genetics and Genomics, 300(1), 79.

Zhao, M., Wang, Qingguo, Wang, Quan, Jia, P. & Zhao, Z. (2013) 'Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives', *BMC Bioinformatics*, 14(S11).

Zheng, G., Dahl, J.A., Niu, Y., Fedorcsák, P., Huang, C.-H., Li, C.J., Vågbø, C.B., Shi, Y., Wang, W., Song, S., Lu, Z., Bosmans, R.P.G., Dai, Q., Hao, Y., Yang, X., Zhao, W., Tong, W., Wang, X., Bogdan, F., et al. (2013) 'ALKBH5 Is a Mammalian RNA Demethylase That Impacts RNA Metabolism and Mouse Fertility', *Molecular Cell*.

Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., Mudivarti, P.A., Wyatt, P.W., Bharadwaj, R., Makarewicz, A.J., Li, Y., Belgrader, P., Price, A.D., Lowe, A.J., Marks, P., et al. (2016) 'Haplotyping germline and cancer genomes with high-throughput linked-read sequencing', *Nature Biotechnology*, 34(3).

Zhou, H., Yin, Z., Ni, B., Lin, J., Luo, S. & Xie, W. (2024) 'Whole exome sequencing analysis of 167 men with primary infertility', *BMC Medical Genomics*, 17(1), p. 230.

Zhou, R., Wu, J., Liu, B., Jiang, Y., Chen, W., Li, J., He, Q. & He, Z. (2019) 'The roles and mechanisms of Leydig cells and myoid cells in regulating spermatogenesis', *Cellular and Molecular Life Sciences*, 76(14), pp. 2681-2695.

Zook, J.M., Hansen, N.F., Olson, N.D., Chapman, L., Mullikin, J.C., Xiao, C., Sherry, S., Koren, S., Phillippy, A.M., Boutros, P.C., Sahraeian, S.M.E., Huang, V., Rouette, A., Alexander, N., Mason, C.E., Hajirasouliha, I., Ricketts, C., Lee, J., Tearle, R., et al. (2020) 'A robust benchmark for detection of germline large deletions and insertions', *Nature Biotechnology*, 38(11), pp. 1347-1355.