Towards the identification of oligogenic disease variants at ecome scale

From oligogenic data to a new priorisation algorithm

Gravel, Barbara

Publication date:
2024

Document Version:
Final published version

Towards the identification of oligogenic disease variants at exome scale - from oligogenic data to a new prioritisation algorithm.

Thesis submitted by Barbara GRAVEL

in fulfilment of the requirements of the PhD Degree in Sciences (ULB - "Doctorat en Sciences ") and in Sciences (VUB)

Academic year 2023-2024

Supervisors: Professor Tom LENAERTS (Université Libre de Bruxelles)
Interuniversity Institute of Bioinformatics in Brussels - Machine Learning Group
and Professor Ann NOWÉ (Vrije Universiteit Brussel)
Artificial Intelligence Lab

Thesis Jury :

Prof. Matthieu Defrance (Université libre de Bruxelles, Chair)
Prof. Wim Vranken (Vrije Universiteit Brussel, Secretary)
Prof. Guillaume Smits (Université Libre de Bruxelles)
Prof. Catharina Olsen (Vrije Universiteit Brussel)
Prof. Yves Moreau (Katholieke Universiteit Leuven)
Prof. Emmanuelle Génin (Université de Bretagne Occidentale)

This PhD thesis has been conducted and written under the supervision of prof. dr. Tom Lenaerts (Université Libre de Bruxelles) and prof. dr. Ann Nowé (Vrije Universiteit Brussel).

The public defense took place on 24th September 2024, at Université Libre de Bruxelles.

Abstract

English

With the advent of high-throughput sequencing technologies, tremendous progress has been made in understanding the relationship between genotypes and phenotypes. Advances in related computational methods have enabled the development of different prioritization methods that help identify in Whole-Exome Sequencing (WES) data which genetic variants are responsible for particular disease phenotypes. However, these methods overlook the fact that a significant proportion of genetic diseases do not follow monogenic inheritance patterns, but are caused by the interaction of variants in a small number of genes. Developping novel computational methods to identifying these more complex combinations of genetic variants, known as oligogenic inheritance, is therefore essential.

In this thesis, we build upon existing approaches to detect oligogenic inheritance models to make this analysis possible at the whole-exome level. First, we develop a novel database that collects information on all oligogenic variant combinations reported in the literature. This database not only aggregates existing knowledge but also introduces a standardized framework for assessing the pathogenicity of variant combinations. Using this database, we develop a first oligogenic priorization tool: the High-throughput oligogenic prioritizer (Hop). This predictor integrates pathogenicity scoring, from a machine learning predictor specific to variant combinations, together with disease relevance scoring, based on knowledge propagation in biological networks, to rank variant combinations based on how likely they are to explain a patient's disease. This tool demonstrates superior performance to existing approaches for ranking oligogenic combinations in exomes. Finally, we investigate the usefulness of these computational tools on real patient data. We apply the Hop predictor on a cohort of patients affected with male infertility, a disease with heterogeneous genetic causes, and investigate the relevance of the prioritized combinations. This analysis validates the ability of Hop to detect oligogenic combinations that were manually identified by clinicians, and also showcases its capacity to identify novel oligogenic signatures in this disease.

In summary, this research demonstrates that it is now possible to directly detect disease causing variant combinations in whole exome sequencing data using computational approaches. By introducing a new data repository, computational tools, and analysis protocols, this research opens the way for easier detection and analysis of oligogenic signatures for genetic diseases.

French

L'avènement des technologies de séquençage à haut débit a permis de réaliser des progrès considérables dans la compréhension de la relation entre les génotypes et les phénotypes. Les progrès des méthodes informatiques ont permis le développement de différents outils de priorisation qui facilitent l'identification, dans les données de séquençage de l'ensemble de l'exome (WES), des variants génétiques responsables des phénotypes liés à certaines maladies. Toutefois, ces méthodes ne tiennent pas compte du fait qu'une proportion importante de maladies génétiques ne suivent pas des modèles de transmission monogéniques, mais sont en fait causées par la combinaison de variants dans un petit nombre de gènes. Il est donc essentiel de développer de nouvelles méthodes informatiques pour identifier ces combinaisons plus complexes de variants génétiques, connues sous le nom d'hérédité oligogénique.

Dans cette thèse, nous nous appuyons sur des approches existantes pour détecter les modèles de transmission oligogénique afin de rendre cette analyse possible au niveau de l'exome. Tout d'abord, nous développons une nouvelle base de données qui rassemble des informations sur toutes les combinaisons de variants oligogéniques rapportées dans la littérature. Cette base de données ne se contente pas d'agréger les connaissances existantes, mais introduit également un système de standards pour l'évaluation de la pathogénicité des combinaisons de variants. En utilisant cette base de données, nous développons un premier outil de priorisation oligogénique : le High-throughput oligogenic prioritizer (Hop). Ce prédicteur intègre un score de pathogénicité, issu d'un prédicteur d'apprentissage automatique spécifique aux combinaisons de variants, ainsi qu'un score de pertinence pour la maladie, basé sur la propagation de connaissances dans les réseaux biologiques, afin de hiérarchiser les combinaisons de variants en fonction de leur probabilité d'expliquer la maladie d'un patient. Cet outil démontre des performances supérieures aux approches existantes pour la priorisation des combinaisons oligogéniques dans les exomes. Enfin, nous étudions l'utilité de ces outils informatiques sur des données réelles de patients. Nous appliquons le prédicteur Hop à une cohorte de patients atteints d'infertilité masculine, une maladie dont les causes génétiques sont hétérogènes, et nous étudions la pertinence des combinaisons classées par ordre de priorité. Cette analyse valide la capacité de Hop à détecter les combinaisons oligogéniques identifiées manuellement par les cliniciens et met également en évidence sa capacité à identifier de nouvelles signatures oligogéniques dans cette maladie. dans cette maladie.

En résumé, cette recherche démontre qu'il est désormais possible de détecter directement les combinaisons de variantes pathogènes dans les données de séquençage de l'exome entier à l'aide d'approches computationnelles. En introduisant un nouveau référentiel de données, des outils informatiques et des protocoles d'analyse, cette recherche ouvre la voie à une détection et une analyse plus faciles des signatures oligogéniques des maladies génétiques.

Nederlands

Dankzij verschillende sequeneringstechnologieën is er enorme vooruitgang geboekt in het begrijpen van de relatie tussen genotypen en fenotypes. Vooruitgang in gerelateerde computationele methoden heeft geleid tot de ontwikkeling van verschillende prioritiseringsmethoden die in volledige exomen (WES) nauwkeurig identificeren welke genetische variant verantwoordelijk is voor een bepaalde ziekte. Niettegenstaande deze vooruitgang, veronderstellen deze methoden vaak dat genetische ziekten passen in een monogenetisch overervingsmodel, wat soms problematisch is. Het is mogelijk dat dezelfde ziekte ook het gevolg is van een combinatie van varianten in een klein aantal genen, wat door die methoden niet gevonden wordt. Het is dus belangrijk om nieuwe computationele methoden te ontwikkelen om deze complexere combinaties van genetische varianten, bekend als oligogenetische combinaties, te identificeren en dit in WES data. Vetrekkende van een algoritme voor het kwantificeren van oligogenetische combinaties in genen ontwikkelen we in deze thesis een nieuw algoritme voor het rangschikken van pathogene combinaties in volledige exomen. Ten eerste ontwikkelen we een nieuwe databank die informatie verzamelt over alle oligogenetische variantcombinaties die in de literatuur worden gerapporteerd om de kwaliteit van de pathogeneciteitvoorspellingen te verbeteren. Deze databank introduceert tevens een gestandaardiseerd raamwerk voor het beoordelen van de combinaties gerapporteerd in de literatuur. Met deze databank ontwikkelen we de High-throughput oligogene prioritizer (Hop). Hop combineert een pathogeniciteitsscore voor de variantencombinatie met een score voor de ziekterelevantie van het genenpaar, gebaseerd op kennis in biologische netwerken. Deze integratie maakt het mogelijk om variantencombinaties te rangschikken op basis van hoe waarschijnlijk het is dat ze de ziekte van een patiënt verklaren. Deze tool demonstreert superieure prestaties ten opzichte van bestaande benaderingen voor het rangschikken van oligogenetische combinaties in exomen. Ten slotte onderzoeken we de bruikbaarheid van onze methode in echte patiëntgegevens. We passen de Hop toe op een cohort patiënten met mannelijke onvruchtbaarheid, een ziekte met heterogene genetische oorzaken, en onderzoeken de relevantie van de geprioriteerde combinaties. Deze analyse demonstreert het vermogen van Hop om oligogene combinaties te detecteren die handmatig door artsen zijn geïdentificeerd, en toont ook zijn vermogen aan om nieuwe oligogenetische kenmerken van een ziekte te identificeren. Samenvattend kunnen we stellen dat het nu mogelijk is om met behulp van Hop ziekteverwekkende variantencombinaties direct te detecteren in volledige exomen. Door de introductie van een nieuwe gegevensopslagplaats, computationele methoden en analyseprotocollen opent dit onderzoek de weg voor eenvoudigere detectie en analyse van oligogenetische handtekeningen voor verschilende ziekten.

Acknowledgements

My PhD journey has been a challenging, enriching but also very fun adventure. All the work presented in this thesis, the questionings, the lessons learned, and the enjoyable times would not have been possible without many people surrounding me during these past four years and who I want to acknowledge here.

First, I would like to thank my PhD supervisors Tom Lenaerts and Ann Nowé for giving me the opportunity to work on this project. Special thanks to Tom for his constant optimism, helpful comments and giving me the opportunity to go travel abroad during this PhD as he knew it would give me extra motivation.

I would also like to thank Gianluca Bontempi for being on my advisory committee on the ULB side and providing relevant feedback and help on machine learning and ranking questions, as well as Pieter Libin for following my PhD progress from the VUB side and giving encouraging comments on my reports.

I thank the members of my jury for agreeing to evaluate my work. Special thanks to Matthieu Defrance and Guillaume Smits, who in addition to being on this PhD jury, also followed my research and provided constructive comments on my work over the years. Thank you to Wim Vranken, Catharina Olsen, Yves Moreau and Emmanuelle Génin for agreeing to evaluate my work and providing valuable feedback to improve this thesis.

I want to extend a huge thank you to Maris Laan and her research team at the University of Tartu for their warm welcome in their group during my three visits to Estonia. I thank especially Maris for her enthusiasm regarding my work and this collaboration, as well as her extensive knowledge of genetics which have made the last chapter of this thesis possible. I extend a special thank you to Avirup Dutta for helping me navigate the datasets as well as cleaning up the VCF files and running analyses from abroad when I needed to. A warm thank you to Kristiina, Ana Grete, Anu, Rain and Mario for showing me around Tartu and making me discover the Estonian culture.

I also want to thank my collaborators at the Erasme hospital: Chantal Depondt for her optimistic outlook on my results and Sarah Duerinckx for relevant feedback on the predictions and visualisation of the results.

I am extremely grateful to the members of the oligogenic team at IBsquare for being such a fun and collaborative team to be a part of. My first thank you goes to Sofia, who supervised me during my master thesis before showing me the ropes of PhD work and providing really helpful support in writing articles and developing new ideas. I thank Charlotte, for the collaborative work in curating articles for OLIDA, her support in the final months of thesis writing (including

careful reading of this manuscript), her inspiration to scientific communication, and the fun and gossipy office chit-chats. Thank you to Alex for being so knowledgeable and resourceful about bioinformatics, the rich scientific discussions, his inspirational attention to detail, and being a great officemate to share afternoon popcorn with. Thank you to Emma for the support on adding Hop in analyses pipelines, her helpful insights into cloud infrastructure, and being a cheerful running buddy during lunchtimes. I also thank Nassim for our collaborative work on developing VarCoPP2.0 and his relaxed attitude that made everything seem easy. Finally, I thank Inas for being a great master student to supervise, and an even better PhD student to work with, and for being a supportive officemate during this last year, especially with her encouraging and constructive comments on writing this manuscript.

I am very thankful to my colleagues from the IBsquare and MLG research groups for making this workplace such a lively and cheerful environment. The birthday celebrations in the kitchen, coffee breaks, lunches and occasional lunch runs made it motivating to come to work every day.

I also want to thank my friends from Brussels and abroad, for the support and the fun adventures. Thank you to my numerous flatmates around Brussels for the breakfasts, tuesday night beers and other fun times.

Lastly I want to thank my family for the support and encouragement through the years, as well as giving me a great excuse to run away to sunny Provence or lively Paris every time Belgium became too gloomy.

Contents

Acronyms and Abbreviations

1KGP 1000 Genomes Project. 40, 63, 69, 91, 94, 119

ACMG American College of Medical Genetics. 12, 13, 176, 177, 182–185, 193
AD Autosomal Dominant. 8, 123, 169, 177
API Application Programming Interface. 48
APS Average Precision Score. 69, 127
AR Autosomal Recessive. 8, 123, 162, 166, 169, 171, 175, 185, 199, 245
ARBOCK Association Rule learning Based on Overlapping Connections in Knowledge graphs. 78, 204
AUC Area Under the Curve. 69, 152

BOCK Biological networks and Oligogenic Combinations integrated as a Knowledge graph. 77, 106, 125, 135, 137, 220
BPO Biological Process Ontology. 54
BRF Balanced Random Forest. 127, 128, 133

CADD Combined Annotation Dependent Depletion. 64, 73, 121
CCDS Consensus Coding Sequence. 42, 163, 164
CCO Cellular Component Ontology. 54
CDF Cumulative Density Function. 69, 70, 136, 137, 139, 141, 143, 152, 153, 155
CHH Congenital Hypogonadotropic Hypogonadism. 29
CNV Copy-number variant. 72, 80, 93, 95, 228
CS Classification Score. 76

DAG Directed Acyclic Graph. 50, 54, 69, 125, 139
DCG Discounted Cumulative Gain. 71
DE Digenic Effect. 23, 72
DIDA Digenic Disease Database. 15, 32, 79, 80
DMD Dual-Molecular Diagnosis. 16, 72
DNA Deoxyribonucleic Acid. 2
DP Read Depth. 163
DS Disease-relevance Score. 113, 136, 141–143, 170, 171, 178, 181, 184, 186, 193, 195–198, 205, 222, 249–253, 256–258, 264–274, 279–281
DSD Disorder of Sexual Development. 29, 181, 196, 198, 209

ESTAND ESTonian ANDrology. 40, 43, 47, 160, 163, 189, 199, 218, 220

FAIR Findable, Accessible, Interoperable and Reusable. 48, 103
FN False Negative. 67
FP False Positive. 33, 67, 73, 130, 132
FS Final Score. 142, 143
FSH Follicle Stimulating Hormone. 43, 203

GARD Genetics and Rare Diseases information center. 48
GEMINI Genetics of Male Infertility Initiative. 44
GnomAD Genome Aggregation Database. 42, 65, 94, 102
GO Gene Ontology. 54, 125, 138, 203, 207, 208, 210, 211, 220
GQ Genotype Quality. 163
GWAS Genome Wide Association Study. 24, 40

HGMD Human Gene Mutation Database. 63, 69
HGVS Human Genome Variation Society. 8
Hop High-throughput oligogenic prioritizer. 111, 113, 142, 148, 150, 152, 154, 156, 160, 162, 218
HPG Hypothalamic-Pituitary-Gonadal. 29, 169
HPO Human Phenotype Ontology. 22, 50, 51, 54, 63, 69, 113, 116, 203, 220

IC Information Content. 125
ICD International Classification of Diseases. 48
IHH Idiopathic Hypogonadotropic Hypogonadism. 181
IM Inheritance Mode. 164, 166, 169–173, 175, 185, 187–190, 194, 200, 205, 209, 211–213, 249
ISPP Inheritance Specific Pathogenicity Predictor. 123

KG Knowledge Graph. 77, 125, 135

LH Luteinizing Hormone. 203

MAF Minor Allele Frequency. 6, 63, 94, 115, 119, 166, 178–180, 182–186, 195, 198, 249–253, 256–258, 264–274, 279–281
MAP Mean Average Precision. 70, 154, 155
MDI Mean Decrease in Impurity. 66, 133
MFO Molecular Function Ontology. 54
ML Machine Learning. 19, 20, 58, 59, 61, 63, 73, 123, 135, 228
MPO Mammalian Phenotype Ontology. 50, 63
MRR Mean Reciprocal Rank. 70, 71, 154

nDCG normalized Discounted Cumulative Gain. 71, 154, 155
NGS Next-Generation Sequencing. 47

NLP Natural Language Processing. 63

NOA Non-Obstructive Azoospermia. 26, 29, 43, 200, 202, 212

OA Obstructive Azoospermia. 26, 28

OLIDA Oligogenic Disease Database. 79, 80

OMIM Online Mendelian Inheritance in Man. 12, 48, 49, 77, 90

ORVAL Oligogenic Resource for Variant Analysis. 23, 75

OZ Oligozoospermia. 29, 43, 200–202, 211, 212

POI Primary Ovarian Insufficiency. 29, 169, 209, 211

PPI Protein-Protein interaction. 23, 53, 55, 57, 73, 75, 112, 125, 220

PR Precision-Recall. 68, 69, 128, 152, 153

PS Pathogenicity Score. 113, 135, 141–143, 170, 171, 180–182, 184, 186, 195–198, 222, 249–253, 256–258, 264–274, 279–281

RF Random Forest. 59–61, 66, 67, 127

ROC Receiver Operating Characteristic Curve. 68, 128, 152, 153

RVIS Residual Variation Intolerance Score. 121, 122

RWR Random-Walk-with-Restart. 56, 57, 63, 113, 135, 136, 140, 220

SNV Single Nucleotide Variant. 6, 25, 80

SPGF Spermatogenic failure. 29, 43, 44

SS Support Score. 76

TN True Negative. 67

TP True Positive. 67, 73

VarCoPP Variant Combination Pathogenicity Predictor. 21, 34, 72, 73, 128, 217

VCF Variant Call Format. 19, 45, 75, 113, 161, 165, 241

VUS Variant of Unknown Significance. 12, 177, 183, 187

WES Whole Exome Sequencing. 4, 5, 22, 30–33, 42–44, 111, 113, 150, 156, 160, 163, 212, 216

WGS Whole Genome Sequencing. 4, 22, 42, 226

XL X-Linked. 123, 169

Glossary

allele Version of the DNA sequence at a particular location. 5, 43

bi-locus combination A combination of variants in two genes that leads to disease. 16, 119

class imbalanced problem Classification problem where the number of instances in a specific class is much larger than the number of instances in the other class. 68

database Structured collection of data which is electronically stored on a computer system. 47

digenic disease Disease phenotype which is caused or modulated by the interaction of variants in two distinct genes. 14

dominant disease A category of disease where a single copy of the abnormal variant allele in a gene from one parent (i.e. in heterozygous state), combined with single copy of its standard allele from the other parent, is sufficient to lead to a disease phenotype. 10

Dual Molecular Diagnosis Cases where two independent monogenic diseases are present in an individual due to the segregation of monogenic variants in two different genes. 16

epistasis The mechanisms by which the combined effect of two variants is different than the expected combined effect. 14

exome The collection of exons, which are the coding parts of the genes. 3, 40

exon Sequence of DNA which codes for a protein. 3

gene panel Set of genes that are known to be associated to a disease and that are used by clinician to perform genetic testing. 49

genetic architecture The amount and type of variation that influences a trait and provides a link between genotype and phenotype. 14, 17

genome The complete set of genetic material present in an organism. 2, 40

genotype The collection of genetic variants in an individual. xx, 5

haploinsufficiency Mechanism by which a single copy of a gene is insufficient to produce the correct phenotype. 10, 123

heterogeneous network Network with different types of nodes . 51

incomplete penetrance A situation where some individuals who carry the pathogenic variant associated with a particular disease, do not show the associated disease trait. 12

indel Insertion or deletion variant in the DNA sequence. 6, 47, 121

intron Sequence of DNA which does not code for a protein but is between the start and stop codon of a gene. 3

Machine Learning The field of computer science that uses statistical techniques to enable a computer system to learn from a particular type of data and create a predictive model that can be used in the future. 19

Mendelian disease A disease that is caused by genetic changes (mutations) in a single gene or due to the abnormalities in the genome and follow the rules of Mendelian genetic inheritance. 12

modifier A modifier gene/variant is a gene/variant that will affect the expression of another gene/variant (e.g. it will alter the phenotype caused by the other gene/variant). 174, 189, 205

Monogenic plus modifier A digenic case, where one variant is present in the primary most pathological gene of the gene pair and alone can infer symptoms of the disease, whereas the variant at the second less detrimental gene can either affect the severity of the symptoms or the age of onset. 16

multiplex network Network with one type of nodes but different types of edges . 51

multiplex-heterogeneous network Network that is both heterogeneous and multiplex, i.e. which has different types of nodes and different types of edges between the same types of nodes . 51

network Collection of nodes, which represent entities, and edges, which represent connections between them. 51

node degree The number of edges connected to a node inside a network. 173

oligogenic disease Disease phenotype which is caused or modulated by the interaction of variants in two or more distinct genes. 13, 14

ontology Standardized and hierarchical vocabulary, which provide a common language for organising knowledge. 50

phenotype The collection of traits observed in an individual. 5

polygenic disease A polygenic disease is a genetic disease caused by variation in a large number of genes. 13

precision medicine Fine-tuning of medical care based on the genotype of the patient. 16

recessive disease A category of disease where more than one abnormal variant alleles (e.g. in homozygous or heterozygous compound state) in a gene are needed in order to lead to a disease phenotype. 10

Single nucleotide variant Variant which involves the change of a single nucleotide in the DNA sequence. 6

True Digenic A bi-locus model, where variants in both genes are needed in order to show symptoms of disease in an individual, whereas those carrying either one of the two variants remain unaffected. 15

variant Sequence of DNA in an individual's genome which differs from the reference genome. 5

wild-type allele The allele that encodes the most common phenotype in the population. 7

X-linked inheritance Mode of inheritance where a disease is caused by a gene on the X chromosome, affecting males and females differently due to their different sex chromosome compositions (XY for males and XX for females). 11

Y-linked inheritance Mode of inheritance where a disease is caused by a gene on the Y chromosome, therefore only affecting male individuals. 11

Chapter 1

Introduction

Over the past decades, there have been tremendous advancements in the understanding of the genetic basis of diseases, and genetic testing and variant interpretation have become routine procedures for a subset of diseases. These advances have been facilitated by huge amounts of data regarding genetic variation in relation to human diseases. Such resources allowed the development of novel computational approaches able to predict the involvement of genetic variants in human diseases and assist clinical researchers in identifying the causal genetic variants. While predicting the pathogenicity of monogenic variants is now well established, the first attempts at detecting oligogenic causes to disease are still recent and remain limited.

In this thesis, we build up on the first developed resources for digenic analyses, to introduce a new database, computational tool and analysis pipeline to enable oligogenic analysis at the whole-exome level. This transition is made possible by the decreasing cost of sequencing technologies, which have enabled the collection of large datasets of human genetic data in relation to diseases at the exome level. Additionally, the development of new computational methods, such as cloud computing, enable secure data sharing (through encryption, access control, and secure protocols), and thus provide the means to develop such computational pipelines.

This first chapter introduces the necessary concepts and knowledge to understand the work presented in this thesis. First, we introduce the basic notions of human genetics, including the organization of the human genome, the sequencing technologies that allow for deciphering of the genome, the different types of genetic variation and how these are linked to human traits. Secondly, we discuss rare diseases and the importance of studying genetic variation in relation to these diseases to better understand these disorders and improve patient care. In the third part, we dive into how bioinformatics approaches and machine learning have helped comprehend and predict the link between genetic variation and diseases. Finally, we introduce basic concepts of the genetics of male infertility, a disease further analyzed in the last chapter of the thesis.

1.1 Foundations of human genetics

1.1.1 The human genome: molecular basis, structure and organization

The study of genetics, i.e. how traits are passed from one generation to the next, goes back to the experiments performed by Gregor Mendel in the mid 19th century, who first introduced the laws of inheritance and suggested that some traits are inherited through "genes", the fundamental unit of heredity [1, 2]. It is only years later, with the experiments of Thomas Hunt Morgan in 1910, that genes were revealed to be located on molecules called chromosomes [3]. The work of Avery, MacLeod and McCarty in 1944 then demonstrated that chromosomes were constituted of Deoxyribonucleic Acid (DNA), which became known as the carrier of genetic information [4]. This molecule was first described in 1869 by Friedrich Miescher, and was shown to be made up of four building blocks, called *nucleotides* (Adenine, Thymine, Guanine and Cytosine), by Phoebus Levene [1, 5]. Finally, it is only in 1953, with the work of Watson, Crick, Wilkins and Franklin, that the full structure of the DNA double helix was discovered, demonstrating complementary base-pairing between the different nucleotides, leading to a shift towards a molecular understanding of human genetics [1, 6].

Knowing the structure of the material that carried genetic information, scientists turned to understanding how this information was encoded. The genetic code was discovered in 1966, demonstrating that specific sequences of nucleotides code for specific amino acids that make up proteins [7]. Because different *codons* or triplets (i.e. sequences of three nucleotides) can code for the same amino acid, the genetic code is said to be *degenerate* or *redundant*. In the following decades, chemists discovered how to "read" this genetic code, by developing sequencing technologies which enabled fast deciphering of the genetic code. These technologies are discussed in more details in the next section (1.1.2).

Following the development of these sequencing methods, the first human genome, i.e. the entirety of the genetic information that makes us who we are, was assembled in 2002 [8]. This revealed that the genome consists of two distinct parts: a nuclear and a mitochondrial component, which are located in two different organelles of the cell. The nuclear component of the genome, on which we focus in this thesis, is made up of approximately 3.2 billion base pairs of DNA, which are organized in 23 pairs of chromosomes. 22 are common between males and females. The last pair typically consists of two X chromosomes in female individuals, and one X chromosome and one Y chromosome in males , although other configurations can exist [8, 9].

Genes are the functional units of the genome and consist of a sequence of nucleotides coding for a functional product (a protein or RNA molecule), which in turn allows for the proper function of biological mechanisms in the cells. A gene typically begins by a *start codon*, which indicates to the cellular machinery where the coding sequence starts, and ends with a *stop codon*, signalling the end of the coding part.

The mechanism by which the genetic code is translated into a protein product is known as the "central dogma of molecular biology" [10]. A gene is first transcribed into a messenger RNA (mRNA), by the transcription machinery, which works directly in the nucleus to transfer the information contained on the DNA double helix onto a single strand RNA molecule. The mRNA then undergoes transformations, such as *splicing*, which removes certain parts of the gene which do not code for the protein. It is then transferred outside the nucleus to be transformed into a protein product, a process known as *translation*, during which each sequence of codon is translated into an amino acid, which are the building blocks of proteins.

The parts of the gene that actually code for a protein sequence are called *exons*, but there are also some non-coding parts between the start and stop codons, which are referred to as *introns* and contain regulatory elements, which modulate the expression of the genes [11]. Furthermore, a gene is usually preceded by a promoter region and can be surrounded by several enhancers and other regulatory sequence elements.

The human genome consists of about 20,000 protein coding genes [12], but also includes other genes coding for RNA molecules that interact with proteins and regulate different processes taking place in the cells [13]. The part of the genome that involves only protein-coding sequences is known as the *exome*, and is the main focus of genetic studies and the work presented in this thesis, although it only represents 2% of the full sequence of the genome [14]. It is nonetheless important to note that new evidence, brought forward by advancements in sequencing technologies, is showing that the non-coding parts of the genome also play an important role in the expression of our traits [15–17].

1.1.2 Sequencing technologies and mapping the human genome

Discoveries in genetics have been greatly facilitated by the development of different methods to sequence the human genome, i.e. determine the order of the nucleic acid residues. These sequencing technologies have been traditionally divided in three "generations", and differ by the number of DNA molecules they are able to process, the size of the DNA fragments they are reading (which are known as *reads*), and the types of genetic variant they are able to detect [18, 19].

The first generation of sequencing technologies dates back to 1977, with the development of *Sanger sequencing* [20]. This method uses special specific chain-terminating nucleotides (ddNTPs) in order to generate fragments of the DNA molecule of different lengths, and then reads the order of nucleotides by looking at the last nucleotide of each fragment (the ddNTP which is tagged to be easily readable). Sanger sequencing produces highly accurate results (99.999% of the read bases are the true bases) and long reads, and remains today the gold-standard for obtaining accurate information about genetic variation. However, it can only sequence one fragment of DNA at a time, and is costly due to the number of chemical reactions required and the cost of ddNTPs [21].

The second generation of sequencing technologies, appeared in the early 2000s, also consists of synthesizing new DNA molecules and reading what nucleotide is added at each step. However, these new methods can sequence huge amounts of DNA at a time because they allow for massive parallelization of the sequencing reaction. Briefly, several DNA molecules are anchored to a specific location and undergo a "wash and scan" cycle: the cells are flooded with reagents containing the labeled nucleotides, that are incorporated in the DNA strands during the synthesis reaction, which is then interrupted in order to scan the cells to identify the added bases, before treating these bases to prepare for the next cycle. The nature of the labeled nucleotides and the scanning method depends on the sequencing technology [22]. These technologies are high-throughput, meaning that they can sequence a large amount of DNA, and have seen their costs dramatically reduce over the years (from 3 billion dollars for sequencing the first human genome to less than 1000 dollars for a complete genome sequence today), meaning that they can be easily applied to larger groups of people [23, 24].

The third generation of sequencing technologies refers to long-read single-molecule sequencing technologies and was initiated around 2010 [25]. The idea was to obtain the same throughput as second generation sequencers, but using a single DNA molecule instead of amplified fragments, and to obtain longer sequencing reads (typically 10-50kb). The former is important to avoid errors occurring during the amplification while the latter would allow reduction in errors happening during the mapping of the reads. Third-generation sequencing has shown to improve the quality of genome assemblies and to allow a better detection of *structural variants* (see Section 1.1.3) [26]. Furthermore, these novel sequencing technologies allow for an easier sequencing of RNA molecules and detection of epigenetic modifications (i.e. modifications that do not alter the DNA sequence but have an effect on the expression or silencing of genes) [27], which are not discussed further in this thesis.

Sequencing approaches can also be distinguished based on the type of variants they detect, as well as the regions of the genome they read. *Panel-based sequencing* refers to the sequencing of specific gene sets, and therefore allows to detect variants in this gene list specifically, as well as in regulatory regions. On the other hand, *Whole Exome Sequencing (WES)* or *Whole Genome Sequencing (WGS)* are usually regarded as less biased, because they allow to detect variants in the full exome or genome of the individual respectively, and can therefore allow for novel gene - disease associations. Although sequencing costs have dramatically reduced, WGS remains about 3 times more expensive than WES [25, 28].

The choice of sequencing method therefore depends on the type of genetic variants and analyses a researcher is looking to perform, as well as more practical considerations such as costs [29]. Although some people advocate for the use of WGS, since it allows to detect variants in a larger portion of the genome and is usually also more accurate, WES is currently more widespread, due to its lower cost and the fact that the majority of relevant genetic variation for genetic diseases has been shown to be located in the coding part of the genome.

Panel based sequencing is still performed in cases where the genes of interest are well known, but is now more often replaced by WES with *in silico* panel testing, i.e. performing WES but then restricting the downstream analysis to a specific gene set. This method allows to reduce the potential false positives obtained from WES, but leaves the door open for analysis of other regions of the exome if needed.

The new generations of sequencing technologies have led to increased accuracy in the sequencing of the genome and the release of more and more precise reference genome [30,31], culminating in the publication of the fully complete human genome, which now also covers the previously unsequenceable and unalignable regions [32]. A first human pan-genome, i.e. a representation of the human genome that takes into account genetic variation between individuals from different populations, was also recently published [33].

1.1.3 Overview of genetic variation

Although the large majority of the genome is common between all human beings, about 0.1% differs between individuals [34]. This is known as genetic variation, and is part of what makes each human unique.

Understanding the relationship between *genotype* — the set of genetic variants carried by an individual — and *phenotype* — the physical traits exhibited by the individual — is one of the main challenges in biology [35]. In the past decades, high-throughput sequencing has revolutionized the field of human genetics with large-scale data collection such as the 1000 genomes project [36, 37] or the Exome Sequencing Project [38] providing huge amount of information on human genetic variation.

Genetic variation comes from differences in DNA sequences, and is continuously generated by the mutational process [39]. Mutations can be either *inherited* from the parents or *acquired*, which are referred to as *de novo* mutations, and happen by chance due to errors in the DNA repair process or mutational events such as exposition to mutagens or radiation. *Germline* variants refer to variants that are present in the reproductive cells and can be passed on to other generations, while *somatic* variants are acquired over the course of an individual lifetime in other cell types [40]. Genetic variation persists in the genome, enabling evolutionary change and diversity in populations [39].

Genetic variation can take different forms, and variants can be classified with respect to a variety of factors. A genetic variant is defined as any specific region of an individual genome where the DNA sequence of the individual differs from the reference genome, encompassing both the location and the specific change in the sequence. When multiple versions of a DNA sequence exist at a particular site within a population, these different versions are called alleles, with each allele representing a distinct sequence at that locus.

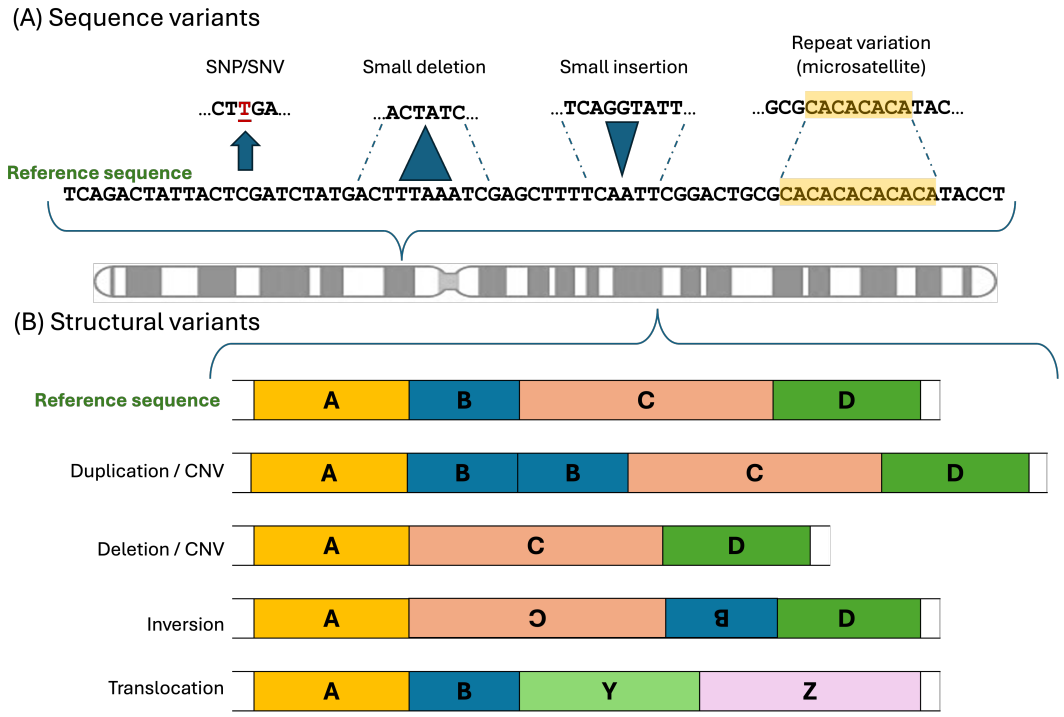


Figure 1.1: Types of genetic variation found in human genomes: (A) sequence variants which involve only a small number of nucleotides and (B) structural variants which involve larger sequences of DNA, where each letters represent different large segment of DNA. Figure adapted from [40]

A first way to classify genetic variants is according to their frequency in the population. This is assessed by computing the Minor Allele Frequency (MAF) of a variant, i.e. the frequency of the less common allele of a variant. Variants with a MAF larger than 1 % are considered common and sometimes called *polymorphisms*, while variants with MAFs lower than 1 % are considered rare [39, 40].

Another way to distinguish between different types of variants is according to the size of the change in DNA sequence they imply (See Figure 1.1).

Sequence variants (Figure 1.1A) refer to variants that span a small number of DNA bases (usually less than 1000 base pairs [40]). Single nucleotide variants (SNVs) are sequence variants that are characterized by the change of a single nucleotide in the DNA sequence: the substitution of a purine by another purine is called a *transition* while the substitution of a purine by a pyrimidine or vice versa is a *transversion* [34]. SNVs are the main source of genetic variation, with the latest phase of the 1000 genomes project reporting information on 84.7 million of them [37]. A typical genome contains 3.5 to 4 million of SNVs [37] with most of them being common: only 1 to 4 % of variants in a typical genome have a MAF smaller than 0.5 % [37]. This means that the majority of the variation present in an individual genome is shared among several individuals in the population. Insertion-deletions (or *indels*) refer to

mutations where a few nucleotides are added/removed at a specific position. Indels which are not multiple of 3 result in a frameshift mutation, which can completely change the sequence of amino acid in the protein product or result in a premature stop codon. Finally microsatellite variants refer to indels in particular repetitive regions of the genetic code.

The remaining variants belong to the global category of structural variants, which imply larger changes in the DNA sequence [34] (Figure 1.1B). They include copy-number variants (where a large part of the DNA sequence is duplicated or deleted), block substitutions (which correspond to the substitution of several adjacent nucleotides), inversions (where the order of nucleotides is reversed in part of the sequence) and translocations (where pieces of a chromosome break and are reattached at different sites than their original ones). Structural variants affect more than 1000 DNA bases, but a typical genome only contains 2,100 to 2,500 of them [37], which represents less than 1 % of total variation in a genome. Structural variants identification and characterization methods have evolved quickly in the past decade but were previously less studied, which is partly due to difficulties in detecting them [41]. New information is therefore likely to allow for a better understanding of the mechanisms and consequences of this type of variation in the near future [42].

Variants also differ by their effect on the individual's phenotype, which is highly dependent on the location of the variant and its change in the DNA sequence. Most variants are located in the non-coding part of the genome and were therefore initially thought to have no direct consequence. However, recent studies report that variants in intronic regions close to the exon edges, for example, can have strong repercussions on the phenotype, since they contain critical splicing elements [34, 43]. *Non-synonymous* variants are mutations in the coding sequence with a direct effect on the protein sequence: if the mutation leads to a change in amino acid it is called a *missense* mutation, while a SNV causing the premature appearance of a stop codon is termed *nonsense*. On the contrary, *synonymous* mutations refer to variants in the coding region which do not lead to changes in the amino acid sequence of the protein product, due to the redundancy of the genetic code.

Non-synonymous mutations are therefore assumed to have a higher potential to be deleterious since they have a direct effect on the protein product of genes, but synonymous mutations have also been associated with diseases [44].

Finally, mutations have different consequences on the phenotype based on their *zygosity*. A given mutation can be *homozygous* if it is present on both chromosomes of the same pair, *heterozygous* if it is only present on one of the two chromosomes (and the other chromosome therefore presents a wild-type allele copy or allele), or *compound-heterozygous* if two distinct mutations occur in the same gene (which means both copies of the gene are mutated but with a different mutation). Finally, a variant can be *hemizygous* if is located on the X or Y

chromosome in a male individual, since the individual only carries one copy of each of the two distinct sex chromosomes. The effect of the zygosity of the mutation can change depending on the inheritance mode of the genes: *Autosomal Dominant* genes are defined as being more tolerant to heterozygous mutations than *Autosomal Recessive* genes (see Section 1.2.2).

Due to their versatile nature, genetic variants can therefore be represented using different coordinates. The more precise description is the genomic coordinates, which gives the chromosome and number of the changed nucleotide, as well as the reference and alternative (i.e. the mutated residue) nucleotides. Another possibility to represent a genetic mutation at the DNA level is to give the coding DNA change. This change is given as the gene name, or transcript ID of the exon, as well as the nucleotide number where the variant is located. Since genes can have several transcripts, this notation is less precise since if the transcript is not provided, it is not always possible to retrieve the actual variant. Finally, variants can also be described at the protein level, by indicating which amino acid of the protein is changed in which amino acid. However, due to the redundancy of the genetic code, it is not always possible to trace back which genetic variant correspond to which protein change.

For example, a change of a C to T nucleotide at a genomic position 68208397 of chromosome 15 can be represented as 15:68208397:C:T using genomic coordinates, CLN6(NM_017882.3):c.679G>A in cDNA coordinates (where CLN6 is the gene in which the variant is found and NM_017882.3 is the transcript used to compute the coding DNA coordinates) and p.Glu227Lys in protein change. This diversity of variant representations has been a challenge in genetic research, since different notations used in different research articles may actually refer to the same variants. In order to address this issue, guidelines for the notations of variants have been put in place by the Human Genome Variation Society (HGVS) Nomenclature [45], and different tools have been developed to convert variant coordinates between the different formats used [46–48].

1.2 Studying the genetic basis of rare disorders

Genetic variation is linked to a wide variety of traits, which is called phenotypic variation. While the majority of genetic variation is linked to benign phenotypes, meaning they result in traits that do not affect health or biological function (e.g., differences in eye color or height), a small fraction of genetic variants can be deleterious and cause diseases. In this section, we delve into how human diseases are studied in relation to genetic variation, in order to obtain a more precise understanding of these diseases and improve the quality of life of people affected by them.

1.2.1 An introduction to rare disease research

There are currently about 7,000 rare diseases reported around the world [49], although it is expected there are many more yet to be identified and classified [50]. While these diseases individually have a small prevalence (a disease is defined as rare if it affects less than 1 in 2,000 individuals in Europe [51]), they together affect about 3.5–5.9% of the world's population which means that 263–446 million persons are currently living with a rare disease [52]. This makes these diseases an important public health issue [53].

The main challenges faced by affected individuals on one hand include the time to achieve diagnosis, uncertain diagnosis and inappropriate treatment for the disease. On the other hand, the main challenge faced by clinicians and researchers when studying rare diseases is the limited sample sizes [54, 55]. With few individuals affected by a particular condition, recruiting an adequate number of participants for clinical studies becomes complicated. This small number of participants can limit the study's statistical power and robustness, and also makes it challenging to generalize the findings, as the small group of participants may not capture the full spectrum of variability in disease presentation, limiting the applicability of the results to the broader population of patients. The general lack of data on rare diseases hinders the current efforts to understand their etiology, progression and design optimal strategies for treatment. Furthermore, rare diseases are characterized by being clinically heterogeneous, meaning that patients with the same condition can exhibit a wide spectrum of symptoms (varying in type, severity and onset). This variability complicates both diagnosis and treatment, as the disease may present with overlapping symptoms across different conditions or follow unpredictable clinical courses. As a result, developing standardized diagnostic criteria and effective treatment protocols becomes significantly more challenging for healthcare providers [56].

To overcome these challenges, the main approach used by researchers is to develop large collaborations and networks bringing together researchers and patients from around the world to study rare diseases [57,58]. This is the case for example of the Undiagnosed Disease Network (UDN), which collects data on more than 5000 patients affected with rare diseases, for whom a diagnosis could initially not be obtained [59]. Through data sharing and their re-analysis, new disease mechanisms were discovered [60]. These initiatives also exist for specific, more common, diseases, such as the Epi25 project collecting data on patients affected with epilepsy around the world [61], the ASPIRE autism spectrum disorder cohort [62] or the Deciphering Developmental Disorder cohort [63].

Additionally, the development of standardized registries and classification of diseases allow to pool together data on diseases with similar phenotypic presentation [64,65]. Orphanet, for example, has enabled the precise recording of rare diseases, together with a classification of the diseases. It also provides links to other databases of diseases and is therefore a key resource in rare disease research, enabling researcher to properly identify gaps in understanding, and prioritize areas for further investigation [64].

Finally, the more recently developed Human Phenotype Ontology, enables researchers to use a unified vocabulary when describing the symptoms of patients [66, 67]. This resource has been shown to enable new discoveries by grouping patients with similar symptoms thanks to this framework [68–70] and is continuously evolving, now providing a huge structured vocabulary to describe disease symptoms in detail [71]. Precise description of the symptoms of the patients through what is called deep phenotyping [72], was shown to be essential to obtain a better understanding of the genetic mechanisms underlying certain diseases [73].

Through the use of such registries and standardized vocabularies, huge amounts of data have been collected and shared among researchers on not only the clinical manifestations of these heterogeneous disorders, but also genetic associations linked to these diseases, paving the way towards obtaining a molecular diagnosis for genetic disorders.

1.2.2 Genetic variation and rare diseases

It is estimated that around 80% of rare diseases are caused by genetic mutations [52], and that genetic factors also contribute to many other disorders which are not considered as rare such as cancer or type 2 diabetes [74]. Genetic diseases can be inherited if the mutation is passed on from the parents to the affected individual or occur *de novo*, if the causative mutation appeared accidentally in the individual [75]. A genetic mutation usually causes a disorder phenotype by impairing a molecular pathway, i.e. the mutation alters a protein product or a regulatory region of a gene, leading to an impaired protein function or gene expression, and thus have implications on the cellular or biological processes possibly leading to a disease phenotype.

Based on the number of copies of the mutation that are needed to show the disease phenotype, diseases are often divided into *dominant diseases*, i.e. diseases for which a single allele (or copy of the mutation) is sufficient to cause the disease, and *recessive diseases*, i.e. diseases that require both copies of the gene to be mutated. These modes of inheritance also apply to single genes, where dominant genes have been shown to lead to impaired cellular function with only one mutation, while recessive genes require mutations in each copy of the gene to lead to a disease phenotype. This can happen through different mechanisms such as *haploinsufficiency* (when one copy of a gene is insufficient to produce the correct phenotype), *gain-of-function* (when a mutation leads to a protein with a new function, as opposed to a *loss-of-function* mutation which leads to a non-functional protein) or *dominant negative* (when the

expression of a mutant protein interferes with the activity of a wild-type protein) behaviours. Some genes are therefore known to be exclusively dominant (i.e. a single heterozygous mutation in the gene will have no direct effect on the protein product), exclusively recessive (i.e. any mutation in the gene will have a direct effect on the protein product), or a mix of both depending on the condition or on the location of the genetic variant [76]. For diseases which are caused by mutations in genes located on specific regions of the X-chromosome, the *X-linked inheritance* term is used for male individuals. Indeed, males only carry one copy of the X chromosome and a mutation in a gene located on this chromosome can therefore act as a dominant mutation since it is the only copy of the gene that is carried.

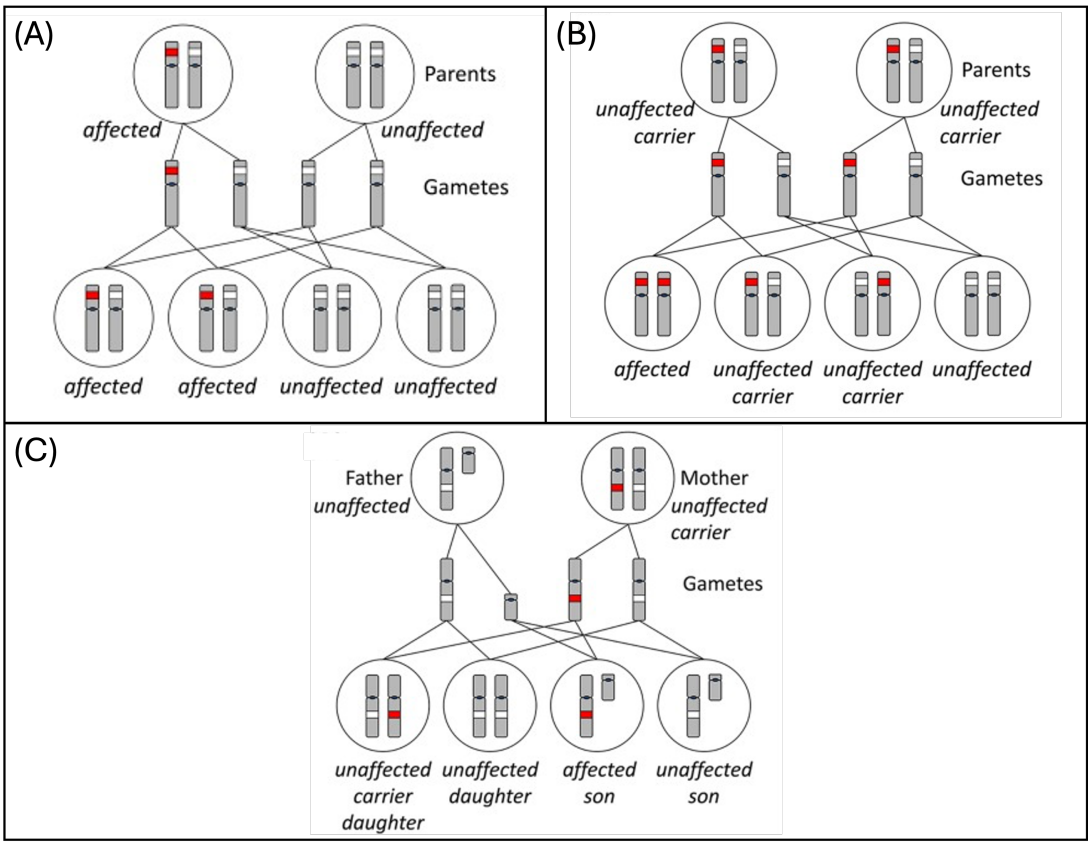


Figure 1.2: Mechanisms of dominant inheritance (A), Recessive inheritance (B) and X-Linked inheritance (C). Figure adapted from [40]

Finally, conditions associated with *Y-linked inheritance* also exist although they are rare since the Y chromosome is very small and only contains comparatively few genes [40]. With the exception of variants in genes which have homologues on the X-chromosome, variants on the Y-chromosomes will be in hemizygous state, and their phenotype will be manifest [40]. Consequently, affected males also have affected fathers, unless a *de novo* mutation has occurred,

and all their sons will be affected. An example of Y-linked condition is spermatogenic failure, which is further discussed in Section 1.4. Other modes of inheritance, such as mitochondrial inheritance can also be associated with some diseases, but are not further discussed in this work.

Genetic variants linked to diseases have been shown to have different effects in different individuals. The *penetrance* of a mutation defines the proportion of individuals carrying that mutation who are affected by the associated phenotype. A mutation is therefore known to have high penetrance if the majority of the individuals carrying this mutation are affected by the diseases. For example, some alleles present in the BRCA1 gene, which have been associated to breast and ovarian cancer in women, have been shown to confer an $\sim$ 80% risk to develop the cancer, therefore exhibiting high yet *incomplete penetrance* (less than 100%) [77]. Even when a mutation is 100% penetrant, i.e. all individual carrying the mutation present the phenotype, this associated phenotype can be of different intensity. This is the concept of *expressivity*, which describes the range of traits that can be associated with a single mutation.

To help clinicians interpret the effect of genetic variants in relation to diseases or phenotypes, guidelines have been developed to define criteria that need to be taken into account. The American College of Medical Genetics (ACMG) guidelines for asssessing the pathogenicity of genetic variants are the most widely used [78]. They lead to the classification of variants in one of 5 classes: Pathogenic (P), Likely Pathogenic (LP), Variant of Unknown Significance (VUS), Likely Benign (LB) and Benign (B) [78]. The class a variant is assigned to depends on different criteria such as frequency in large population databases, functional analysis of the effect of the variants, segregation in pedigree studies and so on. Each type of evidence is assigned a separate criteria and these criteria can be combined to assign the variant to one of the classes. VUS represent variants that could not be categorized as either pathogenic or benign, requires further analysis and should not be used for clinical decision-making [78]. It is important to note that these criteria were defined to assess the involvement of a single variant in a particular phenotype and therefore do not take into account the interaction between variants which are discussed in the next Section (1.2.3).

In order to have a better understanding of the relationship between genotype and phenotype, many resources have emerged to collect data on the genes and variants linked to specific diseases or phenotypes. The Online Mendelian Inheritance in Man (OMIM) database, for example, collects data on all Mendelian diseases and their associated genes and variants, comprising as of today data on more than 8000 distinct phenotypes [79]. The Genomics England PanelApp collects different gene sets of interest that have been linked with specific diseases, in order to aid researchers in their quest for identifying which genetic variants can cause a specific disease [80, 81]. At the variant level, ClinVar compiles reports on genetic variants and their link to diseases or phenotypes [82, 83]. Reports on any variant can be com-

plemented with detailed information supporting the classification of the variant as pathogenic, benign or of unknown significance, and can come from different research groups, in which case the information is aggregated to support for re-evaluation of the data. Each variant is assigned a review score based on the evidence associated and the number of reports on that particular variant. As of May 2024, ClinVar contains data on almost 3 million genetic variants, reported by over 3000 submitters. More recently, the Franklin database[1] has also emerged as an important tool for studying genetic variants in relation to disease phenotype, as it allows to automatically classify variants according to the ACMG criteria for monogenic pathogenicity assessment, as well as providing links to ClinVar reports and other publications discussing the genes andor variants of interest.

1.2.3 The genetic diseases continuum

The increased amount of information on human genetic variation acquired through sequencing technologies has recently completely transformed our understanding of genetic diseases [84].

The traditional framework of genetic diseases makes a distinction between *monogenic diseases* — which are caused by one rare mutation in a single gene with a large effect — and *complex* or *polygenic diseases* — which are caused by multiple mutations in multiple loci. The mutations responsible for complex diseases are assumed to be more common and to have smaller individual consequences. Examples of monogenic diseases include Sickle cell anemia, Cystic Fibrosis and Huntington disease — which are in fact Mendelian diseases, a particular sub-type of monogenic diseases where the mutation is inherited [85]. Examples of complex traits diseases include Alzheimer's disease, multiple sclerosis and epilepsy [86].

Nevertheless, new information has shown that many previously thought monogenic diseases are in fact modulated by a number of mutations in other genes, which are called *modifiers* [87, 88]. This is the case for cystic fibrosis for example, where although mutations in the CFTR gene are almost always causative of the disease phenotype [89], the disorder can be made more or less severe by mutations at different loci [90] and present with very different phenotypes [91].

This brings up the concept of *oligogenic diseases*, which encompasses disorders that are neither monogenic nor complex but rather caused, or modulated, by mutations in a small number of genes [92]. In fact, genetic diseases are now better understood as a continuum with oligogenic disorders bridging the gap between monogenic and complex diseases [93], as illustrated in Figure 1.3.

1. https://franklin.genoox.com-FranklinbyGenoox

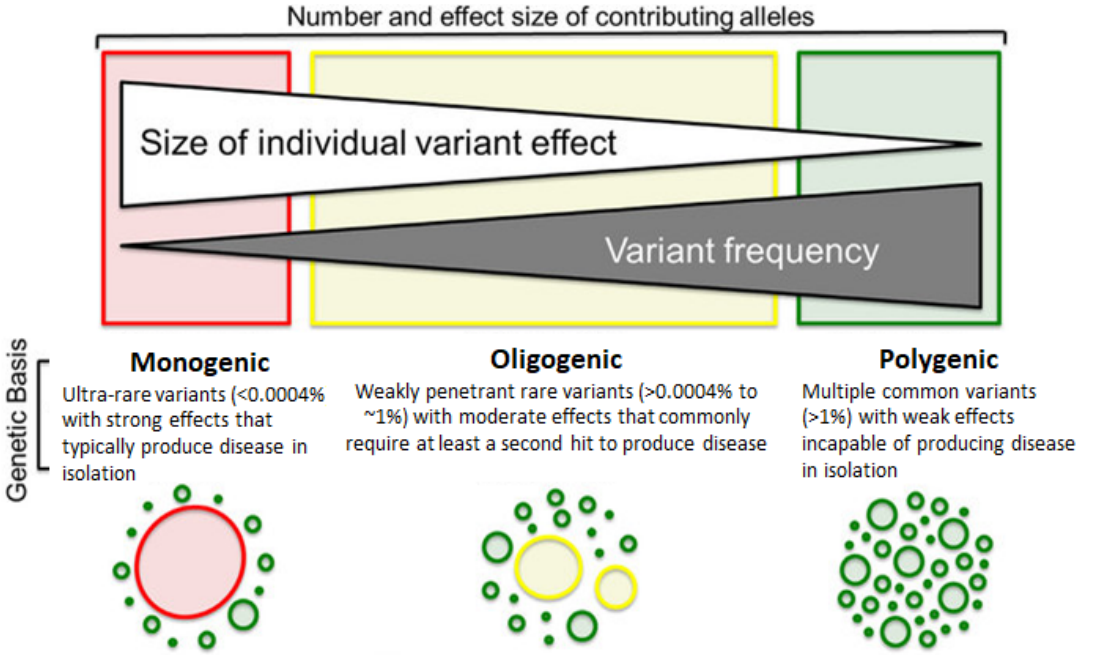


Figure 1.3: The monogenic to complex diseases continuum. Figure adapted from "The genetic architecture of long QT syndrome: A critical reappraisal", Giudicessi *et. al.* , Trends Cardiovasc Med. (2018) [94]

The emergence of the concept of oligogenic diseases brings up new challenges, since the *genetic architecture* of many of the previously established monogenic diseases needs to be re-assessed. However, it is also a great opportunity since it paves the way to get a better understanding of the interaction between genes at a smaller level, which could potentially lead to new insights into the study of polygenic diseases [84, 95].

The simplest case of oligogenic inheritance is digenic inheritance, a term first introduced by Schäffer as "phenotypes whose pattern of inheritance can be better explained by mutations in two loci than a variant in one locus alone" [96]. There have been several well established examples of digenic diseases which include Bardet-Biedl Syndrome [97–99] or long-QT syndrome [94, 100].

Different mechanisms have been described to explain how variants at different genetic locations can act together to cause the disease, and thus explain the molecular mechanisms underlying oligogenic diseases. These mechanisms are referred to under the general term of *epistasis*, which was first defined to describe the case when the effect of two variants together deviates from the expected outcome of combining these mutations [101]. There are different types of epistatic interactions, including compensatory effects (when the effect of one variant cancels the effect of the second variant) or *synergistic effects* (when the effect of the combined variants is stronger than the sum of the individual effects alone) [102, 103].

Furthermore, different epistatic mechanisms have been described, depending on whether the effect is due to the direct interaction between protein products, an additive effect due to the proteins being in the same pathway, or due to their redundant involvement in a specific biological function [84, 92].

A better understanding of digenic diseases is crucial in order to be able to tackle more complex oligogenic diseases. In recent years, a lot of progress has been made in this line of research, partly thanks to the development of the Digenic Disease Database (DIDA), which provided for the first time public access to a collection of annotated data on variants, genes and gene pairs that were identified to be causative of digenic diseases [104].

Based on the data collected in DIDA, we can distinguish three types of variant combinations, depending on the effect they have, as illustrated in Figure 1.4.

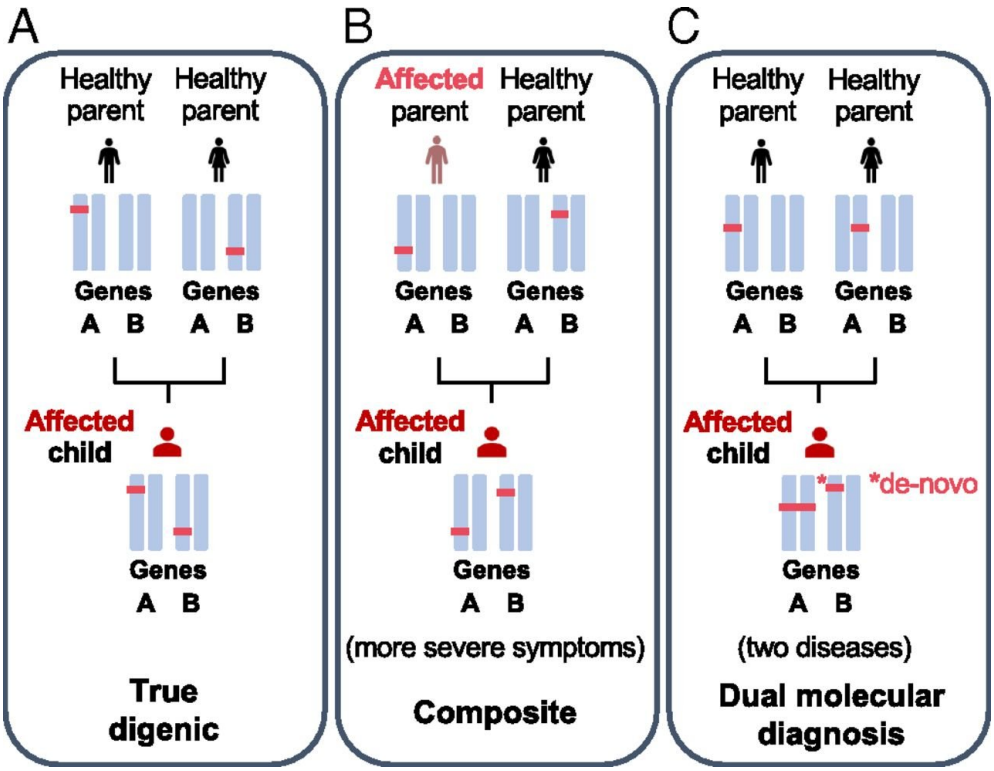


Figure 1.4: Three types of digenic inheritance observed in DIDA. Figure taken from "Predicting disease-causing variant combinations", Papadimitriou *et. al.* (2019) [105].

The first class, which is called *True Digenic*, refers to pathogenic combinations where the presence of two or more variants in two distinct loci is necessary for the patient to develop the disease phenotype. The second class, referred to as *"Composite"*, includes combinations in which a variant on the major gene causes the disease phenotype, while the variant on the

second gene modulates this phenotype by either leading to more severe symptoms or an earlier onset of the disease. This type of combination is sometimes referred to as *"Monogenic plus modifier"*. Finally, in the case of *Dual Molecular Diagnosis (DMD* combinations, variants in both loci are responsible for either distinct or overlapping disordered phenotypes.

Although DIDA was an important step forward in the understanding of oligogenic disease mechanisms, the database included variant combinations based on a set of criteria that was not clearly defined, and has not been updated since 2017. In the literature, variant combinations involving two genes are referred to as both digenic or bi-locus combinations, and both terms are used interchangeably throughout this thesis.

1.2.4 The importance of genetic diagnosis for rare disorders

The large majority (80%) of rare diseases have a genetic origin, but the exact genetic causes remain unknown for a large number of diseases [106]. The experience that patients and their families go through while waiting for a precise diagnosis is termed "diagnostic odyssey" and can sometimes last up to 30 years (from the beginning of symptoms to the correct diagnosis) [107]. It is estimated that about 50% of patients remain undiagnosed. This lack of diagnosis can mean missed opportunities for tailored treatment, and is linked with a substantial burden of uncertainty for families as well as high economic costs due to unnecessary diagnostic procedures [108].

Traditional approaches to diagnosis include detailed clinical evaluation of the patient's symptoms and laboratory tests, but the heterogeneity of rare diseases often makes these methods insufficient [109]. Novel sequencing technologies, such as WES and WGS have therefore proved useful in obtaining genetic diagnosis for cases with atypical or variable phenotypic presentations [110–113]. The diagnosis of previously undiagnosed patients can lead to change in therapeutic strategy, which can in turn lead to improvement of the patients' well-being [114]. These novel methods have led to the emergence of the concept of *precision medicine*, which aims to achieve the personalization of medical care based on an individual's genetic characteristics. Precision medicine, also called "P4" medicine, i.e. personalized, predictive, preventive and participatory medicine, is to be achieved through a better, more precise, stratification of patients, specific tests and the identification of new biomarkers, and will lead to the development of novel treatments through novel drug developments or drug repurposing [115–117].

However, using sequencing for genetic diagnosis also comes with significant challenges [118–120]. Properly assessing the causality of a sequence variant in a particular disease remains difficult, since standards were only recently published [121]. The question of how and what genetic results should be shared with patients also needs to be ethically addressed, especially for the development of preventive care based on interpretation of genetic variants [122]. Furthermore, it is important to note that while large amount of data on genetic variation has

been collected for specific European/Caucasian individuals, the scarcity of data from other ethnicities implies that genetic testing for these individuals will be less reliable [119]. Finally, generalizing these analyses to large population data requires the development of efficient and accurate methods to identify causative variants in sequencing data, which are discussed in the following section.

1.3 Bioinformatics methods for variant interpretation

New sequencing technologies have allowed for a tremendous amount of data on human genetic variation to accumulate in the scientific literature. The increased amount of information on genetic variation calls for proper computational methods in order to analyze and interpret this variation in relation to disease. In the previous years, a wide variety of variant prioritization tools have therefore emerged in the field of bioinformatics [123–125]. These tools have two main purposes: the first one is the discovery of new variant-disease associations and a better understanding of the genetic architecture of some rare genetic diseases [126]; the second one is linked to the emergence of precision medicine, which requires proper integration of "omics" information in disease risk models [127]. In this section, we introduce the different types of bioinformatics methods that have been developed to interpret genetic variation in relation to genetic diseases. We first introduce bioinformatics pipelines which transform raw sequencing data into interpretable variant data, then describe general usage of machine learning in bioinformatics and then focus on the differences between two main types of variant pathogenicity prediction methods: classification tools and prioritization methods. We then dive more specifically into methods that have been developed to identify digenic causes to disease, and finally introduce strategies and frameworks developed to find genetic causes in cohort data, in order to establish what are called "genetic signatures" of diseases.

1.3.1 From sequencing data to variant calls

The first step to any genetic analysis is to transform raw sequencing data into what are called *variant calls*, i.e. to identify which are the specific genetic variants carried by an individual. This process involves several steps, made possible with bioinformatics methods (see Figure 1.5).

The first step of this procedure is to align sequencing reads to a reference genome [129, 130]. Nowadays, three different reference genomes are commonly used in the literature: the GRCh37 assembly, the GRCh38 assembly and the chm13 assembly [131]. The GRCh37 assembly is starting to be outdated, with many resources not maintaining their data in this

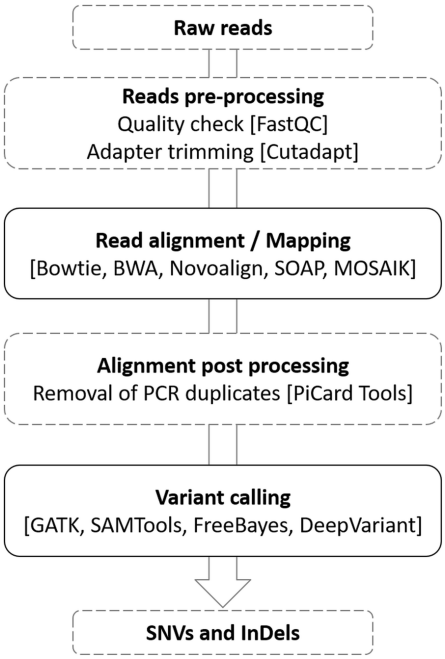


Figure 1.5: Overview of the analysis pipeline from sequencing data to variant calls, highlighting at each step, some of the main algorithms used. Figure from [128].

assembly anymore, while the chm13 is still very novel, which means that many interpretation and analysis tools have not yet been adapted to this assembly. Although tools have been developed to convert variants between assemblies [131], the reference genome used to call the variants has been shown to lead to discrepancies [132, 133].

The main algorithms used to align sequencing reads to a reference assembly include BWA [134], Bowtie [135] or HISAT2 [136]. These algorithms take as input reads generated from sequencing machines, in the format of a FASTA or FASTQ file, and look for the best-fitting positions for these reads on the reference genome by scoring the alignment of the reads to the genome and using this score to select the optimized location [129, 137]. The algorithms output SAM (Sequence Alignment file) or BAM (Binary Alignment file) files, which contain the sequencing reads, the location of their alignment on the genome and the alignment scores. Several errors can arise during alignment, including misalignment due to repetitive regions or structural variants [138, 139], as well as biases introduced by the reference genome [140–142].

From the aligned reads, several tools can be used to call variants and obtain the genotypes of the mutated alleles, after applying a few pre-processing steps to remove duplicate reads, recalibrate base quality scores, and perform local realignment around indels [29]. When analyzing several samples, variants can be called individually, for each sample, or jointly. Joint variant calling has been shown to offer important advantages as it improves the accuracy and reliability of genetic variant detection by analyzing data from multiple samples simultaneously,

reducing errors and enhancing variant quality. It also produces calls for every positions, which allows to differentiate between genomic positions for which a sample is known to have the reference allele with high confidence and positions for which the sample did not achieve sufficient coverage. Additionally, it reduces the problem of variant representation difference and enables the variant caller algorithm to use the genotype information from one sample to infer the genotype information from another [29]. The main tools used to call variants include BCFTools [143], FreeBayes [144], GATK Haplotypecaller [145] and Platypus [146]. SNVs are identified by examining aligned reads at each genomic position to find positions where the majority of reads differ from the reference genome. Indels are found by looking for regions where the aligned reads exhibit insertions or deletions relative to the reference genome. These variant calls are assigned quality scores based on several criteria, to reflect the confidence in the variant call [29].

Finally, the pipeline results in a Variant Call Format (VCF) file, which contains on separate lines different types of information on the genetic variants carried by an individual. VCF files come in two formats: cohort VCF files, which contain information on all variants carried by individuals in a cohort (and are generated by joint calling of the variants) and individual VCF files, which contain information on the variants carried by a particular person. In addition to the information on the variants carried by individuals, VCF files also contain various quality metrics which reflect the confidence in the read alignments and variant calling processes, therefore enabling for careful filtering of potentially false positive variant calls (see Section 3.1.3 for more details).

1.3.2 Machine learning models in bioinformatics

Machine Learning (ML) is the process of constructing models by applying sophisticated statistical and computational methods to extract complex patterns from data [147]. The main advantage of this type of models, with respect to traditional statistical methods, is that they are sometimes able to discover unexpected associations between input data and predictions. However, in order to have a good performance, these models usually require quite a large amount of training data [148].

Increased amount of biological data and recent advances in computational modelling have led to the application of machine learning models to several biological problems. These applications cover various biological domains, with models built to solve the gene finding problem (i.e. finding protein coding regions in human DNA), predict protein structure or understand the link between genes and disease [147, 149, 150]. There exist different categories of ML, which are described in more details in Section 3.4: supervised learning – where the learning happens from a pre-defined dataset for which the output to predict is known, unsupervised learning – where the model tries to learn the underlying structure of the data – and reinforcement learning – where the learning occurs through interactions with the environment [151]. In the

supervised classification setting, which is often used when trying to predict the pathogenicity of a variant or gene, a model is trained on a set of instances for which the class label is known. A dataset is composed of *instances*, which are represented as a vector of characteristics, referred to as *features*. This dataset is then usually split into a training set, testing set and validation set. The features in the training set are used to determine a set of classification rules which can establish the class of an instance. These rules are fine-tuned by testing them in the testing set. The final model can then assign class labels to new instances in the validation set based on their feature values and this set of classification rules. More details on the ML procedure are presented in the Materials & Methods chapter of this thesis (Section 3.4).

There exist different types of machine learning models and their performance depends on the type of data and problem at hand. The most commonly used in bioinformatics include K-nearest neighbour, Support vector Machine, neural networks and decision tree based models [147]. The K-nearest neighbour algorithm is an example of what is called “lazy learning” and attributes a class to an instance by taking the majority class of the K-nearest points in the feature space, with the nearest points determined by some distance function between the feature vectors. Support vector machine tries to find a hyperplane that maximizes the separation between the two classes, the kernel trick making it applicable to non-linear problems. Neural networks use a network of connected artificial neurons (called nodes) organized in layers where the output of a node is used as input for another node. The weights of the connections between nodes are adapted during training and this network can then be used to attribute an output class to any input feature vector. Finally, in the case of decision trees, the decision logic is modelled using a tree structure: at each node, a test is applied on the input features and the outcome of this test takes the classification algorithm to a specific child node, this process is then repeated until it reaches a leaf node corresponding to a decision. Random forests, which are used in this work, are ensembles of decision trees: each individual decision tree is trained on a different subset of the training set and the final output of the Random Forest is obtained by majority vote of the decision trees [152].

Interpretability of machine learning models has become a huge topic of interest in the past few years, especially with regards to applications in bioinformatics. The definition of “interpretability” itself can be debated [153], but an interpretable model typically needs to have one of the 2 following properties: transparency, i.e. one should be able to understand globally how the model works; or *post-hoc* explanation, i.e. even if there is no general understanding of how the model works, one should be able to have information on why a particular instance was attributed a particular prediction. This second understanding of interpretability is becoming increasingly popular since it allows to interpret “black-box” models (i.e. ML models for which the decision process is not straightforward) after the fact and therefore not compromise predictive power for interpretability [154]. Many *post-hoc* algorithms are therefore constantly being developed, that are either general or specific to model types [155].

In bioinformatics, there are three major reasons as to why interpretability is important [154]. The first one is to improve the performance of the model and therefore understand the potential bias that could exist in the training set or the way the model is built. The second reason is to obtain novel insights on biological problems. As mentioned previously, machine learning models are sometimes able to make unexpected associations between the input and output data, and therefore allow to detect novel biological patterns. However, this is only possible if the user has some understanding of how the model generates a particular prediction. The last reason for wanting to have an interpretable model is to be able to be confident about the prediction. This is especially important in clinical applications where the user needs to be able to interrogate and understand the prediction made by such a model, especially when it is involved in "high stakes" situations where a prediction error can have severe consequences [156, 157].

1.3.3 Classification tools for pathogenicity

Traditional approaches to variant pathogenicity prediction, used by predictors such as Sort Intolerant From Tolerant (SIFT) [158] or PolyPhen2 [159], are based on sequence conservation and protein structure to predict the consequence of a missense change on protein function. These predictors are therefore limited to missense variants and have a relatively high false positive rate, which means that for an individual, they tend to output a large number of potentially damaging variants, instead of identifying the actual disease-causing one [160].

New variant classification tools, which have been more successful, are meta-predictors, which means that they integrate information from many other tools to generate an overall pathogenicity score. Examples of such tools include Combined Annotation-Dependent Depletion (CADD) — which uses a Support Vector Machine model to score variants based on annotations from 63 different tools [161] — and the Rare Exome Variant Ensemble Learner (REVEL) — which integrates 18 individual pathogenicity scores from 13 different prediction tools into a Random Forest predictor [162].

More recently, deep learning approaches have also shown very good performance on variant pathogenicity classification. Some of these predictors also appear to function better than the aforementioned approaches because they rely on unsupervised learning, which prevents them from learning biases introduced in manually curated datasets [163]. For example, EVE relies on a variational autoencoder to learn the latent rules that underlie a multiple sequence alignment [164], and AlphaMissense supplements this approach with structural context from AlphaFold (a deep-learning based protein structure prediction model) [165, 166].

Most of the methods that have been developed so far are therefore focused on single variant or single gene prediction. However, increased knowledge on digenic diseases has made the application of such methods to variant combinations possible. The first predictor targeting directly combinations of variants, the Variant Combination Pathogenicity Predictor (VarCoPP),

was developed in 2019, and classifies variant combinations as either pathogenic or neutral [105]. A gene pair pathogenicity predictor was also recently published in 2021, and uses a similar model as VarCoPP but with the use of novel pair features to identify gene pairs that are likely to be involved in a digenic disease [167].

1.3.4 Prioritization methods

Variant prioritization is the process of ranking or selecting the subset of variants from an individual genome that is the most likely to be causative of the patient's disease. Variant prioritization tools can be divided into different categories depending on the types of variants they predict, the level of knowledge required to predict these variants (what variant characteristics are used) and the type of prediction they output (ranking of the variants or selection of potentially damaging variants) [123]. Three main categories of prioritization methods have shown most success: text-mining-based, network-based or machine-learning-based. The latter is becoming increasingly popular as more biological data becomes available [168].

Although they can in theory identify novel genetic causes to disease, prioritization tools are more oriented towards achieving the precision medicine goal: based on variants characteristics and background information about the patient's phenotype, their primary objective is to identify the variants that are most likely to be causative of the disease and thus help obtain a molecular diagnostic. These tools are becoming particularly important as sequencing is now moving in the WES and WGS era (see Section 1.2.4).

Numerous methods have been developed to perform this task [160]. They differ by their information sources, methods of integrating these sources, types of variants or genes they prioritize. One source of information that has proven particularly useful in both gene and variant prioritization is phenotype information, which is now collected and standardized in the Human Phenotype Ontology (HPO).

These so called "Phenotype-driven" approaches to prioritization have been shown to outperform other methods, and are based on the idea that diseases causing similar symptoms are likely to be caused by similar genes [169]. These methods take as input both the patient's phenotype (encoded as HPO terms) as well as the patient's genotype (in the form of a VCF file or other standard variant file). Most of these tools involve scoring using a phenotype relevance component and a pathogenicity component, which are integrated to obtain a final ranking. The pathogenicity component evaluates the potential of a variant or gene to be involved in a disease mechanism, and is based on the MAF of the variants [170], pathogenicity predictions from single variant classifiers [171, 172] or based on the variant effect [173]. The phenotype-relevance component scores how likely it is for a gene to be involved in the phenotype of the patient. It is based on previous knowledge about gene-phenotype associations, which is expanded to score all genes using various methods such as network propagation methods

in Protein-Protein interaction (PPI) networks or biomedical ontologies [171, 174] or through cross-species comparison using semantic similarity [170, 175–177]. They are based on the guilt-by-association principle and exploit previous knowledge about gene-phenotype associations to identify genes with similar profiles.

In particular, network-based methods have been extensively used, and rely on the intuitive aggregation of different sources of information into a network [178]. Different studies have shown that disease genes are organized into disease modules, which means that novel gene disease associations can be discovered by looking for genes that are closely related to known disease-genes in different types of networks [179, 180]. Based on this hypothesis, several prioritization methods have been developed that uses proximity to known disease genes in different networks to help prioritize variants or genes [174, 181–183].

A recent review about single-variant phenotype-based prioritization tools [125] identifies as top performing single-variant prioritization methods AMELIE [184], a text-mining based tool, LIRICAL [185], which uses likelihood-ratio framework based on the Exomiser predictor and X-rare [186], which incorporates different variant pathogenicity methods with different phenotype relevance scores into a machine learning model. Nonetheless, these methods are tailored to identify single variant causes to disease, and may thus not be useful for the detection of more complex patterns of inheritance.

1.3.5 A first generation of digenic predictive tools

Although several tools have been developed in order to predict the deleteriousness of variants in single genes (see Section 1.3), the first variant combinations pathogenicity predictor was created in 2019 [105]. The VarCoPP model is a machine learning predictor which, for the first time, directly predicted the pathogenicity of variant combinations in gene pairs. It was trained on the combinations found in DIDA as positive disease-causing instances [104], and on combinations found in individuals of the 1KGP project as negative instances [37]. The predictor uses a set of 11 features at the variant, gene and gene pair level to make predictions, and was shown to have good performance in both cross-validation and on an independent set, consisting of combinations reported as disease causing in later update of the DIDA database [105].

In order to make the study of oligogenic diseases more accessible, an online platform was developed, which integrates the VarCoPP predictor, as well as the Digenic Effect (DE) predictor [187], which predicts the probability of a variant combination to be one of the three main types of digenic models (see Section 1.2.3). The Oligogenic Resource for Variant Analysis (ORVAL) allows for user-friendly exploration of the pathogenicity of oligogenic variant combinations [188]. It takes as input a list of variants, filters them according to the types of variants that the predictors can integrate, annotates them with features from different biological levels, creates all possible variant combinations between the filtered variants and predicts them using the two

tools. The predicted disease-causing variant combinations are then visualized as a network, to facilitate module detection. An explanation of the different predictions is available by using the *treeinterpreter* module, which evaluates how each feature contributed to the decision of the classifier. This platform has been used in numerous studies to help identify novel oligogenic causes to disease [189–193].

Although the performance of VarCoPP is promising, it still has some limitations. The model has a very complex structure, which not only prevents the expansion of the training set, but also makes it computationally heavy, leading to increased prediction time, especially for larger datasets such as exome sequencing data. Additionally, the model was trained on DIDA, which contains data that was not evaluated based on its quality, and is not up-to-date. Finally, although it performs well for the analysis of gene panels, the false-positive rate makes the analysis of unfiltered datasets impossible, since the predictor returns too many irrelevant combinations as potentially disease-causing.

In addition to VarCoPP, a first prioritization tool for for oligogenic variant combinations was also developed: OligoPVP [194]. It was developed based on DeepPVP, a single variant phenotype-driven prioritization tool using a neural network [176]. The predictions of DeepPVP are then combined with knowledge about PPI to rank oligogenic combinations in two and more genes (see Section 3.5.4 for a detailed description of the tool).

1.3.6 Cohort analyses

The tools and methods described above all aim at predicting the pathogenicity of either specific variants or predicting the most likely pathogenic variants in a patient's exomes. However, they do not propose a method to detect rare disease signatures in several patients affected with the same disease for example, or provide a way to analyze directly whole cohorts, which are becoming increasingly available.

Approaches to find relevant genes and variants for particular diseases within whole cohorts have also been developed. Genome Wide Association Study (GWAS) for example, has been shown to be efficient to find association between common variants and diseases [195]. However, variants associated with rare diseases usually have very low frequency, and are therefore not likely to be found in several patients within the same cohort. Rare variant association strategies have therefore been developed, and rely on aggregating variant information inside biologically meaningful regions such as exons or genes [196–199].

Most association tests are case-control studies, where the goal is to identify genes for which patients with a particular phenotype are found to have more variants compared to individuals who do not present with the phenotype. These tests are called burden tests, and are run by simply aggregating the number of variant carriers in a biological unit (often a gene), and assess the difference in frequency of carriers between the cases and the controls [196–199].

Different burden tests have been developed to consider covariates (e.g. age, sex or body mass index), model family structure, and take into account the fact that variants in the same gene can have opposite effects [200–202]. The main assumption behind burden tests is that a large proportion of variants in the regions are causal [196], which is why the choice of which variants to include in such burden tests is important [197]. Classification and prioritization methods described above can be used as a first step to select only variants predicted as disease-causing to be counted in the burden [197].

Finally, some association tests have been designed to take into account genetic interactions and thus are potentially able to detect digenic inheritance models. Some are based on case-control studies as mentioned before, but also integrate gene-gene interaction score or dimensional reduction techniques to test for interaction between all possible SNVs [203–205]. More recently, a case only method was also developed, and was shown to be able to identify several digenic combinations in a cohort of patients with craniosynostosis [206, 207].

1.4 An introduction to the genetics of male infertility

Infertility is estimated to affect between 8 and 12% of couples worldwide, with a prevalence increasing in the past few years [208]. In roughly 50% of cases, males are either the primary cause or contribute significantly [209]. Male infertility is a complex condition which can be caused by many factors, and affects about 7% of males worldwide [210].

In this section, we introduce basic concepts and current knowledge about male infertility genetics, to provide a background for the study of the oligogenic basis of male infertility which is presented in Chapter 6 of this thesis' work.

1.4.1 A multifactorial condition

The biological mechanisms underlying male infertility are highly diverse. Proper functioning of the male reproductive system relies on several key steps. The hypothalamic–pituitary axis must function to stimulate the testes through follicle-stimulating hormone (FSH) and luteinizing hormone (LH). This signal enables the spermatogenic and androgenic compartments of the testes to produce sufficient amounts of functional spermatozoa (spermatogenesis) and testosterone (androgenesis). Finally, sperm must be transported via the male reproductive ductal system.

Based on these main steps, the etiology of male infertility is categorized into four main groups [211]: (i) quantitative spermatogenic failure, (ii) ductal obstruction or dysfunction, (iii) alterations of the hypothalamic–pituitary axis and (iv) qualitative spermatogenic disturbances.

Quantitative spermatogenic failure accounts for about 75% of cases, and encompasses conditions where the issue is the insufficient production of spermatozoids due to primary failure of spermatogenesis. This can lead to *oligozoospermia*, which refers to low number of spermatozoids in the ejaculate (concentration <15 million per mL), or *Non-Obstructive Azoospermia (NOA)*, the most severe case, which refers to the complete absence of spermatozoid production. Primary spermatogenic failure can be congenital or acquired [211]. One important cause of quantitative spermatogenic failure is cryptorchidism, one of the most common congenital abnormalities among men, which is presumed to be at least partly genetic. This condition refers to the failure of one or both testes to permanently descend, which can lead to azoospermia in 98% of untreated bilateral cases and 13% of unilateral cases [212].

Azoospermia can also result from ductal obstruction or dysfunction. In this case, it is called *Obstructive Azoospermia (OA)*, and it is the second main cause of infertility, affecting 40% of all azoospermic men [213]. Obstructive azoospermia occurs when there is a blockage or abnormality in the male reproductive ductal system, preventing the transport of sperm. Common causes include congenital abnormalities, infections, and prior surgeries.

Furthermore, both oligozoospermia and NOA can result from defects in the hypothalamic-pituitary axis, known as hypogonadotropic hypogonadism or secondary spermatogenic failure. Hypogonadotropic hypogonadism occurs in 2/3 of cases from Kallman syndrome [214], a genetic disorder characterized by disturbances in the endocrine system.

Finally, male infertility can result from qualitative spermatogenic disturbances, where the main issue is the motility, morphology, or viability of sperm. These disturbances may be caused by genetic factors, hormonal imbalances, environmental factors, or lifestyle choices such as smoking and excessive alcohol consumption. Abnormal sperm morphology and motility can significantly impair the ability of sperm to fertilize an egg, leading to infertility.

1.4.2 Genetic basis of male infertility

Due to the variety of biological factors that can play a role in male infertility, the genetic basis is also very heterogeneous. It is estimated that about 15% of male infertility cases have genetic explanations [215], although genetic factors are thought to play a bigger role since 72 to 75% cases are currently categorized as idiopathic, which means that the cause of the disorder can not yet be identified (Figure 1.6A) [216–218].

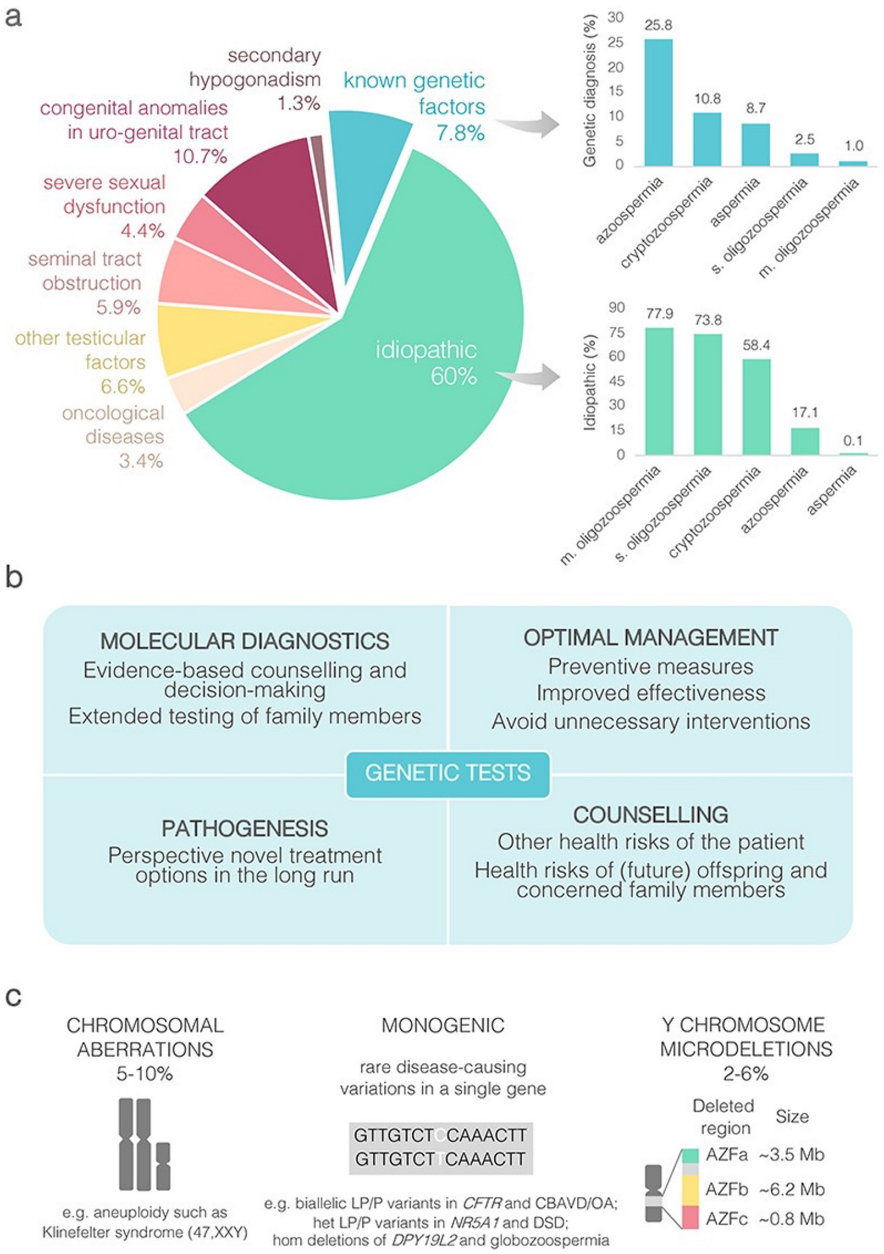


Figure 1.6: Overview of current knowledge in male infertility genetics. (a) Aetiology of male infertility and proportion of genetically diagnosed and idiopathic cases in different patient subgroups (based on data from [216]). (b) Different ways genetic testing can help better manage patients. (c) Current established genetic causes of male infertility. Figure from [218]

The genetic causes of male infertility have been linked to chromosomal alterations, Y chromosome microdeletions and monogenic mutations in specific genes (Figure 1.6c). Chromosomal alterations mostly affect sex chromosomes, with the most common known genetic cause of infertility being Klinefelter's syndrome, which is characterized by extra X chromosomes in male individuals, resulting in a 47,XXY, 48,XXXY or 49,XXXXY karyotype [219]. Most men with Klinefelter's syndrome present with oligozoospermia or azoospermia, resulting from abnormal testicular development or hormonal imbalance [219].

Several studies have associated microdeletions in the long arm of chromosome Y [220, 221], which is known as "azoospermia factor" to oligozoospermia [222] and azoospermia [223] (Figure 1.6c). Three regions have been identified within this chromosomal region as critical for spermatogenesis: AZFa, AZFb and AZFc [221]. They comprise of genes and transcriptional units which are almost exclusively expressed exclusively in testicular tissue [215]. AZFc microdeletions often result from homologous recombination between repeated sequences, which are frequent in this chromosome, impacting genes crucial for spermatogenesis [224]. These deletions are linked to phenotypic variability, from total absence of germ cells to severe oligozoospermia.

Monogenic variants causing male infertility have also been discovered (Figure 1.6c). The number of genes associated with the disease has been growing fast [218], from a first sytematic review in 2019 identifying 78 distinct genes [225], to 108 genes reported in 2021 [226], while an article published this year found close to 500 genes reported to be associated to the disease [227]. A large proportion of monogenic cases are associated with mutations in the CFTR gene, which when presenting with homozygous mutations, causes cystic fibrosis [228]. The presence of mutations that do not completely impair the function of this gene have been shown to cause OA by inducing incomplete formation of the Vas Deferens and Congenital Bilateral Absence of the Vas Deferens (CBAVD) [229]. CBAVD is present in 6% of males affected with OA, and in about 2% infertile patients [229].

Several genes are also known to cause Kallmann syndrome, which is the major cause of hypogonadotropic hypogonadism, leading to abnormalities in the secretion of hormones that stimulate spermatogenesis. The main gene associated to Kallmann syndrome is KAL1 (also known as ANOS1), a gene located on the X-chromosome which encode a protein with a central role in the migration of GnRH-secreting neurons to the hypothalamus. However, other genes have also been shown to be involved in the inheritance of Kallmann syndrome, such as genes involved in pathways related to the development and migration of gonadotropin-releasing hormone (GnRH) neurons [230–232]. In particular, 10 to 12% of Kallmann syndrome cases have been associated with oligogenic inheritance models [233, 234].

Overall, the genetic basis of male infertility remains poorly understood, with more than 500 genes linked to the phenotype in mice [235]. The genes identified as associated to the disease are involved in many of the key steps leading to SPGF, including the Hypothalamic-Pituitary-Gonadal (HPG) axis, the development of reproductive organs, sperm motility, spermiogenesis, spermatogenesis and meiosis [236–238]. Moreover, links between genetic causes of SPGF and Primary Ovarian Insufficiency (POI) have been documented [218]. These genes are therefore linked to a heterogeneous spectrum of phenotypes, including Disorder of Sexual Development (DSD), Congenital Hypogonadotropic Hypogonadism (CHH) as well as the extremely rare defects of sperm morphology (teratozoospermia) and motility (asthenozoospermia) [218, 226]. Monogenic causes for quantitative spermatogenic failure, including NOA and severe OZ, remain difficult to identify, in part due to the fact that pedigree studies for such disorders are limited, since the patients are usually the only affected individual [218, 237].

Recently, several cohorts of patients affected by male infertility have been recruited and sequenced, in order to obtain novel insights into the genetics of this condition [227, 239–241]. They provide a unique opportunity to develop novel approaches to the molecular diagnostic of male infertility such as searching for digenic and oligogenic causes, which have been identified in CHH and DSD [242, 243], as well as test for potential comorbidities such as cancers or cardiometabolic diseases [218, 244]

1.4.3 Genetic screening and implications for treatment and family planning

Properly identifying the cause of the infertility in each case is essential in order to assist patients in their desire to have children. In particular, understanding the mechanisms of the disease can help with prevention and treatment, lead to be better clinical management to avoid unnecessary interventions, and help with the assessment of potential health risks for relatives as well as with the assessment of potential comorbidities to the patients (Figure 1.6b).

Guidelines have been developed to precise clinical management of the condition and have started to include genetic testing as part of the clinical work-up, yet it remains limited to a small number of genetic tests [245]. Clinical testing usually begins with semen analysis, which allows to evaluate the quantity of spermatozoids and therefore to classify the individual as either azoospermic, oligozoospermic or normozoospermic [246]. Results from this analysis are evaluated based on WHO standards [247], and the patient is then referred to a reproductive specialist. The most common therapeutic intervention is Assisted Reproductive Technology (ART) [246]. In some cases, hormone levels are also tested, to assess whether SPGF could be due to secondary hypogonadism. If this is the case, hormonal treatments are also possible [248].

Genetic testing is for the major part limited to chromosomal aberrations, Y chromosome microdeletions, and mutations in the CFTR gene, which, taken altogether, only explain about 10% of all infertility cases [216, 246]. The results of such genetics tests can directly have prognostic impact for the patients and their offspring, since men diagnosed with Klinefelter's Syndrome have lower chances of successful ART, while men with CFTR mutations have increased risks to have children affected with cystic fibrosis [217]. Recently, WES has also been applied in an attempt to diagnose more cases, although the results are more complex to interpret and lead, in the best case scenario in $\sim$ 10% diagnostic rate [227].

Overall, there is therefore an important need to improve clinical genetic testing of infertile men in order to be able to develop precision medicine approaches to the disease management [249, 250]. Recent advancements in the identification of novel genes involved in the disease shows great promise for a better understanding of the molecular basis of SPGF which will translate in better clinical management [250].

Chapter 2

Research objectives and outline

With WES becoming part of the standard diagnostic pipeline for patients with suspected genetic disorders, developing appropriate methods to interpret this data is becoming essential. In particular, such methods should also be able to detect oligogenic inheritance models, as such patterns of inheritance are likely to explain part of the cases with missing diagnoses.

In this chapter, we define the problem we are addressing with this thesis, as well as precise our research questions and two main hypotheses. First, we hypothesize that the development of a prioritization tool for the detection of oligogenic variant in WES data is possible. Second, we hypothesize that applying this tool in a cohort of patients affected by male infertility can generate interesting insights into the genetics underlying this highly heterogeneous disease.

Finally, we also provide the thesis outline and highlight the key publications and contributions associated to this work.

2.1 Problem definition

Millions of individuals are affected by rare diseases worldwide. Most of these disorders are thought to have a genetic basis, but the exact genetic causes remain unresolved in the majority of patients, leading to errors in diagnosis and treatment. Despite significant advances in the field of human genetics, facilitated by the development of new sequencing technologies and the advent of novel machine learning methods to predict the pathogenicity of genetic variants, the majority of patients remain undiagnosed leading to inadequate treatment and significant burden.

Part of the missing heritability problem might be due to the fact that genetic diseases are actually caused and modulated by the interaction of variants in multiple genes, through epistatic mechanisms [251, 252]. Although such oligogenic inheritance models have started to receive attention from researchers, they remain hard to identify and predict due to an insufficient amount of data and limitations in the state-of-the-art oligogenic prediction methods. With the increased usage of WES, digenic predictive methods should be improved so they are also able to tackle this type of data. Yet, since this type of sequencing produces large amount of variants, the number of combinations to evaluate increases tremendously, calling for proper software to assist clinicians in such task.

Getting a better understanding of the genetic mechanisms underlying heterogeneous diseases is especially important, as it brings forward opportunities to improve treatments through precision medicine. This is particularly true for male infertility, a relatively common condition which can be caused by a variety of factors. In infertility cases with a genetic basis, a precise understanding of the mechanisms and genes involved is key to provide an appropriate treatment to the patient. A certain proportion of patients exhibiting oligogenic models of inheritance have already been discovered for this disease, and the establishment of larger cohorts for this disease opens the way to new discoveries.

To create a robust oligogenic predictor for WES, there is a need for better data regarding oligogenic variant combinations. While DIDA was an important first step, it has not been maintained over the years, is limited to combinations of variants in two genes, and does not provide an assessment of the quality of the curated entries (see Section 1.3.5). The number of articles reporting oligogenic causes to disease – including variants in two or more genes – have been significantly increasing over the past few years, suggesting the possibility to improve digenic predictions through the collection of a larger dataset of higher quality. This, in turn, will lead to improved diagnostic accuracy and treatment options for patients with rare genetic diseases.

2.2 Research questions and objectives

This project is centered around answering three main research questions:

- How can we improve the quality of datasets on pathogenic oligogenic combinations?
- How can we develop an efficient oligogenic prioritization tool that is able to detect, in patient's exome data, which variant combinations are most likely to explain the patient's phenotype?
- How useful will this tool be in identifying meaningful oligogenic signatures in whole cohorts?

First, we hypothesize that, using machine learning approaches trained on high quality data, we can build a prioritization method that sufficiently reduces the number of False Positive (FP) cases to make the detection of pathogenic variant combinations in exomes feasible. Secondly, we hypothesize that this tool is relevant to the analysis of whole cohorts, as it will allow to uncover oligogenic inheritance models underlying specific patient groups. By addressing these two hypotheses, we aim to achieve two main research objectives.

The first goal of this project is to advance oligogenic predictions and develop an oligogenic variant prioritization method. Prioritization methods are needed (and go beyond simple pathogenicity prediction) as they provide an objective mechanism to identify the most relevant gene variants (as opposed to panels). This will be the first prioritization tool that will directly target the prioritization of variant combinations (instead of individual variants), by exploiting the relationship between the involved genes and variants. This variant combination prioritization approach will allow for an unbiased analysis of WES of patients with diseases suspected to be attributed to oligogenic models.

The second goal of this project is to generate insights into the genetic nature of male infertility by identifying meaningful oligogenic signatures for the disease. We will use our novel prioritization tool on the WES data of infertile men in order to generate novel insights on the genetic architecture underlying this heterogeneous disease. The analysis of this cohort will lead to the establishment of a computational pipeline to detect oligogenic causes in whole cohorts and the discovery of specific novel candidate genes and gene pairs for male infertility.

2.3 Thesis outline

The project is divided into 3 main parts, which correspond to three separate chapters: collecting high-quality oligogenic ground-truth data, creating an oligogenic prioritization tool using machine learning approaches trained on the collected ground truth data, and analyzing patients cohort data with the newly developed method with the aim to detect meaningful oligogenic signatures.

Ensuring high-quality oligogenic data for predictions

In order to develop efficient and reliable machine learning predictors, a certain amount of high-quality data is required. Since DIDA, which has previously been used as the basis for different oligogenic predictors, had not been updated for some years, and the quality of the variant combinations was not properly assessed, the first part of this research was the development of a novel database for oligogenic variants. The OLIgogenic diseases DAtabase (OLIDA) [253] is presented in Chapter 4, which collects curated data on all oligogenic variant combinations reported in the scientific literature. In addition to providing even more digenic cases than those that were present in DIDA, OLIDA also contains variant combinations in more than two genes, as well as combinations which involve copy-number variants (CNVs) which were not present in DIDA. Through its thorough curation protocol, OLIDA also assigns confidence scores to every combination in the database, based on a quantification of the quality of evidence that supports the involvement of the variants in the disease.

Investigation into the prioritization of oligogenic variant combinations

The second part of this research investigates how to create a prioritization tool specifically aimed at the detection of oligogenic variant combinations. In order to do that, we will follow a similar approach to the most successful single-variant prioritization methods, by combining a variant combination pathogenicity score with a phenotype relevance score for gene pairs. The variant combination pathogenicity score will be obtained from an improved version of VarCoPP, which is trained using the OLIDA dataset, a new set of features and a new model structure to achieve better predictions. The phenotypic relevance score is computed using propagation algorithms in a heterogeneous network integrating a various biological networks. These scores are then combined into a final prioritization score that is used for ranking. The performance of the method is assessed by using synthetic exomes, i.e. exomes from patients of the 1KGP to which we insert variants from an oligogenic variant combination not used in the training.

Chapter 5 presents two novel predictive methods: the VarCoPP2.0 predictor, an improved version of the original VarCoPP model which predicts the pathogenicity of digenic combinations faster and more accurately [254], and Hop, the High-thoughput oligogenic prioritizer, a novel prioritization method which integrates the predictions of VarCoPP2.0 together with disease information in order to rank digenic combinations directly in whole-exome sequencing data [255].

Generating novel insights into the genetics of human diseases

The third part of the project investigates how we can use the methods developed in the second part in order to generate novel insights into the genetic mechanisms underlying male infertility. To do this, we develop a computational pipeline for the analysis of a cohort of individuals affected with male infertility recruited at the university of Tartu in Estonia. We first set up a filtering protocol to retain the variants of interest and then apply the predictors developed in Chapter 5 to detect novel oligogenic causes to disease.

Based on these predictions, we analyse whether the Hop predictor is able to accurately capture the oligogenic combinations that were already manually discovered, before we perform a more exploratory analysis to discover new causes to disease. In particular, we demonstrate that Hop allows to identify gene pairs that appear to be more frequent in patients compared to controls, and also identifies potentially relevant novel genes that could be implicated in the male infertility phenotype.

Finally, we present a discussion and conclusion of our work in Chapter 7, where we highlight the main scientific contributions of this thesis, as well as the limitations of our research which opens the way to new research questions and projects.

In general, the “we” pronoun is used in the chapters describing contributions of this thesis. This pronoun refers to personal individual work with a few exceptions which are the results of collaborations and which are specified in the relevant sections. In Chapter 4, the development of OLIDA and its curation protocol resulted from a collaboration with two other researchers, and the curation work involved 4 people (including myself). Furthermore, in Chapter 5, the development of VarCoPP2.0 was a collaboration with two other colleagues, and the integration of Hop and VarCoPP2.0 in the ORVAL platform was also a collaborative work.

2.4 Scientific publications

- **Barbara Gravel***, Charlotte Nachtegael*, Arnau Dillen, Guillaume Smits, Ann Nowé, Sofia Papadimitriou, Tom Lenaerts, Scaling up oligogenic diseases research with OLIDA: the Oligogenic Diseases *Database, Database, Volume 2022, 2022, baac023, https://doi.org/10.1093/database/baac023* [253] * Joint first authors
- Nassim Versbraegen, **Barbara Gravel**, Charlotte Nachtegael, Alexandre Renaux, Emma Verkinderen, Ann Nowé, Tom Lenaerts and Sofia Papadimitriou, Faster and more accurate pathogenic combination predictions with VarCoPP2.0, *BMC Bioinformatics 24, 179 (2023). https://doi.org/10.1186/s12859-023-05291-3* [254]

- Sofia Papadimitriou, **Barbara Gravel**, Charlotte Nachtegael, Elfride De Baere, Bart Loeys, Miikka Vikkula, Guillaume Smits, Tom Lenaerts, Toward reporting standards for the pathogenicity of variant combinations involved in multilocus/oligogenic diseases, *HGG advances*, 4(1), 100165. https://doi.org/10.1016/j.xhgg.2022.100165 [256]
- **Barbara Gravel**, Alexandre Renaux, Sofia Papadimitriou, Guillaume Smits, Ann Nowé, Tom Lenaerts, Prioritizing oligogenic combinations in whole exomes, *Bioinformatics*, Volume 40, Issue 4, April 2024, btae184, https://doi.org/10.1093/bioinformatics/btae184 [255]

2.5 Open data and softwares

This work has led to the release of one open dataset and two bioinformatics software.

Open data

- OLIDA - OLIgogenic diseases DAtabase [253]
  - **Description:** Website platform to access all the curated data on oligogenic variant combinations
  - **Availability:**
    - Web platform with API: `https://olida.ibsquare.be`
    - Tables, with the different versions: `https://zenodo.org/records/10732286`, `https://doi.org/10.5281/zenodo.10731105`
  - **License:** CC-BY-NC 4.0

Open softwares

- VarCoPP2.0 - Variant Combination Pathogenicity Predictor 2.0 [254]
  - **Description:** ML predictor to predict the pathogenicity of variant combinations in gene pairs
  - **Availability:**
    - Web platform: `https://orval.ibsquare.be`
    - Code to reproduce the paper results: `https://github.com/oligogenic/VarCoPP2.0.git`
  - **License:** CC-BY-NC 4.0
- Hop - High-throughput oligogenic prioritizer [255]
  - **Description:** Prioritization tool that ranks in a patient's exome variant combinations based on how likely they are to cause the patient's disease.
  - **Availability:**

    * Code to reproduce the related publication's results: https://github.com/oligogenic/HOP.git
    * Python package to run the tool on new data: https://pypi.org/project/oligopipe/
  – **License:** CC-BY-NC 4.0

2.6 Supervised Master Theses

- Bosch, I. Knowledge graph embeddings for the prediction of pathogenic gene pairs. *MSc. Thesis. MSc in Bioinformatics and Modelling - Université Libre de Bruxelles (2023)* - Thesis co-supervised by Gravel,. B., Renaux, A., Lenaerts, T.
- Vidal Bankier, D. Oligogenic Markers of Antiseizure Medication Response. *MSc. Thesis. MSc in Medicine - Université Libre de Bruxelles (2024)* - Thesis co-supervised by Gravel, B., Depondt, C., Smits, G., Lenaerts, T.

2.7 Fundings

This PhD project was supported by the Fonds de la Recherche Scientifique (F.R.S-F.N.R.S) with the Fund for Research Training in Industry and Agriculture (40008622), and the Research Foundation-Flanders (F.W.O.) Infrastructure project associated with ELIXIR Belgium (I002819N).

Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

We also thank the Foundation 101 Genomes (f101g.org) for fruitful collaboration,creative exchange and scientific support.

Chapter 3

Materials & Methods

In this chapter, we introduce the material and technical elements necessary to understand the work presented in the thesis.

First, we describe the datasets of human genetic variation that are used in this work and the general format used to collect variant data (Section 3.1). Second, we present the various biological and biomedical databases and ontologies, which have, through the years, collected relevant information about genetic variants and human diseases (Section 3.2). We then focus on introducing concepts of network biology, and how this type of data representation allows to generate new knowledge, but also to visualize results in a manner that is helpful for clinicians (Section 3.3).

In a fourth part, we lay out machine learning concepts which are necessary to understand the tools and methods reported in this thesis work, highlighting the differences between classification and prioritization algorithms for variant interpretation (Section 3.4). Finally, we introduce in more details the existing resources which collect, analyze, predict and interpret oligogenic diseases prior to this thesis' work (Section 3.5).

3.1 Datasets of human genetic variation

Research in genetics has been facilitated by the collection of large amounts of data regarding human genetic variation. This has been allowed by the decreasing costs of sequencing technologies and the launch of large consortiums to study rare diseases (see Section 1.1.3). In this section, we describe the datasets of human genetic variations that are used in this thesis' work. First, we introduce background population datasets which collect data on individuals that are not linked to particular diseases, and second, we introduce the ESTonian ANDrology (ESTAND) cohort, which collects data on patients and control individuals for studying male infertility.

3.1.1 Large population datasets

Several large population datasets have been developed over the years, providing data on background genetic variation across different populations. We here describe a subset of them which are used in this thesis to generate synthetic exomes or annotate genetic variants based on their allele frequency in these reference populations.

The 1000 genomes project

The 1000 Genomes Project (1KGP) is the first large international effort to provide a detailed catalogue of human genetic variation [36,37]. Initiated in 2008, the project aimed to sequence the genomes of 1000 individuals from around the globe. The dataset now comprises of deep sequencing exome and genome data of 2504 healthy individuals, from 26 different populations belonging to 5 main continental populations: Europeans, Africans, East Asians, South Asians and Americans (Figure 3.1). This dataset has enabled significant insights into the diversity of human genetic variation [37], as well as facilitated genotype imputation to support GWAS studies, provided background frequencies to filter non-pathogenic variants from large sequencing projects and improved our understanding of population genetics [257].

This dataset has been extensively studied and used throughout the years. The data has been sequenced several times, as well as mapped to the different assemblies of the genome (see Section 1.1.2), making it one of the large population dataset with highest accuracy in terms of variant calls [259]. Since it is publicly available, and contains data on a wide diversity of populations, it can also be used as background to infer genetic ancestry of individuals in other cohorts and account for population stratification [260, 261].

In Chapter 5, we use this dataset as background information for the training and evaluation of two predictors of pathogenicity of variant combinations.

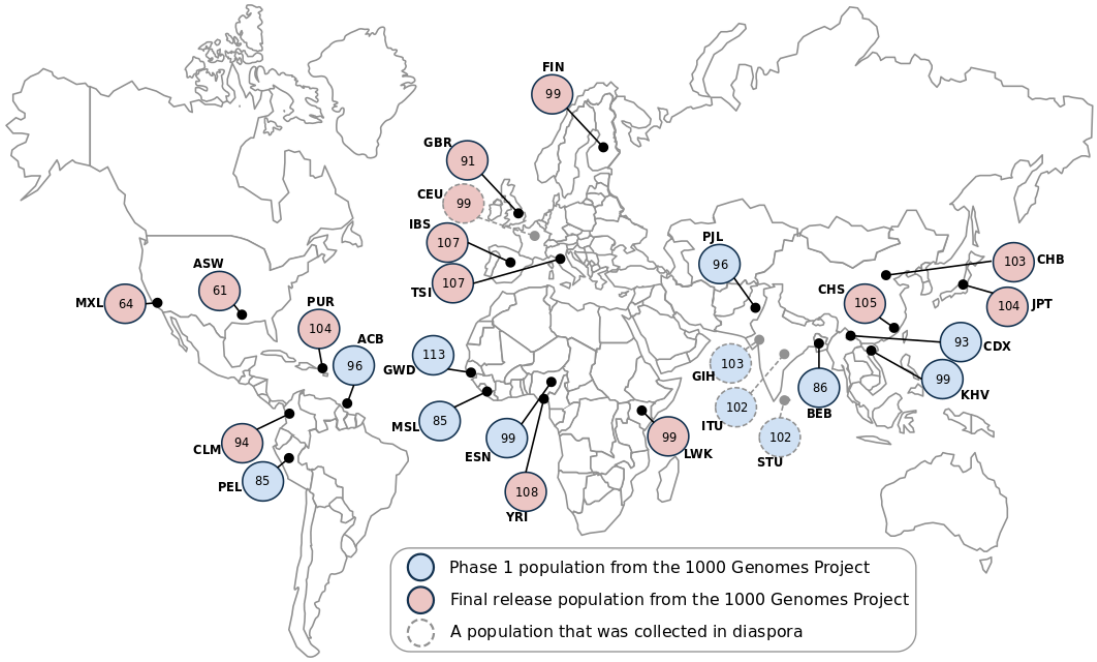


Figure 3.1: Worldwide locations of population samples with the whole genome data from the 1000 Genome Project. Each circle represents the number of genome sequences publicly available. ASIA: BEB Bengali in Bangladesh; CDX Chinese Dai in Xishuangbanna, China; CHB Han Chinese in Bejing, China; CHS Southern Han Chinese, China; GIH Gujarati Indian in Houston,TX; ITU Indian Telugu in the UK; JPT Japanese in Tokyo, Japan; KHV Kinh in Ho Chi Minh City, Vietnam; PJL Punjabi in Lahore, Pakistan; STU Sri Lankan Tamil in the UK. AFRICA: ACB African Caribbean in Barbados; ASW African Ancestry in Southwest USA; ESN Esan in Nigeria; GWD Western Division, The Gambia; LWK Luhya in Webuye, Kenya; MSL Mende in Sierra Leone; YRI Yoruba in Ibadan, Nigeria; EUROPE: CEU Utah residents with Northern and Western European ancestry, USA; FIN Finnish in Finland; GBR British in England and Scotland; IBS Iberian in Spain; TSI Toscani in Italy; THE AMERICAS: CLM Colombian in Medelin, Colombia; MXL Mexican Ancestry in Los Angeles, USA; PEL Peruvian in Lima, Peru; PUR Puerto Rican in Puerto Rico. Each circle represents the number of sequences in the final release. The dotted circles indicate populations that were collected in diaspora. Figure and caption from [258]

The UK10K project

The UK10K project is a large scale project designed to characterize genetic variation with high coverage in the United Kingdom using WES and WGS technologies.

The first goal of the project was to assess how genome-wide genetic diversity influences various quantitative characteristics, by collecting genotype and phenotype data on 3,781 healthy individuals from two British cohorts of European ancestry: the Avon Longitudinal Study of Parents and Children (ALSPAC) [262] and TwinsUK [263]. These two cohorts were sequenced using WGS and different phenotypes were measured such as the Body Mass Index, blood pressure, etc.

The second goal of the UK10K project was to identify causal mutations for more than 6,000 individuals affected with rare diseases, severe obesity and neurodevelopmental disorders [264, 265]. These individuals were sequenced using WES and are not discussed further since this data is not used in this thesis.

The WGS data from the healthy population offered interesting insights, identifying 24 million novel variants and highlighting the importance of increasing sample sizes for improving power of genetic association studies. Additionally, it led to the release of a large dataset of population variation [264, 265].

In chapter 5 of this thesis, we use data from several samples of the UK10K project as background or template data to generate synthetic exomes. We had access to data of all individuals from the ALSPAC cohort. Since these individuals were sequenced using WGS and we wanted to test a tool designed for analyzing WES data, we first filtered this data to remove non-coding regions that are further than 150 bases from exon edges based exonic positions from the Consensus Coding Sequence (CCDS) gene list downloaded from the University of California Santa Cruz (UCSC) genome browser [266, 267].

The Genome Aggregation Database

The Genome Aggregation Database (GnomAD) is a resource developed by a large consortium of researchers to share aggregate exome and genome sequencing data from a variety of large-scale sequencing projects, making summary data available for the wider scientific community. It is an extension of the Exome aggregation Consortium (ExAC) database, which contained only exome variant information [268, 269].

Upgraded very recently to version 4, GnomAD now contains data from 807,162 individuals, which represents a 10-fold increase compared to the previous version, largely due to the addition of data from over 400,000 individuals from the UK Biobank [270]. Since this update is very recent (November 2023), the GnomAD database version used throughout this thesis is the preceding one and we will thus focus this description on this previous release (version 3.1).

The GnomADv3 dataset contains genetic variation data from the genomes of 76,156 individuals, with a significant proportion of individuals being of European origin ($\sim$ 40,000 genomes), and from 125,748 exomes. This database is an *aggregate database*, meaning that it does not allow retrieval of exome or genome data from specific individuals but instead provides extensive information on genes and their aggregated variation (i.e. one can get information on the variants present in a gene, but not the identifier of the individual carrying these variants). In particular, the frequency of specific alleles can be obtained in the different populations collected in GnomAD, as well as the total number of variants of different types that can be found in a gene of interest [271].

GnomAD has therefore proven to be a valuable resource in assessing *genetic constraint* in the genome, i.e. the extent to which certain genetic variants are restricted due to their deleterious effect [272, 273]. In particular, it is now routinely used to filter variants based on allele frequency, as it is the largest such repository. In addition, several gene-intolerance-to-mutation scores have been developed based on the data collected in this database [273,274].

3.1.2 The ESTAND male infertility cohort

The ESTonian ANDrology (ESTAND) male infertility cohort was recruited at the University of Tartu in Estonia, by the research group led by professor Maris Laan. This cohort was put together in order to investigate and discover new genetic causes to male infertility, a disease for which the genetic etiology is still poorly understood (see Section 1.4).

Overall, there were 844 ESTAND cohort participants which were subjected to WES. The cohort included 521 patients affected by idiopathic Spermatogenic failure (SPGF) (median age 34) and 323 normozoospermic control men (median age 31). Each individual's phenotype was assessed using the same criteria, and considered for instance extremely low sperm counts, abnormal hormonal parameters (e.g., FSH >10 IU/l; 438 of 521 SPGF cases, 84%), reduced total testis volume (<30 ml; 318 of 521, 61%), cryptorchidism and hypospadias.

Controls had sperm parameters in the normal range (median sperm count 303.1 x 10^6 per ejaculate) and were recruited during an ongoing pregnancy. Sperm quality parameters (motility and morphology) and time to pregnancy were not considered as exclusion/inclusion criteria for the control group as this information was not available for most patients. Given this limitation, the control group may include men with other forms of male infertility apart from quantitative SPGF, which was the focus of forming this study group.

The patient group included 185 individuals affected with Non-Obstructive Azoospermia (NOA), i.e. individuals with no spermatozoids in the ejaculate, 181 Oligozoospermia (OZ) cases (median sperm count 2.0x10^6 per ejaculate), i.e. individuals with low sperm counts, and 155 men with cryptorchidism and SPGF (NOA or OZ).

Patients diagnosed with known genetic and non-genetic causal factors for male infertility, such as secondary hypogonadism, seminal tract obstruction, cytogenetic abnormalities, Y-chromosome microdeletions, sexual dysfunction, androgen abuse, severe traumas and operations in the genital area or chemo- or radiotherapy were excluded from this study. SPGF due to obstruction was assessed during the andrological workup and screening of the WES dataset for the CFTR (autosomal recessive) and ADGRG2 (X-linked) genes for variants linked to congenital obstructive azoospermia and congenital absence of the vas deferens. None of the study participants carried biallelic or hemizygous disease-causing variants in CFTR or ADGRG2, respectively. Therefore, congenital obstructive azoospermia could be excluded.

The patient group includes five patients from previously published family studies: (i) one pair of brothers with the same genetic finding [275], and (ii) one pair of first cousins and their joint first cousin once removed with variable genetic findings [190]. The rest of the study group represented unrelated singleton SPGF cases and 322 normozoospermic men.

The sequencing was performed in three different centers, and details on the sequencing pipelines presented here come from [227].

The 323 control individuals were sequenced as part of the Genetics of Male Infertility Initiative (GEMINI) project at the Huntsman Cancer Institute High-Throughput Genomics Core Facility at the University of Utah in Salt Lake City, UT, USA. DNA libraries were constructed using high molecular weight DNA using the Illumina DNA Prep with Enrichment kit (cat#20025523). PCR-amplified libraries (500 ng) were enriched for exonic regions by hybridization with the IDT xGEN Exome Research Panel v2 (cat#10005152), and libraries were sequenced using 2×150 bp paired-end sequencing on Illumina NovaSeq6000 with S4 flowcell.

A small subset of the ESTAND patients (82 individuals) were also sequenced as part of the GEMINI project. However, these samples weres sequenced at McDonnell Genome Institute of Washington University in St. Louis, MO, USA. Briefly, WES was performed using an in-house exome targeting reagent capturing 39.1Mb of exome and 2×150 bp paired-end sequencing on Illumina HiSeq 4000. Average exome coverage was 80X across sequenced individuals and platforms.

The WES data of ESTAND subjects generated by the GEMINI pipeline (n=405) were jointly reanalyzed with an additional 1,855 WES datasets (of other populations) as part of GEMINI "Phase II". This reprocessing, which entailed all steps from read mapping to genotype calling, was performed at the Utah Center for Genetic Discovery, using the UCGD pipeline (v2.13). Raw sequencing reads were processed and aligned to GRCh38 assembly in an alternate contig aware manner using BWA-MEM and duplicate read marking with SAMBLASTER. Joint genotype calling of all samples was done using the Genome Analysis Toolkit (GATK; v3.6.077).

The major part of the ESTAND patients (447 individuals) were sequenced at the Next Generation Sequencing Service laboratory of the Institute for Molecular Medicine Finland (FIMM), Helsinki, Finland. For each sample, 50 ng of gDNA was processed according to the Twist Human Core Exome EF Multiplex Complete kit manual. Libraries were pooled to 8-plex reactions according to concentration. The exome enrichment was performed using Twist Comprehensive Exome probes. The captured library pools were quantified for sequencing using the KAPA Library Quantification Kit and LabChip GX Touch HT High Sensitivity assay. Sequencing was performed with the Illumina NovaSeq system. The read length for the paired-end run was 2×101bp. Sequencing resulted in an average of 59 million reads per exome with an average median target coverage of 66X. Primary sequencing analysis and variant calling were performed using the Illumina DRAGEN BioIT Platform. Only subjects with normal sex-chromosomal ploidy estimations derived from the DRAGEN output were taken forward to exclude undiagnosed 47,XXY, or other sex chromosomal abnormalities prevalent in infertile men.

A small number of samples from the initial cohort had their variants aligned to the GRCh37 assembly of the genome, and we found a few duplicate samples in the cohort. After removing the GRCh37 samples and merging the duplicates, we therefore had access to the data of 510 patients (81 sequenced as part of the GEMINI project and 429 sequenced at the FIMM) and 322 controls (sequenced as part of the GEMINI project).

3.1.3 The Variant Call Format (VCF)

Sequencing data is processed through a series of bioinformatics tools in order to transform the raw reads in what are called *variant calls*, i.e. the identified DNA bases where the genetic code of the individuals sequenced differ from the reference sequence (see Section 1.3.1). The output of bioinformatics variant-calling pipelines is usually a Variant Call Format (VCF) file, which contains for each sample sequenced, the different variants carried by the individuals as well as several metrics that quantify the quality of the variant calls and the accuracy of the sequencing technology for that particular mutation.

These measures have been standardized for easier interpretation of the variant calls and are useful to filter the VCF files in order to only retain the variants associated with higher confidence levels. We here introduce the structure of a VCF file, as well as a few of the key metrics that are present in these files, what they represent, and how they can be used to filter variants.

A VCF file contains a header section and a data section (Figure 3.2a). The header section contains a certain number of meta-information lines, which includes data on potential filtering steps previously applied to the VCF file, as well as information on how the VCF file was obtained (e.g. the reference file used to call the variants). It also includes a tab-delimited field definition line, which contains the list of samples included in the file. Each subsequent line contains data on variants, their genotype for the different samples as well as additional information as defined in the metadata section of the header.

The genotype information is encoded as numbers. A “0” represents the reference allele, while other numbers (e.g. 1 or 2) indicates the presence of the first or second alternative allele specified in the “ALT” column. A dot means that the genotype is missing for that specific allele in that specific sample.

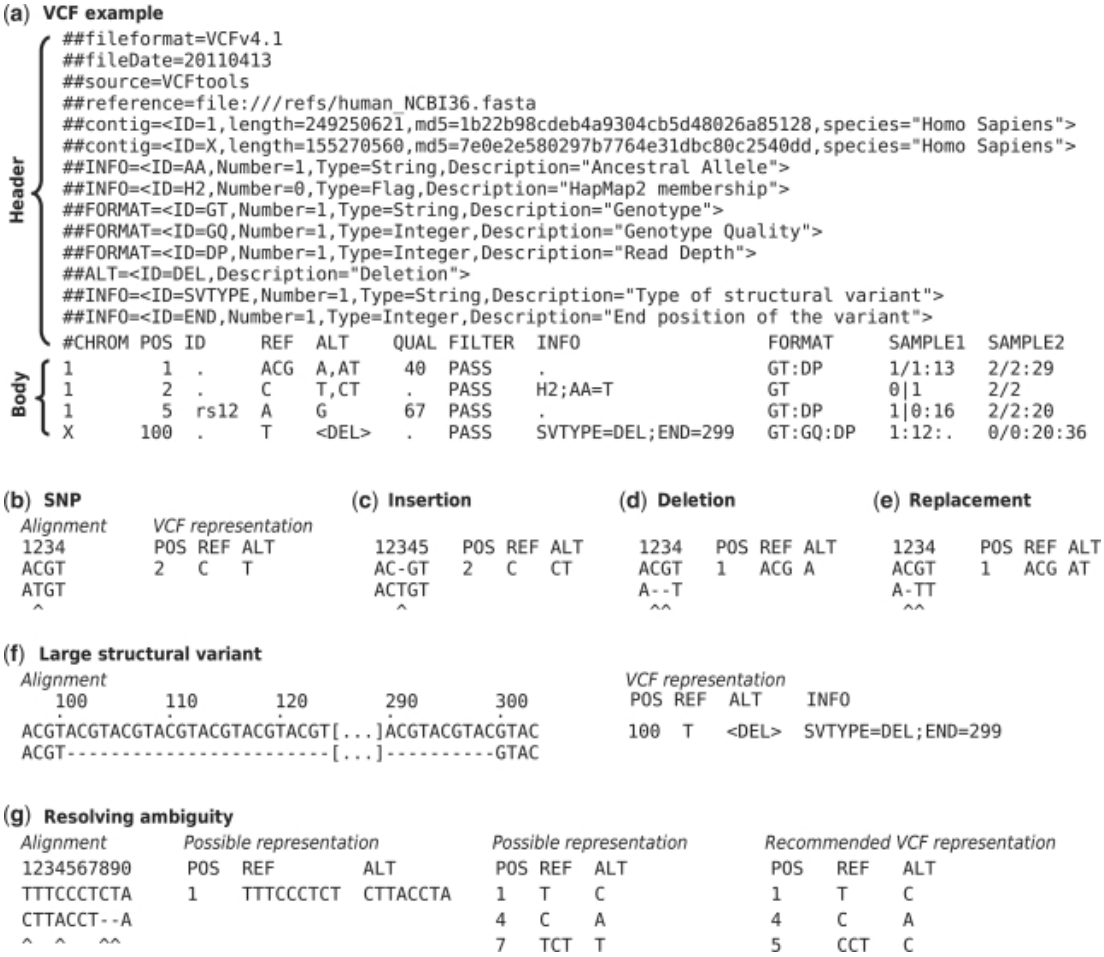


Figure 3.2: The VCF format specification: (a) example of a valid VCF file, (b–f) Alignments and VCF representations of different sequence variants: SNV, insertion, deletion, replacement, and a large deletion. Figure from [276]

Different types of variants can be represented in a VCF (see Figure 3.2(b-f)). It is important to note that for indels, different representations of the variant can exist (Figure 3.2g), which can be standardized using specific tools [277].

Some important metrics regarding variants contained in a VCF file include:

- **Read Depth (DP):** this represents the number of times that specific DNA base was covered during the sequencing process. This measure is also known as sequencing depth or depth of coverage, and a minimum value of 10 is usually required to consider a variant as relevant.
- **Genotype Quality (GQ)**: this represents the probability that the genotype call is wrong under the condition that the site is being variant.
- **Allele Count (AC)**: This represents the allele count across all samples for the genotype, in the same order as the listed alleles.

In Chapter 6, we develop a filtering pipeline for VCF files of the ESTAND cohort, where individuals were sequenced in different centers. In order to compare the datasets and obtain meaningful results, we therefore have to remove all variants which were obtained in some sets due to differences in sequencing coverage kits, as well as all variants which had low sequencing depth or genotype quality. The details of the protocol and the various filters investigated, based on the VCF information are described in Chapter 6, as they are part of the development of our analysis protocol.

3.2 Biomedical databases for disease analysis

With the advent of Next-Generation Sequencing (NGS) technologies, unprecedented amount of data have been generated to associate gene information to specific diseases. Nowadays, there are more than 100 novel gene disease associations per year [278]. To make sense of all of this information, several databases and ontologies have been developed to collect this data into an organized format. In this section, we introduce the features that are common to biomedical databases as well as describe several databases that link genes to diseases and that are used in this work to annotate data or serve as training data for the development of algorithms.

3.2.1 Main characteristics of biomedical databases

A database is a structured collection of data, which is electronically stored on a computer system. It's main function is to effectively manage, retrieve, and manipulate extensive amounts of information. Generally, databases are comprised of *tables* or files, each containing *rows* or records of data with predefined *attributes* or fields. The main objective of a database is to offer a centralized, structured, and systematic method for both storing and accessing data. This facilitates various user tasks including data entry, retrieval, modification, and analysis.

In order to be useful to the biomedical community, there are therefore certain key characteristics that are shared by the majority of biomedical databases. Recently, the importance of such characteristics has been highlighted and standardized with the FAIR principles of data management, which refer to the fact that all data resources should be Findable, Accessible, Interoperable and Reusable (FAIR) [279]. These data management principles apply to data across all fields of research, but we here focus our description on biomedical databases for studying human diseases, which are used in this work.

First, these databases are usually centered around one or a few main entities, such as human diseases, genes or variants, and potential links between these entities. In order to clearly define these entities, they are assigned identifiers, which are specific to the database, and allow for easy **finding** of the information.

These databases are made openly available by enabling easy access to the data through a website or an Application Programming Interface (API), which allows for download and querying of the database. This fulfills the **"Accessible"** criteria of the FAIR principles. It is important to note that in addition to supporting data sharing by making data accessible, many biomedical databases nowadays also enable collaboration between researchers, by providing mechanisms for data deposition, sharing, and collaborative annotation, enabling transparency and reproducibility in scientific research. This is essential to correct and avoid errors in the data collected in such resources [280].

In addition to providing the actual data, biomedical databases usually include a wide range of annotations. This additional information is usually obtained from other biomedical databases, and the use of the identifiers from these external databases therefore allows for **interoperability** between different resources. This is especially important for biomedical databases since several databases co-exist to collect information on the same entities (for example, there exist different databases for rare diseases, see Section 3.2.2 below), and enabling connections between these different resources is essential in order to access all information available. This is done by Orphanet for example, which links each disease to their identifiers in the OMIM, Genetics and Rare Diseases information center (GARD) and International Classification of Diseases (ICD) datasets [64, 281, 282]. Some annotation can also come from specific bioinformatics tools such as variant pathogenicity predictors (see Section 1.3.3).

Finally, these databases usually provide clear details about the way the data was collected, annotated and integrated in the resource. This means that the details of the curation protocol or text-mining strategy are available in the documentation, and that the sources (such as the doi of the articles where the information was obtained) are also made available. This also implies proper versioning of the data, so that it is possible to refer to a specific version of the database and retrieve the data it contained. Overall, these characteristics make a dataset **reusable**.

When designing OLIDA in Chapter 4, these characteristics were taken into account in order to ensure our database followed the FAIR principles.

3.2.2 Disease databases

We here describe several important databases collecting data on human diseases and associated genes which are used in this work.

The Orphanet database, focuses on collecting data and classifying rare diseases in order to help with diagnosis, care, and treatment of patients suffering from rare diseases. It contains curated information from the medical literature on 6,172 rare diseases, of which 71.9% are presumed genetic [52]. The repository is structured and provides a hierarchical classification of diseases, together with annotation regarding age of onset, phenotype manifestations, mode of inheritance, associated genes etc. Orphanet also provides links to many other databases collecting information on diseases.

Online Mendelian Inheritance in Man (OMIM) is the primary repository for collecting genes, disease phenotypes and the relation between them [281]. It is an extension of the original Mendelian Inheritance in Man created in 1966 [283]. The data is curated from peer-reviewed biomedical literature based on criteria that keep evolving, but prioritizes articles that provide detailed information on the relation between gene and phenotype. The data is centered around two entities: the gene and the phenotype. Each entity is associated with several relevant external links to allow for easy interoperability between the different databases. As of April 14th 2024, OMIM contains data on 4,902 genes which are associated to a disease and 7,519 phenotypes for which a molecular cause is known.

The Genomics England PanelApp, contains curated gene panels that collects information on disease gene panels and allow them to be shared and evaluated by the scientific community [80, 81]. Gene panels are typically used by clinical researchers to perform genetic testing, i.e. they are sets of genes known to be associated to a particular disease and are usually the starting point for clinicians searching for potential pathogenic variants. In such a context, including a gene that is not appropriate or not including a gene that is potentially relevant can lead to false-positive discoveries or missed diagnoses respectively. This resource therefore aims to standardize the decision-making process of which genes to include when performing gene panel based analysis. The development of this platform enables researchers to exchange and collaborate on the establishment of standard panels. This allowed to create different levels of confidence associated with genes in a panel based on reviews from experts (there are currently three different sets of gene for the panels, associated with three different confidence levels). Additionally, for each disease panel, the inheritance mode of the gene is reviewed and specified.

3.2.3 Phenotype description and the Human Phenotype Ontology

In addition to the aforementioned databases, for which the primary focus is the collection, classification and sharing of data regarding rare diseases and their genetic associations, an ontology was also developed to focus on the particular symptoms caused by these diseases: the Human Phenotype Ontology (HPO). It was initially published in 2008 by a team of researchers in Berlin and has since grown into an internationally maintained resource.

The idea behind developing such an *ontology*, i.e. a standardized and hierarchical vocabulary, was to be able to better characterize rare diseases that are phenotypically similar. This could in turn help generate new discoveries of gene-disease associations, by investigating pathways that appeared to be disrupted in different diseases with similar symptoms. Furthermore, the semantic relations between the terms can be used by computational algorithms to generate new knowledge and associations. Although it was initially designed to characterize rare diseases, the HPO vocabulary has now been extended to describe more common diseases [284].

The HPO is structured as a Directed Acyclic Graph (DAG), which is similar to a hierarchy except that a child term can be related to more than one parent term. However, cycles (cyclic paths in the graph) are not allowed. In addition to this structure, the HPO includes annotations, which link the genes to the phenotypes and vice-versa. These annotations are obtained from the associations between genes and diseases collected in databases such as OMIM, and the associations between diseases and HPOs. Therefore, a gene that is associated with two different diseases A and B will be associated with all phenotypic terms associated with diseases A and B. These associations are available for download on the HPO website.

Nowadays, the HPO contains more than 18,000 terms and 150,000 annotations. The terms stem from one of 7 main parent nodes: phenotypic abnormality (which has the most descendants), clinical modifier, past medical history, biospecimen phenotypic feature, mode of inheritance, blood group and frequency. Each term possesses a unique identifier which identifies it in the ontology DAG, and can be used to obtain its annotations.

As mentioned in Section 1.3.4, several computational algorithms have made use of the HPO to prioritize genes or variants that are likely to cause disease. Some are based solely on this ontology, but some also integrate information from phenotypic ontologies from other species such as the Mammalian Phenotype Ontology (MPO) [285] and the Zebrafish Model Organism Database [286, 287]. Semantic mappings between these different ontologies have been developed to facilitate their integrations in algorithms [288–291], and algorithms exploiting these different ontologies have been shown to perform better than when relying on the HPO alone [292].

The HPO is a community driven resource which is continuously growing. The latest update, described in a recent article from January 2024, even includes a translation of subsets of terms in 10 languages [293]. It also adds more than 2,000 new phenotypic terms, including more precise terms for describing male infertility phenotypes, which are relevant for our work [293]. In order to facilitate the integration of the HPO in clinical practices, several tools have been developed to make annotation with HPO terms more simple [294–296], including a text annotator tool from the Monarch Initative, which allows to directly annotate sentences from scientific publications with relevant HPO terms [297, 298].

This text annotation tool and the HPO are used in Chapter 5 to develop the first phenotype-driven oligogenic prioritization tool.

3.3 Biological networks

All components of the cellular machinery work by interacting with each other and should therefore not be studied in isolation but rather in the wider context of their associations with specific biological pathways and protein products. The use of networks to represent knowledge and study biological entities is therefore essential and has attracted a wide interest from the biological community in the past decades. In this section, we introduce the concept of networks, showcase different types of biological networks which are used in this work and explore different metrics that can be used to measure distances or similarity profiles within a network.

3.3.1 An introduction to networks

A *network* or *graph* is defined as a collection of *vertices* or *nodes* and *edges*, mathematically defined as $G = (V, E)$, where G represents the network, V the set of all vertices and E the set of all edges. Vertices typically represent different entities (such as genes or proteins) and edges represent relations between them (for example protein-protein interactions).

Networks can be *directed* or *undirected*, depending on whether the edges have a direction that defines a starting node and ending node or not. Networks can also be *weighted* or *unweighted*, depending on whether the edges have an associated weight, i.e. a quantification of the strength of the association.

Networks comprising of nodes of different types, such as a network connecting human genes to human diseases, are called *heterogeneous networks*, while networks for which all nodes are of the same type but with different types of edges, such as a network of genes where edges represent gene-gene interaction and coexpression, are called *multiplex networks*. Finally, a *multiplex-heterogeneous network* contains different node types but also different edge types between the same node types.

A network can be represented by its *adjacency matrix* W, which is a matrix of size $v * v$ where v is the number of nodes in the graph. Each value in the matrix represents the presence/absence of an edge from the "row node" to the "column node" and its weight if the network is weighted. For undirected networks, the adjacency matrix is thus symmetric while it is not necessarily the case for directed networks.

Investigating the topology of a network, i.e. the way the nodes and edges are arranged, is important to characterize the properties of the network. Several metrics can be used to describe the topology of a network, and are described in Table 3.1.

Metric	Description
Node degree	Number of edges that link a node to any other nodes in the network.
Node betweenness	Number of shortest paths between any pair of nodes in the graph that pass through the node.
Node closeness	The minimal sum of all distances in the network for a node.
Node clustering coefficient	The ratio of the number of edges among a node and its neighbors and the maximum possible number of edges among all of them. Basically this denotes the number of node "triangles" that pass through a node.
Edge betweenness	The number of number of shortest paths that pass through an edge connecting node i with node j
Network diameter	The minimum number of links that separate the two most distant nodes in a network.
Shortest path length	The number of links needed to connect every pair of nodes through their shortest path.
Global clustering	Number of closed triplets over total number of triplets

Table 3.1: Metrics that are used to characterize the topology of a network.

Networks can be organized into *modules*, which are groups of nodes that are more connected than other groups of nodes. This was shown to be the case for diseases in biological networks [299]. Algorithms exist to detect such modules or communities [300], with some commonly used ones being the Louvain algorithm [301] and Infomap [302].

3.3.2 Types of biological networks

In biology, networks are useful as they allow to represent different types of interactions which are essential to understand the big picture of cellular organisation. When studying relations between genes and diseases, several types of networks are typically used, and we here present a non-exhaustive list of the most commonly used types, as well as the different sources of data that exist for these networks. Networks exist at the different levels of biological organization, from the genetic interaction all the way to phenotypic similarities [179]:

- **Gene co-expression:** Genes are expressed in particular tissues and in particular time frames. Positively co-expressed genes refer to genes that are expressed in the same tissue and the same time frame, sometimes due to the fact that they are necessary for the same biological processes. On the other hand, genes can also be negatively co-expressed in which case a gene is silenced while the other one is expressed. In a co-expression network, each node represent the transcription product of a gene and each edge reflects the correlation between these products. Understanding gene co-expression can therefore lead to better understanding of particular biological mechanisms and gene function [303, 304]. Co-expression data is obtained from gene expression analysis, where the amount of RNA transcript molecules from different genes is quantified using RNA sequencing or microarrays. This analysis is performed on multiple samples and can be tissue specific or bulk. Gene expressions are then correlated to obtain co-expression data. Databases collecting information on human genes expression and co-expression data include GTEx [305, 306], GeneFriends [307] and CoexpresDB [308] to name a few.

- **Protein-Protein interaction (PPI) networks:** PPI networks collect information on the proteins that physically interact with each other, and are one of the richest types of biological networks. Each node usually represents a protein, and an edge between them represents a physical interaction, which can be associated with a particular weight representing the strength or confidence in the interaction. PPI networks have been shown to be mostly scale-free [309]. Data on PPI can be obtained experimentally (using biophysical methods or high-throughput assays such as Yeast-two-hybrid) or using computational approaches (empirical based on existing data or theroretical based on biophysical properties) [309]. There exist a huge number of PPI databases, with the most commonly used being STRING [310], Intact [311], Mentha [312] and comPPI [313].

- **Pathways networks :** Biological pathways are a series of action that leads to a product or a change in the cell. Components of pathways often interact to result in more complex biological processes, and are therefore better studied using networks. Such networks can be represented by pathway nodes, which are connected to each other. For example,

a node representing DNA replication, would be connected to mitosis, as it is involved in this larger biological process. These pathway nodes can then also be linked to the different genes and proteins which are involved. This can create pathway co-membership networks linking genes that are involved in the same pathways. Examples of sources for pathways information include the widely used REACTOME database [314], the KEGG database [315] and Pathway Commons [316].

- **Gene Ontology (GO) networks:** The Gene Ontology (GO) is a standardized vocabulary to describe the function of genes and gene products. Similarly to the HPO, it can be represented as a DAG where each term stems from one of three main parent terms: Molecular Function Ontology (MFO), which describes the primary activity of gene products at the molecular level (e.g. binding to specific molecules or catalyzing reactions), Biological Process Ontology (BPO), which is used to characterize the biological objective of the gene product (e.g. cell growth or signal transduction), and Cellular Component Ontology (CCO), which specifies the subcellular location, structure or complex where the gene product is active (e.g. cytoplasm or mitochondria). The Gene Ontology DAG is already a graph structure, with nodes being GO terms and edges representing semantic matchings between these terms. However, this graph is usually transformed to study gene relations, by linking gene nodes to the GO terms they are associated with, and linking GO terms together through semantic similarity measures [317]. Many algorithms have been developed to perform this task [318], including GOntoSim [319] and GOGO [320].

There exist many more types of biological networks, such as phenotypic similarity networks, which are based on HPO terms (see Section 3.2.3) that can be linked to the genes they are associated with and to other HPO terms through semantic similarity measures, as well as interaction of genes with drug compounds, sequence similarity networks etc.

3.3.3 Distance and similarity measures

Measuring distances and/or similarities between elements in a network is essential, as it can allow to characterize the links between specific entities and discover potential new interactions between biological entities. We here introduce several important distance and similarity measures which are used in this work and which have been developed for network analysis in general or specifically biological networks.

The biological distance

The biological distance is a measure that was developed specifically for assessing connections between gene pairs in a PPI network. It was established by a team of researcher to generate the Human Gene Connectome in 2013 [321, 322].

The biological distance was initially computed using the STRING database for PPI information, but can be applied to any PPI network with weighted edges.

First, the direct biological distance is defined as the distance between two genes that share a direct connection in STRING, which is the inverted STRING confidence score, and can be summarized by the following equation:

$$D_{i,j} = \frac{1}{S_{i,j}} \tag{3.1}$$

where $S_{i,j}$ represents the strength of the interaction in the STRING database (which ranges between 0 and 1) [310].

To obtain the biological distance between any two genes, the Dijkstra algorithm is used to obtain the shortest path in the network, i.e. the path that connects the two genes and minimizes the sum of the direct biological distances [323]. The biological distance between two genes is then defined as the number of edges in that path (i.e. the degree of separation between the two genes) multiplied by the sum of direct biological distance:

$$B_{i,j} = \begin{cases} D_{i,j}, & \text{if } C = 1 \\ C(D_{i,1} + D_{1,2} + \ldots + D_{C-1,j}), & \text{if } C > 1 \end{cases} \tag{3.2}$$

where $D_{i,1}$ is the direct biological distance (Equation 3.1) between gene i and gene 1 (the first gene on the path between genes i and j, as predicted by the Dijkstra algorithm), gene 2 is the second gene on the path, and gene $C-1$ is the last gene with a predicted direct connection to gene j.

The biological distance was shown to be very useful to identify new candidate genes for a particular disease based on known gene-disease associations [321]. It is also used as a gene pair feature in the VarCoPP predictor (see Section 3.5.2) and an updated version using the most recent version of STRING is also used in this work as a gene pair feature for the VarCoPP2.0 predictor (see Chapter 5).

Jaccard similarity

Jaccard similarity, also known as Jaccard Index, can be used to measure the similarity and diversity within sample sets [324]. It is defined as :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.3}$$

where A and B are two different sets.

Although it was developed to measure similarity between sets, it is also widely used in network analysis to define the similarity between nodes based on the number of neighbours they share for example [325, 326]. In Chapter 5, we compare the use of this metric with more complex algorithms for measuring similarity between nodes in a network, in order to assess which one is more efficient for ranking gene pairs based on their disease relevance.

Propagation algorithms and the Random-Walk-with-Restart (RWR) algorithms

Network propagation algorithms are methods used to spread or transfer information, through the nodes and edges of a network, by iteratively updating node values based on their neighbours' values (Figure 3.3) [327]. They have been shown to be very useful to predict protein function [328, 329] and detect novel disease-gene associations using biological networks [174, 330, 331].

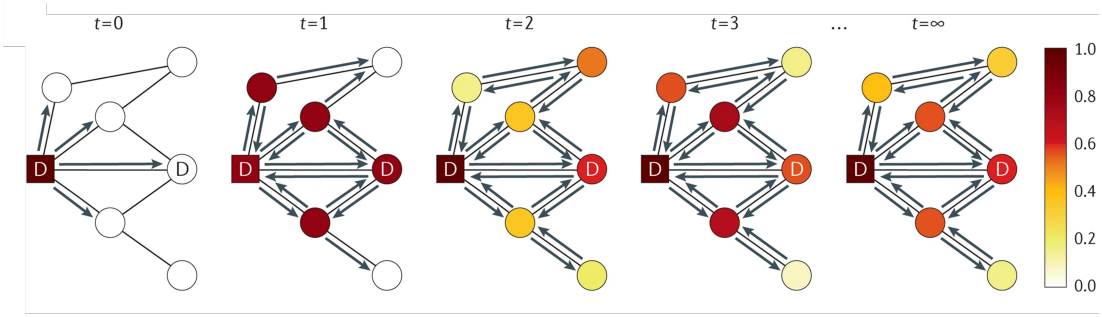


Figure 3.3: Step-by-step illustration of network propagation. The process is shown at different time points until convergence ($t = \infty$). The arrows depict the direction of the propagation at each step, and the node are coloured according to the amount of information they receive. The square node represents the "seed" node, and the 'D' label represents the known or predicted nodes associated with a disease phenotype. Figure from [327]

Random-Walk-with-Restart (RWR) is a type of network propagation algorithm where the information from a set of user-defined seed nodes is disseminated through a network, and which outputs a score for each node in the network, which can be interpreted as a measure of proximity to the set of seeds (Figure 3.3). This algorithm is known to take into account both local and global network topology when performing the propagation [327]. It was first

applied to disease-gene prioritization on a PPI network in 2008 [174]. The main intuition behind applying an RWR algorithm for disease-gene prioritization is the "guilt-by-association" principle, which is based on the fact that genes that are involved in the same diseases are usually close in biological networks.

The RWR is formally defined as :

$$p^{t+1} = (1-r)Wp^t + rp^0 \tag{3.4}$$

where W is the column-normalized adjacency matrix of the graph and p^t is a vector in which the i-th element represents the probability of being at node i at time t. The parameter r is the restart probability, which is set between 0 and 1 and controls the "spread" of the walk on the network. If the restart parameter is high, the walker will often go back to the seed nodes, and only nodes close to the seeds will be traversed, while if the restart parameter is closer to 0, the walker will be able to reach nodes further from the initial seeds. The p^0 vector represents the initial starting point, and is here constructed by setting equal probabilities to all seed nodes, with the probabilities summing to 1, which is equivalent to having the random walker start randomly from any of the seed nodes.

The "classic" RWR algorithm has been further expanded to compute random walks on heterogeneous graphs, multiplex graphs and multiplex-heterogeneous graphs (see Section 3.3.1 for a description of these graph types) [332, 333]. These extensions are based on considering each layer of the graphs independently and allow for the walker to "jump" from one layer to the other, with the goal of keeping the topology of each individual network. They use the same equation as for the classic RWR, but the W transition matrix is modified to reflect the heterogeneous and/or multiplex nature of the graph. These extensions are said to better integrate the intrinsic properties of the different layers of heterogeneous and multiplex networks.

The main parameter to set for all types of RWR algorithms is the restart parameter r in Equation 3.4, and controls the ratio of exploration vs exploitation of the prior knowledge (i.e. the set of seeds). The restart value of 0.7 is often shown to present a good exploration vs exploitation ratio for biological networks [169, 174, 332, 333], but we still explored how this parameter affected our results when using this algorithm in Chapter 5. For the RWR-heterogeneous and RWR-multiplex-heterogeneous algorithms, additional parameters need to be tuned, which represent the probability of jumping between layers, as well as the probability of restarting in any of the multiplex layers (when a multiplex layer is at hand). These parameters were not shown to have a strong effect on the results of the algorithms and the probability of jumping between the different heterogeneous and/or multiplex layers is therefore usually set to 0.5, a value which we adopted for this thesis.

In Chapter 5 of this thesis, we use a RWR approach to score the disease-relevance of gene pairs. We test different seed types and algorithm types to assess which method has the best performance for ranking gene pairs based on disease relevance.

3.4 Machine Learning concepts

In this section, we explain important concepts in the field of Machine Learning (ML), which are necessary to understand part of this thesis' work. We first introduce what is ML and the ML procedure, and then go into the details of two main types of algorithms, and their application to the variant pathogenicity prediction problem. We then highlight the importance of feature engineering and selection for developing a ML predictor, and explain how these features can be used to understand how the model functions and the predictions it generates. Finally, we explain how to evaluate the performance of a ML model, depending on the problem at hand and the available data.

3.4.1 The machine learning procedure

ML refers to the study and development of algorithms that can learn from data. ML algorithms are mainly divided into 4 categories: supervised learning, semi-supervised learning, unsupervised learning and reinforcement learning [151]. *Supervised learning* implies training a model based on a labeled dataset, so that the trained algorithm can predict as accurately as possible the label based on a set of characteristics or features which are provided in the dataset. *Unsupervised learning* refers to the detection of patterns from unlabelled data, where the algorithm's goal is to try to determine the underlying structure of the data, by performing clustering for example or dimensionality reduction. *Semi-supervised learning* can be defined as a hybrid between the aforementioned approaches: the aim is to predict the label of the sample, but the algorithms can make use of both labelled and unlabelled data. Finally, *reinforcement learning* is an approach where learning happens based on the environment instead of a pre-defined training set. Software agents or machines learn by performing actions in the environment and evaluate the optimal behavior based on rewards and penalties. This type of ML is often used in decision making scenarios.

In this work, we use supervised learning algorithms. The supervised learning procedure is typically divided in two stages: the learning phase and the prediction phase. This process is usually preceded by a preliminary preparation phase, which includes problem definition (and thus type of learning), data collection and pre-processing. Although this stage is not always included as part of the ML procedure, ensuring a high-quality dataset is essential in order to obtain reliable predictions. In this pre-processing step, the dataset is collected and

annotated with different features, which will be used by the model. Features can be numerical or categorical, and are the attributes of the data that will be used by the model to make predictions. The choice of which features to include in a dataset is therefore essential and is described in more details in Section 3.4.3.

During the learning phase, one or several models are trained and evaluated to find the optimal parameter values such that the model can generate the most accurate predictions. This phase implies splitting the dataset into a training and testing set, with the training set being used to optimize the model parameters and the testing set being used to evaluate its performance.

Finally, during the testing and prediction phase, the optimized model from the learning phase is used to predict an external validation set, for which the true labels are known in order to evaluate its performance on external data. The model can then be used to generate predictions on new data.

3.4.2 Types of supervised learning algorithms

Different types of algorithms can be used depending on the problem the algorithm is trying to solve. In supervised learning, we distinguish two main problems: the regression problem, where the variable to predict or estimate is a continuous variable, and the classification problem, where the variable to predict or estimate is a categorical variable. For the work presented in this thesis, we will focus on algorithms that have been designed to solve the classification problem and will additionally describe algorithms aiming to solve the ranking problem, i.e. ordering items in a set based on their relevance to a particular query.

Classification algorithms

In a classification problem setting, the goal is to estimate the class or category an item belongs to, based on a series of values called *features* (see Section 3.4.3). In the context of this thesis, the classification problem at hand is to determine whether a particular genetic instance (e.g. a variant, a gene or a combination of variants) is pathogenic or benign. Such problem is known as *binary classification*, because it entails assigning the instance of interest to one of two classes. In supervised ML, the classification problem is approached by training an algorithm on a training set, which includes a certain number of instances for which the output class is known, and then using the trained algorithm to determine the class of new instances for which the class is unknown.

Several categories of algorithms have been defined to perform such a task and each time, attempt during the training phase, to determine for which values and combinations of feature values an instance should be classified as one or the other class. In Chapter 5, we use a Random Forest (RF) algorithm to develop a predictive model for the classification of variant combination in gene pairs, and thus describe the functioning of this particular algorithm in this section.

RFs are based on decision trees, which are aggregated in order to improve accuracy and avoid overfitting of the data. A decision tree model has a hierarchical structure starting from a *root* node, which is connected through edges to *child* nodes (Figure 3.4a). Each node in the graph represents a rule or test, that is applied during the classification process and orients the decision process to the next node based on the test outcome. This process is repeated until the *leaves* are reached, i.e. the nodes that are not linked to any child nodes and are associated with the class label which aims to be predicted. Each rule is based on a condition set on a particular feature of the dataset. A decision tree can therefore be seen as a series of questions breaking down the decision-making process [334].

A decision tree is trained starting with all training instances at the root node. It is then grown by selecting for each node, the feature and threshold that best splits the data, and setting the rule of the node to that condition. The training data is then divided based on that rule and the process is iterated for each resulting child node, until achieving one of three conditions: (i) a terminal node is reached where the data cannot be split anymore (i.e. all instances in that node belong to the same class), (ii) further splits do not improve predictions, or (iii) a maximum preset condition is reached (e.g. reaching maximum preset depth).

Two impurity metrics are generally used to identify the feature that best splits the data at each step in the tree growing process: the Gini impurity or the entropy. The Gini impurity is defined as :

$$Gini(t) = 1 - p(t)^2 - (1 - p(t))^2 \tag{3.5}$$

Where $p(t)$ is the proportion of instances belonging to the positive class at node t (and $1 - p(t)$ is thus the proportion of instances belonging to the negative class).

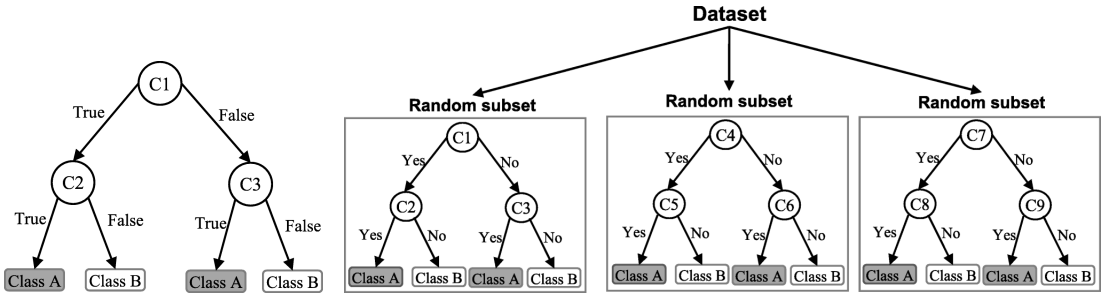
The entropy is defined as :

$$H(t) = -p(t)log(p(t)) - (1 - p(t))log((1 - p(t))) \tag{3.6}$$

Where $p(t)$ is the proportion of instances belonging to the positive class at node t (and $1 - p(t)$ is thus the proportion of instances belonging to the negative class).

In both cases, the aim is to choose the feature that performs the split that will minimize the weighted average of the impurity of the resulting child nodes.

Decision trees are used in biological applications and have the advantage of being able to model non-linear interactions between features as well as being easily interpretable, since the list of rules associated with each node can be easily retrieved. However, they are prone to overfitting and can struggle to model high-dimensional data.



(a) Decision tree binary classifier

(b) Random forest binary classifier

Figure 3.4: Schematic representation of a decision tree binary classifier (a) and a random forest binary classifier (b). Figure from [335]

RF overcome these limitations by aggregating multiple decision trees into a single model (Figure 3.4b). Each tree in a RF is trained on a subset of the training data that is selected randomly and independently (i.e. through *bootstrapping*), and can thus consider different features at each split. When predicting a novel instance, each decision tree outputs a class independently and the output of the RF is the majority vote over all the trees. By this ensembling approach, RFs have been shown to be more robust and generalisable than single decision trees. They have been extensively used in bioinformatics and clinical applications, in particular because they are able to handle class-imbalanced problems.

Prioritization algorithms

As discussed in section 1.3.4, in variant pathogenicity prediction, prioritization tools refer to tools for which the aim is to rank genes or variants according to their relevance to a particular disease or phenotype.

The problem of ranking instances, i.e. sorting elements in a list according to their relevance to a particular query, can also be approached using ML algorithms. This learning problem is known in the literature as “preference learning” or “learning to rank” and has been the focus of several reviews [336–338]. Three types of ranking tasks are usually distinguished: *label ranking*, which deals with the ordering of labels to assign to a particular instance, *instance ranking*, which involves ordering a set of instances based on their relevance to a particular query, and *object ranking*, where objects are ordered based on their properties and not necessarily in relevance to a particular query [336]. Instance and object ranking are sometimes used interchangeably, although object ranking is normally more general as it does not involve ranking linked to specific queries. Examples of applications of ranking algorithms include search engines, recommender systems and ranking genes for relevance to disease [336, 339].

In the case of instance ranking, which includes the variant prioritization task, the problem can be formalized as follows: given a query q and a set of n instances $(D_1, ..., D_n)$, we want to fit a model that will output a score s_i to each instance D_i based on its relevance to the query. We assume there exist a true relevance score y_i which we aim to estimate. Different algorithms have been proposed and can be further categorized in three approaches based on the choice of loss function for the learning. In the *pointwise approach*, each single instance is viewed independently, and the model assigns the instance a score which is an estimate of the true relevance score. The loss function is the sum of the distance between each prediction and the true relevance score, transforming the ranking problem into a regression problem. The main issue with this approach is that the exact relevance for each instance in a ranking is not always known, as it is the case for the gene prioritization approach, where some genes are known to be relevant to a disease but assigning a relevance measure to other genes is not straightforward. For this reason, the *pairwise approach* uses the idea of relative preference, and the training happens by comparing the relevance of each two instances in the set to perform the ranking. This transforms the ranking problem in a binary classification problem. The main limitation of this approach is that pairwise ranking is not sensitive to the fact that relevant instances should be found at the top of the list (see Figure 3.5). Using metrics that integrate the rank of the relevant instances in the loss function is difficult, since these metrics are non-differentiable.

Figure 3.5: Illustration of the limitation of the pairwise approach for instance ranking. Panels A and B show rankings of 2 relevant and 6 irrelevant instances with the same number of incorrect pairwise rankings (6). In the ranking task, having relevant elements at the top of the ranking as in panel A is usually favored, showing that the pairwise approach might not always yield the best ranking.

Finally, the *listwise approach* aims to address the aforementioned limitations by directly optimizing the loss function based on the ranking of the complete list. In order to overcome the fact that sorting is a non-differentiable operation, different strategies have been developed such as smoothening the ranks [340], using iterative methods [341, 342] or computing a loss based on permutation [343].

Although learning to rank algorithms have received significant attention in the ML community, they have not yet reached many applications in the bioinformatics field [339]. In particular, while there have been uses of learning to rank algorithms for the prioritization of genes relevant to diseases [344, 345], there are, to our knowledge, no variant prioritization models using learning to rank approaches.

In practice, the variant prioritization problem is therefore more often treated as a supervised classification problem, although the evaluation of the performance differs slightly, as discussed in Section 3.4.4. We here provide a brief overview of some variant prioritization algorithms to showcase a few of the main approaches used specifically for variant prioritization. For more information, these predictors have been the subject of several review articles [125, 160, 346–349].

The Exomiser suite of algorithms prioritize variants by integrating a variant score together with a disease-relevance score [169]. The disease relevance score is obtained through either a random-walk approach [350], a phenotype comparison approach using human data only [69], a phenotype comparison approach integrating data from other model organisms [170] or an integration of all three [292]. The variant score is based on the MAF and the maximum predicted pathogenicity score from a series of predictors (MutationTaster, Polyphen2 and SIFT in the original publication [69, 170]). The disease-relevance score depends on the method selected. In the case of the ExomeWalker approach, it is computed by RWR proximity to a set of seed genes. In the case of the Phenix approach, it is computed using semantic similarity between the HPO terms of the patients and the HPO terms of known disease genes. Finally, in the case of the PHIVE (PHenotypic Interpretation of Variants in Exomes) approach, the disease-relevance score is computed as the semantic similarity between the HPO terms of the patient and the HPO and MPO terms associated to specific genes, while a mix of these three scoring methods is used in the hiPHIVE approach. The pathogenicity and disease-relevance scores are then combined through a logistic regression to generate a final ranking. This logistic regression model was trained on an equal number of known pathogenic mutations from the HGMD and neutral variants from the 1KGP.

AMELIE (Automatic Mendelian Literature Evaluation) is a recent tool which aims to provide clinicians with both a ranking of the gene most likely to cause the disease of a particular patient and references to articles providing additional information on that gene [184]. The algorithm relies on a knowledgebase generated by Natural Language Processing (NLP) techniques trained on curated databases. A classifier is trained on 27 features including variant pathogenicity prediction, phenotype matching between article terms and HPO terms provided with the patient and various gene and article relevance scores. The training data includes a set of simulated patients using data from OMIM, and a set of randomized patients as negative instances [184].

Finally, the PVP and DeepPVP algorithms are trained using positive and negative data from the ClinVar database, and rank variants based on variant features, disease mode of inheritance and semantic similarity matching between the patient's HPO terms and the HPO terms associated with each gene [175, 176]. A tool based on this framework was also developed to prioritize oligogenic variants, and is described in Section 3.5.4.

The aforementioned predictors are thus trained as classifiers. However, their aim is to identify the disease-causing variant for each individual patient, using information from the patient's disease. The difference between these tools and other pathogenicity predictors such as CADD (see Section 1.3.3) is that they are evaluated differently, by testing the ability of the tool to find the disease-causing variant of a patient in the top prioritized instances. This led us to the following definition of the difference between pathogenicity predictors and prioritization tools.

In this thesis, the main difference between what we consider as *classification algorithms for pathogenicity*, and *variant prioritization methods* is therefore not in the training of the algorithm but rather lies in the scope of the method and in the evaluation of its performance. We refer to algorithms which aim to predict whether a specific variant (or variant combination) is likely to be disease-causing as a classification tool. On the other hand, tools that aim at identifying among all variants in a patient's exome which variant is the most likely to cause the patient's disease are addressed as prioritization methods. The differences in performance evaluation are detailed in Section 3.4.4.

3.4.3 Feature selection and interpretation

In addition to selecting the instances used in the training set and choosing a model structure, identifying which features are relevant to the problem at hand is one of the key steps to building a ML predictor. We here discuss two important steps of feature selection: the manual identification of relevant characteristics for a particular problem, and feature reduction algorithms, a computational approach to selecting the computationally relevant features from the set of theoretically relevant ones. Finally, we also introduce feature interpretation methods as a mean to understand how a particular model makes predictions.

Manual feature selection and feature engineering

Manual feature selection refers to searching for the type of information that can be used to characterize and annotate the instances in a dataset before presenting it to a ML model. This process is highly dependent on the problem at hand and the dataset, but several general considerations can be taken into account, especially for applications in the biomedical field where model interpretation is important.

First, features should be selected based on their relevance to the problem at hand. This allows not only to develop a more accurate predictor but also to be able to obtain biological insights by interpreting feature importances (see Section 3.4.3 below). Assessing the proportion of instances which cannot be annotated by a particular feature (e.g. certain variant pathogenicity scores can only annotate specific types of variants) is also essential, since using such features will thus lead to many missing values, which can introduce important biases in the predictions [351]. The maintenance of the source of a feature as well as its updates' frequency can also be of interest for the maintenance of the predictor. This is especially important in bioinformatics where data is accumulating fast and sources of information are being regularly updated. For example, many scores have been developed to assess gene's tolerance to mutations. These scores are obtained from large databases of variant frequencies, which have undergone significant changes over the past few years. A score measured over old versions of GnomAD and not updated with the latest releases is not likely to provide valuable insights today.

Finally, feature engineering is important to consider in some cases, and refers to the application of transformations to datasets in order to create machine readable characteristics. For example, network properties of genes or variants can be transformed into features, by computing several relevant topology or distance metrics (see Table 3.1 in Section 3.3).

These considerations were taken into account when designing the initial feature set for the VarCoPP2.0 predictor in Chapter 5, and are further discussed in this context.

Feature reduction algorithms

Careful manual feature selection does not necessarily yield the optimal set of features for the algorithm to learn. Computational feature selection methods, or feature reduction algorithms, have thus also been developed. These algorithms aim to identify the most relevant set of features from the original feature set, in order to improve the predictor's performance and reduce overfitting. Feature reduction methods are traditionally divided in three categories: filter approaches, wrapper approaches and embedded approaches [352].

Filter approaches rely on the intrinsic properties of the data (e.g. the distribution of the data), score features based on their relevance and output a subset to be presented to the model. The advantage of such methods is that they are fast and can thus be applied to large datasets, as well as independent from the model structure, which means that different model types can then be assessed. This can also be a disadvantage, since the selection method will not generate a feature set that is optimized for the model at hand.

Wrapper approaches, on the other hand, include the model in the feature search. In this setup, various feature sets are evaluated together with the model structure, by training and testing the model with each feature set. The search for the optimal feature sets is thus "wrapped" around the model training. Since the set of all possible subsets of features to evaluate is usually

too large, heuristic search methods are used to explore the feature sets. These methods can be divided in two categories: deterministic search and randomized search. The key advantage of wrapper approaches is that the optimal set of features identified is therefore tailored to a specific model structure, and that the search takes into account not only features dependencies, but also interactions between the features and the model. A big drawback of such approaches is that they are computationally heavy. In Chapter 5, we use a wrapper approach to select the optimal feature set from a set of manually curated features, to train a RF algorithm.

Finally, *embedded approaches* refer to methods where the feature selection is part of the model. These are usually faster than wrapper approaches, but only exist for specific models.

Understanding feature importances and feature interpretation methods

In order to better understand a ML predictor and the process behind its predictions, *feature interpretation* methods can be used. These methods have two main goals: the first one is to uncover general data patterns that have been identified by the model, which can lead to new biological insights; and the second one is to be able to understand decisions made by the model, thus explaining how a prediction was made.

To gain general insights about the model, *feature importance* measures can be used. These measures typically rely on permutations of the features in the model which allow to identify which features are more essential. These measures are typically model dependent, although they can have similarities [353]. We here discuss in more details feature importances for RF algorithms which are used in Chapter 5 of this thesis.

In RF algorithms, decision trees are grown by splitting the data based on values of specific features (see Section 3.4.2). Measuring feature importance can therefore be done directly based on these splits. A naive measure of feature importance would therefore be to count the number of times a feature was used in a split over all trees present in the forest. However, a split at the root of a tree and a split at the end of a tree do not have the same value, which is why more elaborate measures have been developed.

The main measure used is the *Mean Decrease in Impurity (MDI)*, also called *Gini importance*. Since the trees of a random forest are grown by selecting the splits that generate the maximum decrease in impurity (see Section 3.4.2), the features that are used to generate splits with larger decrease in impurity are therefore considered more important. The MDI measure reflects this idea, and computes for each feature X_i, the sum of all impurity decrease measures of all nodes in the forest at which a split on X_i has been conducted, normalized by the number of trees. For classification, the impurity is usually measured by the Gini impurity defined in Equation 3.5 [354].

To better understand a particular prediction, *"post-hoc" feature interpretation* algorithms can be used. Unlike feature importance measures, these methods compute, for each predicted instance separately, the contribution of each feature to the attributed class. They provide detailed information on the link between the extent and the type of influence (i.e. positive or negative) a variable had on the output prediction. In the case of decision trees and RF models, the *treeinterpreter*[1] algorithm can be used, which transforms the decision path in each tree as a sum of feature contributions.

It is important to note that these methods are more relevant when the selected features can have a biological interpretation, and can thus provide novel insights into the biological mechanisms underlying the process that one is trying to predict.

3.4.4 Performance evaluation

Finally, an important aspect to consider when developing predictive methods for classification or ranking, is how to accurately evaluate their performance. Different strategies can be used based on the scope of the methods, but also on the type and amount of data available. The idea of a good performance evaluation approach is to measure the performance of the predictive tool in a setting that is similar to "real- world" usage of the tool.

Classification problems

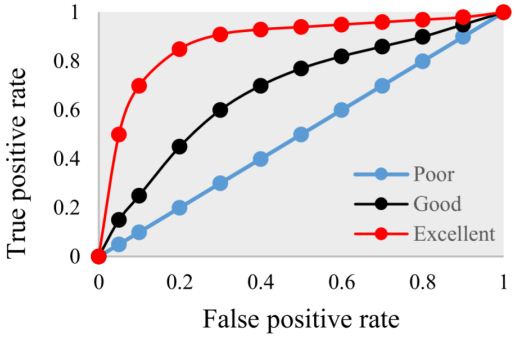
Cross-validation is a procedure that allows to evaluate the performance of a model using only the training data. The training set is split into a certain number of different groups, called *folds*. The instances belonging to a group are removed from the training set, the model is trained on the remaining instances and its performance is assessed on the "left out" group that was not used for training using different performance metrics (see below). This process is then repeated for all the pre-defined folds. The final performance is then measured as either the average of the metrics in the different folds (micro-average) or by aggregating the predictions of the model in each fold and computing the performance metrics on this full prediction vector (macro-average). However, performance evaluation using only the cross-validation procedure can introduce bias since the predictor is then trained on all instances in the training set [355]. Ideally, the use of an independent test set, which is never used in the training part, is always important in order to have an accurate evaluation of the performance of the model.

For classification tools, the main performance metrics are based on the confusion matrix (Figure 3.6a), which is computed by counting the number of instances which are correctly classified in both positive and negative classes (True Positive (TP)s and True Negative (TN)s) and the number of instances which are misclassified in both classes (False Positive (FP)s and False Negative (FN)s). The most common metrics are summarized in Figure 3.6c.

1. https://github.com/andosa/treeinterpreter

	Actually positive	Actually negative
Predicted positive	True Positive (TP)	False positive (FP)
Predicted negative	False Negative (FN)	True Negative (TN)

(a) Confusion matrix



(b) Comparing ROC curves

Metric	Formula	Description
Error rate	$\frac{FN + FP}{TP + FP + FN + TN}$	Proportion of mis-classified instances
Accuracy	$\frac{TN + TP}{TP + FP + FN + TN}$	Proportion of correctly predicted instances
True Positive Rate (TPR)	$\frac{TP}{TP + FN}$	Proportion of actual positives that are correctly classified. Also called sensitivity or recall.
False Positive Rate (FPR)	$\frac{FP}{FP + TN}$	Proportion of actual negatives that are mis-classified
Specificity	$\frac{TN}{TN + FP}$	Proportion of actual negatives that are correctly classified. Also called true negative rate
Precision	$\frac{TP}{TP + FP}$	Proportion of true positives out of all predicted positives. Also called positive predictive value
F1-Score	$\frac{2\,TP}{2\,TP + FP + FN}$	Harmonic mean of precision and recall.
Matthews Correlation Coefficient (MCC)	$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	Metric that returns a value between -1 and 1, with a high value (close to 1) meaning that all classes are predicted well while a value close to 0 means that the classifier is random.

(c) Performance metrics for classification

Figure 3.6: Measures used to evaluate the performance of a two-class classifier. (a) The confusion matrix is the basis of the different measures used and contains counts of the number of instances according to the class they belong to and the class label they are attributed by the predictor. (b) Receiver operating characteristic curves plot the true positive rate as a function of the false positive rate for different probability thresholds and give an idea of the predictive power of a classifier. A straight line indicates that the classifier essentially performs as random class assignment. (Figure from [356]) (c) Description and formulas of commonly used performance metrics based on the confusion matrix.

For class imbalanced problems, which are common in biomedical problems [357, 358], and where the number of instances in one of the two classes is much larger than in the other class, some of the most commonly used metrics such as precision and recall can become biased and thus be strongly misleading [359, 360]. In this case, Receiver-Operating characteristic (ROC) curve analysis [361] and Precision-Recall (PR) curve analysis is generally favoured [362]. ROC analysis is done by plotting the sensitivity (or True positive rate) and specificity

of the predictions for various threshold values for the different classes. The threshold value is varied between 0 and 1 and represents the value for which any instance with a probability above the threshold is predicted as belonging to the positive class and any instance with a prediction probability below the threshold is predicted as belonging to the negative class. The area under the curve (AUC) then measures the probability that the model correctly predicts an instance out of a randomly chosen pair of case and control samples. The AUC ranges from 0.5 (for a completely random classifier) to 1 (for a perfect classifier with 100 % accuracy, see Figure 3.6b). On the other hand, the PR curve plots the precision as a function of the recall (or sensitivity or TPR) for different threshold values. The AUC under the PR curve can also be computed and is called the Average Precision Score (APS).

The definition of which class is considered as positive or negative is arbitrary but in the field of pathogenicity prediction, the disease-causing class is usually considered as positive (i.e. the one we want to detect) and the neutral class as negative.

Prioritization tools

We here focus on methods used to assess the performance of variant prioritization tools. Prioritization tools typically output a ranking of the pathogenic variants, with the most likely disease-causing variants ranked first, and where a single variant is expected to be relevant. In order to assess the performance of such tools, they have to be tested on data from a patient's exome. However, while large databases of pathogenic variants are available (e.g. ClinVar [83], HGMD [363]), the exome data of the individuals carrying these pathogenic variants is often not available, which means that no real test data can be used.

In practice, the performance of prioritization tools is thus evaluated in synthetic exomes. Exomes from healthy patients are obtained from in-house projects or large databases of genetic variation (e.g. 1000 Genomes Project (1KGP)), and known pathogenic variants are inserted into these exomes to be detected by the prioritization method. For phenotype-driven tools which use HPO terms as additional input, the HPO terms associated with the disease caused by the pathogenic variant of interest are used. In order to simulate real clinical data, these HPO terms are sampled at random (up to a maximum of 5 to 10 terms). To test the robustness of the method to phenotype annotations, some authors have tried to add random HPOs (to simulate errors in phenotype description) and replace some terms by their parents in the DAG (to simulate imprecision in phenotype description) [185]. More recently, the performance of these tools has also been assessed in large cohorts of patients, which are well annotated, as this is becoming the new standard in the field [125, 364].

The performance of the prioritizers is then measured as the percentage of synthetic exomes where the known pathogenic variants are found in the top k ranked variants for different values of k. A Cumulative Density Function (CDF) is often computed and plotted, to visualize the percentage of cases with causal variants ranked within the top k (varied between 1 and

a maximum value) by each method (see Figure 3.7). This cumulative display is better for the illustration of results generated under continuously changing conditions than the recall @K measure (which measures at a specific rank K how many relevant instances have been found).

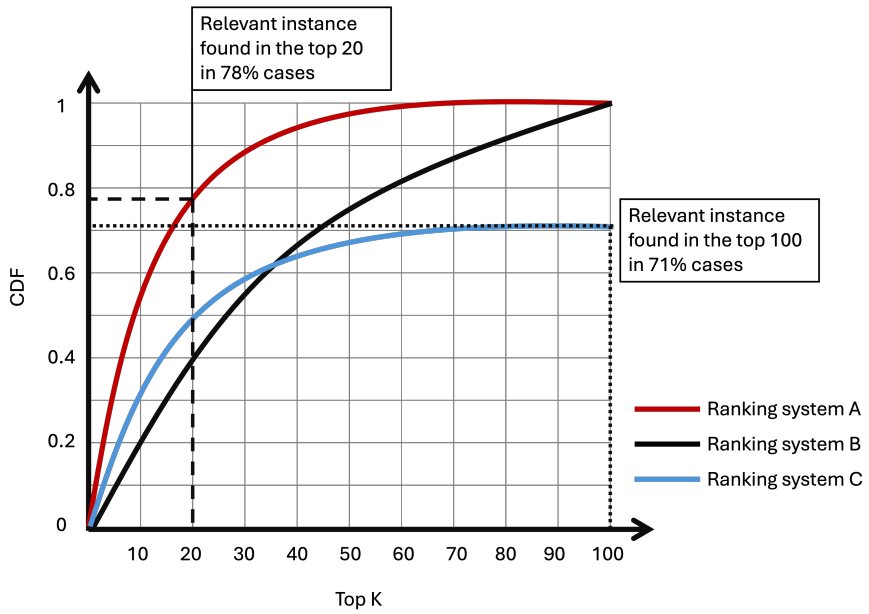


Figure 3.7: Illustration of CDF curves for 3 ranking systems and how to read them. The best performing system (system A) retrieves the relevant instance in all cases in the top 70. Ranking system B performs better than system C when looking at the top 100, but performs slightly worse when focusing on the top 20. The dotted lines illustrate how to read the plot to obtain the proportion of cases for which the relevant instance(s) are retrieved in the top K.

In addition to the CDF curve, other metrics have been developed for evaluating the performance of a ranking algorithm [365]. We here provide a brief overview of the main metrics, which we will use in Chapter 5, to compare how different measures of performance affect the evaluation of our developed methods. The *Mean Reciprocal Rank (MRR)* computes the average of the reciprocal rank (i.e. $1/rank$) of the first relevant item across all ranked lists of items. This measure is usually assessed in the top K elements, and the MRR therefore ranges from 0 (if no relevant element is found in this top K) to 1 (in the case where the top element of each ranking is relevant). The main limitation of this measure is that it does not take into account the rank of the other elements, but it is suitable for scenarios where a single instance is relevant, and is easy to interpret.

The *average precision* is computed as the average of the precision values for each relevant item in the list. The *precision* in a ranking is measured as the proportion of relevant items in the top K. Base on the average precision, the *Mean Average Precision (MAP)* can be computed to measure the performance of a tool by taking the mean of the average precision in each

ranking produced by that tool. This metric therefore integrates the position of the retrieved items as well as the precision and the recall, but is less intuitive to compute and more difficult to interpret. For problems where there is only one relevant item per list, this measure simplifies to the MRR.

The *normalized Discounted Cumulative Gain (nDCG)*, computed from the *Discounted Cumulative Gain (DCG)* measure, is based on the assumption that highly relevant instances are more useful when appearing at the top of the list. To compute the *DCG*, each item is assigned a gain based on its particular relevance y_i. A popular gain measure is defined as $G_i = 2^{y_i} - 1$, since the exponent puts a lot of emphasis on highly relevant items. Based on its rank i, each item is also assigned a discount, which acts as penalty to the gain. This discount is often computed as $D_i = log_2(i+1)$. The DCG is then computed as:

$$DCG = \sum_{i=1}^{k} \frac{G_i}{D_i}$$

This measure is usually normalized based on the highest possible DCG, which would be obtained when all items are ranked correctly, in order to obtain the *nDCG*. We note that for ranking problems where there is a single relevant instance in each list (such as the variant prioritization problem), this measure is simplified to :

$$DCG = \sum_{i=1}^{k} \frac{1}{log_2(i+1)}$$

The choice of metrics and evaluation set is therefore essential to accurately assess the performance of a tool in a realistic scenario. In Chapter 5, we assess different methods to evaluate the performance of our oligogenic prioritization tool, using both the prioritization and classification metrics described in this section.

3.5 Available oligogenic resources

As introduced in Section 1.3.5, several databases, tools and resources had already been developed before the work introduced in this thesis. In this part, we discuss these resources and their limitations in more detail, highlighting how these were used as the building blocks of the novel methods and resources developed in this thesis.

3.5.1 DIDA: a first repository of data on digenic diseases

The Digenic Diseases Database, published in 2015, was the first repository of data regarding digenic diseases. The authors were inspired by the review paper from Schäffer on digenic inheritance [96], and collected from the scientific literature information on digenic variant combinations, i.e. combinations of variants in 2 genes involved in a digenic disease. Information on variants, genes, variant combinations and the diseases they underlie was obtained by manual curation of articles. DIDA was updated in the years following its publication and, at the time of writing, contains information on 258 variant combinations, linked to 52 distinct diseases. Bardet-Biedl syndrome, familial haemophagocytic lymphohistiocytosis and familial long QT syndrome are the most represented disorders in the database. Three main types of digenic models are typically found in DIDA, corresponding to the main types of digenic inheritance observed (see Figure 1.4 in Section 1.2.3): True Digenic models, Monogenic + Modifier combinations and Dual-Molecular Diagnosis (DMD) cases.

DIDA has been used to train several machine learning predictors which aim to predict and help understand the cause of digenic diseases. The Variant Combination Pathogenicity Predictor (VarCoPP) was the first model to classify variant combinations as either pathogenic or neutral ([105] and see Section 3.5.2), while the Digenic Effect (DE) predictor was trained to initially distinguish between True Digenic and Monogenic + modifier cases [366] and was later extended to also identify DMD cases [187]. A first oligogenic prioritization tool ([194] and see Section 3.5.4) was also benchmarked on the DIDA dataset, and a predictor classifying gene pairs as likely to be involved in a digenic disease or not was also recently published [167].

These pathogenicity prediction tools in turn helped find novel oligogenic causes to diseases [189, 192, 367, 368], highlighting the need for continuous collection of data in this area. Nevertheless, DIDA is not perfectly suited for that purpose. Indeed, the database's architecture was limited to variant combinations in gene pairs (digenic), excluding all other types of oligogenic combinations. Moreover, certain types of variants, such as Copy-number variant (CNV)s were not collected in the database. Furthermore, although DIDA gathers information on whether bi-locus combinations are supported by familial and functional evidence, the criteria for including any particular combination was not clearly defined. With the advent of guidelines for reporting disease causality of genetic variants [121], this type of evidence should be clarified and more precisely assessed.

In this work, we present a complete overhaul of the DIDA database, which resolves the aforementioned limitations, and significantly expand the content of the database (Chapter 4).

3.5.2 VarCoPP, a first predictor of variant combination pathogenicity

The Variant Combination Pathogenicity Predictor (VarCoPP) is a ML model which predicts the probability that a digenic variant combination is pathogenic. It was trained using instances from the first version of the DIDA database as positive instances, and variant combinations obtained from individuals from the 1KGP as neutral instances. VarCoPP was trained as an ensemble model, in order to tackle the class imbalance problem that arose from the training data. The final predictor consisted of 500 RFs, which were trained on 500 different training sets, and which predictions were aggregated in a majority vote (Figure 3.8).

Each random forest learns from a set of 11 features which were selected from an original set of 21 features using recursive feature elimination, collected from different biological levels:

- At the variant level, the flexibility difference, hydrophobicity difference and CADD scores of pathogencity are used. The flexibility and hydrophobicity differences are computed as the change in flexibility and hydrophobicity between the wild-type and mutated amino acid, based on flexibility and hydrophobicity scales from [370] and [371] respectively. CADD is a pathogenicity predictor for single variants (see Section 1.3.3 and [161] for more details).
- At the gene level, both genes are annotated with haploinsufficiency and recessiveness probability scores. The haploinsufficiency probability is the probability that a single copy of a gene is insufficient to maintain normal function; it is computed by a predictive model trained on different properties of known haploinsufficient and haplosufficient genes [372]. The recessiveness probability is the probability that a given gene will cause a disease phenotype following its homozygous inactivation; it is here again computed by a predictive model trained on the different properties of known recessive disease genes and loss-of-function tolerant genes [373].
- At the gene pair level, the biological distance between the two genes, which measures how far genes are in a PPI interaction network is used (see Section 3.3.3).

VarCoPP was shown to perform very well in both cross-validation and independent validation settings. It achieves a TP rate of 0.88 and FP rate of 0.11 in cross-validation, and correctly predicts 20 of the 23 pathogenic variant combinations in the independent validation set.

Although the performance of VarCoPP is promising, it still presents with some limitations. The model has a very complex structure, which not only prevents the expansion of the training set, but also makes is computationally heavy, leading to increased prediction time, especially for larger datasets such as exome sequencing data. Additionally, the model was trained on the

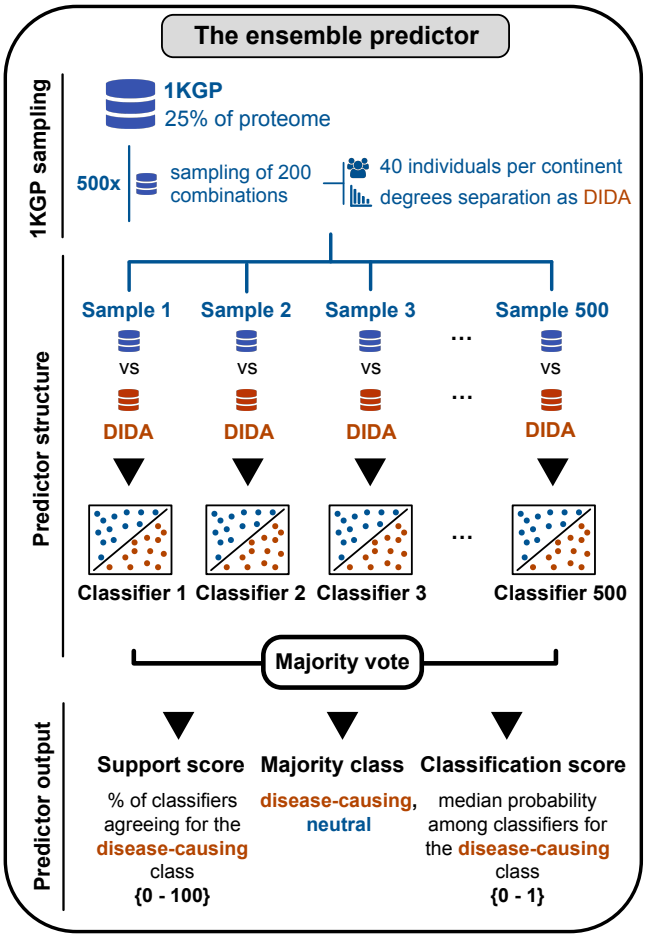


Figure 3.8: To solve the class imbalance problem inside the training set of VarCoPP, 500 random 1KGP samples, each containing 200 bi-locus combinations, were extracted using two types of stratification: each sample contained an equal amount (40) of bi-locus combinations from individuals of each continent, as well as an equal distribution of degrees of separation (i.e. a metric of protein-protein interaction distance) between the genes of each pair, following the degrees of separation distribution of DIDAv1. Each 1KGP sample was used against the complete DIDAv1 set to train an individual classifier that gives a class probability for each bi-locus combination. Based on a majority vote among the individual classifiers, the output of VarCoPP for each tested bi-locus combination is the final class ("neutral" or "disease-causing"), a Support Score (i.e. the percentage of the classifiers agreeing about the pathogenic class, SS) and a Classification Score (i.e. the median probability among the individual predictors that the bi-locus combination is pathogenic, CS). Figure and caption from [369]

DIDA database, and the quality of the training data is therefore questionable. Finally, although it performs well for the analysis of gene panels, the false-positive rate makes the analysis of unfiltered datasets impossible, since the predictor returns too many irrelevant combinations as potentially disease-causing.

In this work, we develop a new version of VarCoPP, which is trained on a larger and more confident dataset, integrates more calibrated features from the different biological levels, and employs a less complex model structure to achieve higher performances in both cross-validation and independent tests (see Chapter 5).

3.5.3 ORVAL: Bringing oligogenic predictions to the public

The Oligogenic Resource for Variant Analysis (ORVAL) web platform was developed following the creation of the VarCoPP and the Digenic Effect predictors [105, 187], in order to make predictions using these tools more accessible to researchers and clinicians, as well as provide an interactive way of exploring the results generated by these tools [188]. This resource was first developed in 2019, as a web platform, where a user can input a list of variants of interest to be predicted, either in the format of a single patient VCF file, or as a list of genomic coordinates and zygosties. The user is also encouraged to select a gene panel in order to filter the variants to specific genes of interest, since VarCoPP is more suited for this kind of analysis (see Section 3.5.2). Once an analysis is submitted, the variants are filtered based on the filtering criteria used for the training set of VarCoPP, annotated with the features from the different biological using an in-house database, and all possible combinations of variants in two genes are generated and predicted using VarCoPP. Predicted pathogenic combinations are then also analysed with the Digenic Effect predictor.

In addition to predicted scores and classes for each variant combination according to the two aforementioned predictors(Figure 3.9D and F), ORVAL offers different ways of visualizing the data in their biological context. The predicted variant combinations can be visualized as a network, to facilitate module detection (Figure 3.9A). An explanation of the different predictions is available by using the *treeinterpreter* module (Figure 3.9E), which evaluates how each feature contributed to the decision of the classifier. Finally, information from external databases such as PPI connections between genes in the same module and pathway enrichment analyses are also made available to provide additional information on the biological mechanisms that can explain oligogenicity (Figure 3.9B&C).

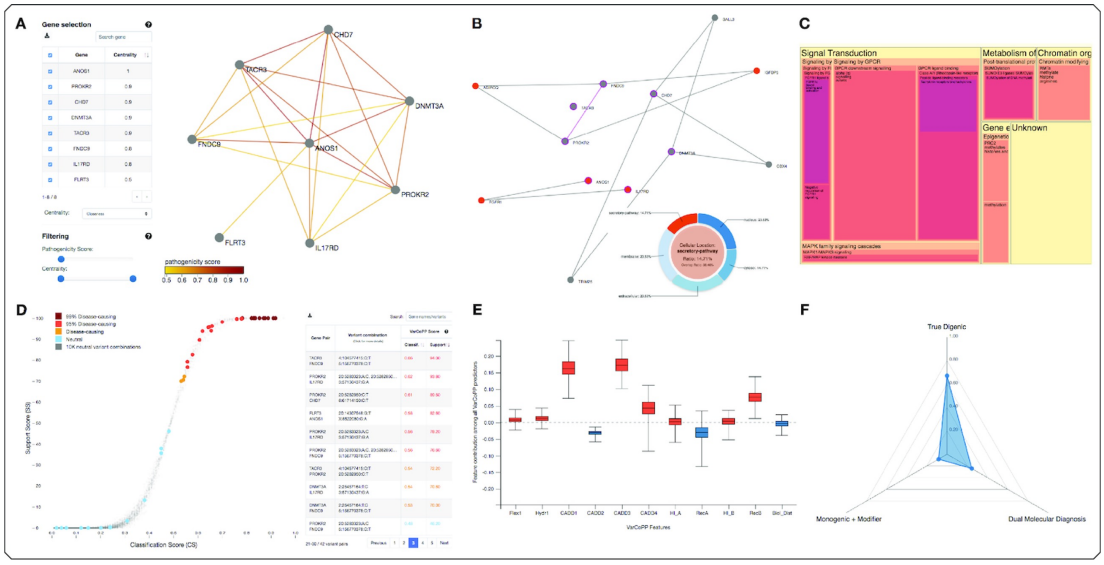


Figure 3.9: Examples of outputs generated by ORVAL. (A) Interactive oligogenic network built from all gene pairs having at least one predicted candidate combination. (B) A protein–protein interaction network where the central nodes circled in purple represent the proteins from a selected oligogenic module and the external nodes are the first-level interactors. (C) A Treemap representing the ontology sized proportionally to the number of mapped genes from the oligogenic module and colour according to the level on the ontology hierarchy. (D) An S-plot representing the classification of all digenic combinations as being neutral (in blue) or potentially disease-causing (from orange to dark red) depending on the predicted VarCoPP CS and SS. (E) A boxplot chart, displayed for a specific digenic combination, showing the contribution of each predictive features into the disease-causing class (in red) or neutral (in blue). (F) A spider plot, displayed for a specific digenic combination, showing the probabilities for each class of digenic effect predicted by the DE Predictor, with a highest probability for True Digenic. Figure from [188].

This platform has received great interest to the biomedical research community, and has been used in numerous studies to help identify novel oligogenic causes to disease [189–193]. The underlying annotation database was recently updated to include annotations for variants called using the Hg38 assembly. In this work, we integrate our novel predictive methods in this platform, to allow for easy use of the methods by the biomedical community.

3.5.4 OligoPVP: a first attempt at oligogenic prioritization

OligoPVP is the first developed oligogenic variant prioritization tool, which was trained on single variants data using deepPVP, but was tested on combinations from the DIDA database. In oligoPVP, the score of a combination is computed as the sum of the DeepPVP scores for the individual variants that make up the combination if they are located in genes that are connected in the PPI network. Therefore, one of the strong limitations of the tool is that variant combinations that involve genes that are not connected by PPI will never be scored and prioritized.

OligoPVP was tested using combinations from DIDAv2: the 189 variant combinations annotated with HPO terms were used to generate 189 synthetic exome (by inserting the variants of each combination in an exome from the 1KGP), and the predictor was applied to these exomes. The results indicate that out of the 87 DIDA combinations which are deemed "interacting" (i.e. the genes are connected in a PPI network), oligoPVP identifies 57 (65 %) as top hit and 59 (67%) in the top 10 hits and outperforms single-variant prioritizers.

However, the interacting combinations of DIDA represent only 53% of all DIDA combinations, which means that almost half of the combinations are actually not prioritized by the tool. The number of combinations which cannot be prioritized is supposed to diminish with time, as new knowledge about PPIs becomes available. Nevertheless, oligoPVP remains based on scores from a single variant tool, and might thus also fail to recognise variants which are less rare, or do not have the same properties as monogenic disease-causing variants.

3.5.5 BOCK: contextualizing oligogenic combinations in biological networks

Biological networks and Oligogenic Combinations integrated as a Knowledge graph (BOCK) is a resource that was developed at the Interuniversity Institute of Bioinformatics in Brussels by Alexandre Renaux, in parallel to this thesis' work and was published in 2023 [374]. We here introduce it as a completed work, although it was based on the work presented in Chapter 4, because it was then used as such in the development of our prioritization tool in Chapter 5 of this thesis.

This Knowledge Graph (KG) integrates different types of bioinformatics resources, together with information contained in OLIDA (see Chapter 4 or [253]) into a single heterogeneous network. It integrates protein-protein interactions from Mentha database [312], coexpression between genes from post-processed Gtex data collected in the TCSBN database [375], sequence similarity data between genes from STRING [376], pathway information from [314], gene ontology terms and their linked to genes from the Gene ontology [377, 378], protein families from the InterPro database [379], protein complexes from CORUM [380], phenotypic data from HPO [67], disease data from Orphanet[2] and Online Mendelian Inheritance in Man (OMIM) [79] and oligogenic combinations from OLIDA [253]. The overall architecture of the network and the number of different instances present is summarized in Figure 3.10.

2. https://www.orpha.net/

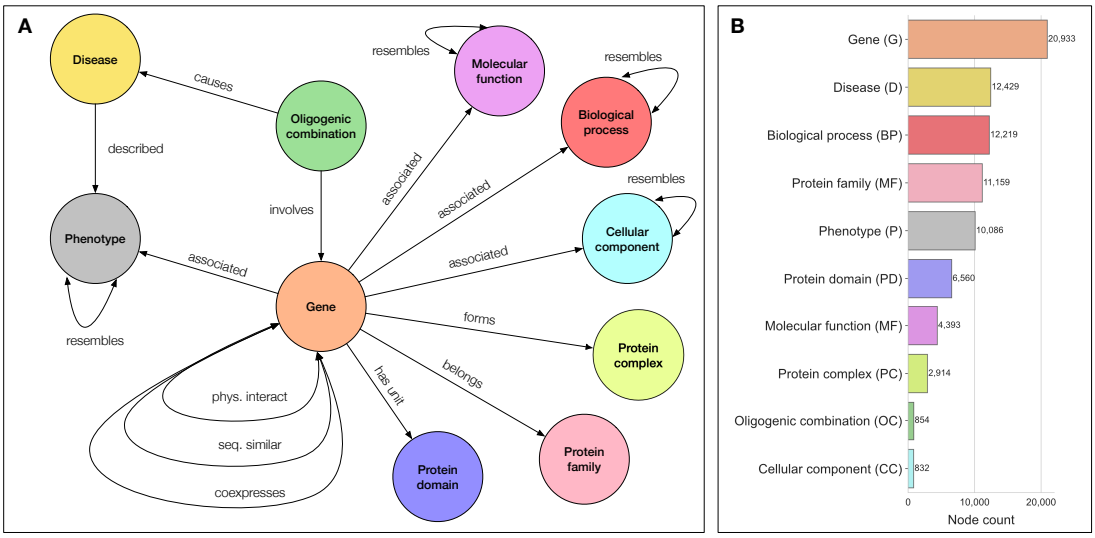


Figure 3.10: Summary representation of BOCK showing the 10 different types of nodes that are present in the graph, as well as the 17 different edges that link these nodes in panel A, and the node counts in panel B. Figure from [374].

3.5.6 Explaining digenic disease mechanisms with ARBOCK

Based on this knowledge graph, a rule based classification model was trained to identify specifically the characteristics of paths linking gene pairs that are known to be associated to an oligogenic disease. This model, called Association Rule learning Based on Overlapping Connections in Knowledge graphs (ARBOCK), allows to not only predict the probability for a gene pair to be involved in an oligogenic disease with relatively high performance, but it is also fully interpretable. This means that for any gene pair predicted to be involved in a disease, the model can extract the relevant paths from the graph that contributed to the pathogenic prediction.

These paths can be visualized in the form of a small network, which can help validate the prediction of the model, therefore improving trust in the prediction, but can also help to gain insights into the patho-mechanisms of the disease. This model can therefore be very useful to validate oligogenic findings and/or generate new hypotheses for digenic models. The model can be run with or without taking into account phenotype nodes in the graph as they were shown to lead to biases in the explanations and predictions. In Chapter 6, we use ARBOCK to investigate potential novel digenic combinations that were found in a cohort of patients affected with male infertility.

Chapter 4

Improving the quality of ground-truth oligogenic data

In this chapter, we present the Oligogenic Disease Database (OLIDA), a new repository of data on oligogenic variant combinations linked to human diseases. This database significantly expands the Digenic Disease Database (DIDA), providing for the first time access to variant combinations in more than two genes which are reported in the literature as associated to a human genetic diseases. In addition to the inclusion of new types of combinations and variants, the database also provides a protocol to assess the quality of the evidence linking this type of variant combinations to disease. Although such quality criteria existed for the association of monogenic variants, this is the first time that such criteria are developed for oligogenic variant combinations.

The novel repository contains unprecedented amounts of data on variant combinations reported to be causative of genetic diseases, and is then used in the remaining of the thesis to train novel predictive methods and assess their performance on new unseen data.

We first introduce the basis of the established curation protocol, describing in details the procedure to assign manual curation score and knowledge curation scores, and then describe how the database content was kept up-to-date over the years, by performing yearly rounds of curation. In the second part, we dive into statistics of the database, the effects of the confidence scores and how the content evolved over the last three years. We then introduce the first set of quality standards for reporting oligogenic variant combinations in the literature, which was based on the curation protocol. Finally, we highlight the interface of the database and how its ease of use has enabled various usage around the world.

4.1 Motivation and objectives

The Digenic Disease Database (DIDA) [104], published in 2015, was an important step forward in the study of oligogenic diseases, by enabling the development of various predictive tools and serving as the standard repository for digenic data (see Sections 1.3.5 and 3.5.1). As the number of articles identifying digenic and oligogenic causes to disease is continuously increasing, this database also started to exhibit some limitations. First, the architecture of the database only allows for the inclusion of digenic combinations, with a maximum of two variants per gene, and is limited to certain types of variants (SNVs and small indels). Secondly, due to the long manual work that is required for data curation, DIDA had not been updated in a while, with the latest data being collected in 2017. Finally, the criteria for inclusion needed to be re-evaluated, and precise standards for evaluating the quality of the data were required in order to objectively quantify the level of evidence associated with a variant combination.

In this chapter, we aim to address these limitations by creating a novel repository of variant combinations reported in the literature as associated to oligogenic diseases: the Oligogenic Disease Database (OLIDA). This database is created by manually curating scientific articles using a precise and detailed curation protocol, which assigns to each combination a confidence score, based on the evidence that supports its involvement in disease. Furthermore, it includes combinations of variants in more than two genes, without a maximum set number of variants per gene, as well as oligogenic combinations involving CNVs, a type of variant which was not present in DIDA. The main goal of OLIDA is to create a comprehensive repository of data on oligogenic diseases, as well as provide, with the confidence scores, a way to filter this data based on its quality. The curation protocol introduced is also the basis of a first set of recommendations for identifying and reporting oligogenic variant combinations involved in diseases. As such, a secondary aim of OLIDA is to improve the quality of the data regarding variant combinations associated to oligogenic diseases.

The results presented in this chapter have been published in:

- “Scaling up oligogenic diseases research with OLIDA: the Oligogenic Diseases Database” **Barbara Gravel** and Charlotte Nachtegael, Arnau Dillen, Guillaume Smits, Ann Nowé, Sofia Papadimitriou, Tom Lenaerts, Database, Volume 2022, 2022, baac023, doi:10.1093/database/baac023
- “Toward reporting standards for the pathogenicity of variant combinations involved in multilocus/oligogenic diseases” Sofia Papadimitriou, **Barbara Gravel**, Charlotte Nachtegael, Elfride de Baere, Bart Loeys, Miikka Vikkula, Guillaume Smits and Tom Lenaerts, Human Genetics and Genomics Advances, 2022, doi:10.1016/j.xhgg.2022.100165

The development of the curation protocol was the result of discussions between myself, Charlotte Nachtegael and Sofia Papadimitriou, and the curation of the scientific articles was divided equally among the three of us, with Inas Bosch replacing Sofia Papadimitriou for the latest version of the database (curation of articles published in 2023). The development of the website for the database was the work of Charlotte Nachtegael and Arnau Dillen, while my contributions focused on the pipeline for annotating the data and attributing the knowldege scores in the post-curation step.

4.2 General premises of the curation protocol

The developed curation protocol is inspired from the criteria from the paper of Schäffer on digenic inheritance [96] and the paper from McArthur on the causality of genetic variants in diseases [121]. The main premise of the protocol is that, for a combination to be associated to a particular disease, it is required to have sufficient **genetic and functional** evidence that links the **joint effect** of the variants to the disease phenotype. Genetic evidence can be obtained from pedigree studies (familial evidence) or cohort studies (statistical evidence), and should show that the segregation of the variants involved in the combination is linked to the phenotype and that these variants do not co-occur by chance. The functional evidence can be obtained by functional studies at the gene-combination (gene evidence) and the variant combination (variant evidence) levels, and should show that the joint effect of the variants have a deleterious effect on the function of the cell, and that his effect is linked to the observed phenotype.

In order to assess the quality of the evidence that supports the involvement of a combination in disease, we assign to each type of evidence a confidence level (with the corresponding score in parenthesis) as follows:

- Strong (3): If there is strong evidence of the joint pathogenic effect of the variants leading to the observed phenotype.
- Moderate (2): If there is evidence that the oligogenic combination has an effect on the observed phenotype, but some information on the underlying mechanism is missing to constitute a definite proof of oligogenicity.
- Weak (1): If there is evidence that the variants are relevant for the observed phenotype, but evidence is missing to show that the cause of the phenotype is indeed oligogenic and that all of the variants need to be present to cause the phenotype.
- Absent (0): If the information presented is not enough, according to our criteria, to attribute a Weak confidence level.

The curation pipeline is summarized in Figure 4.1.

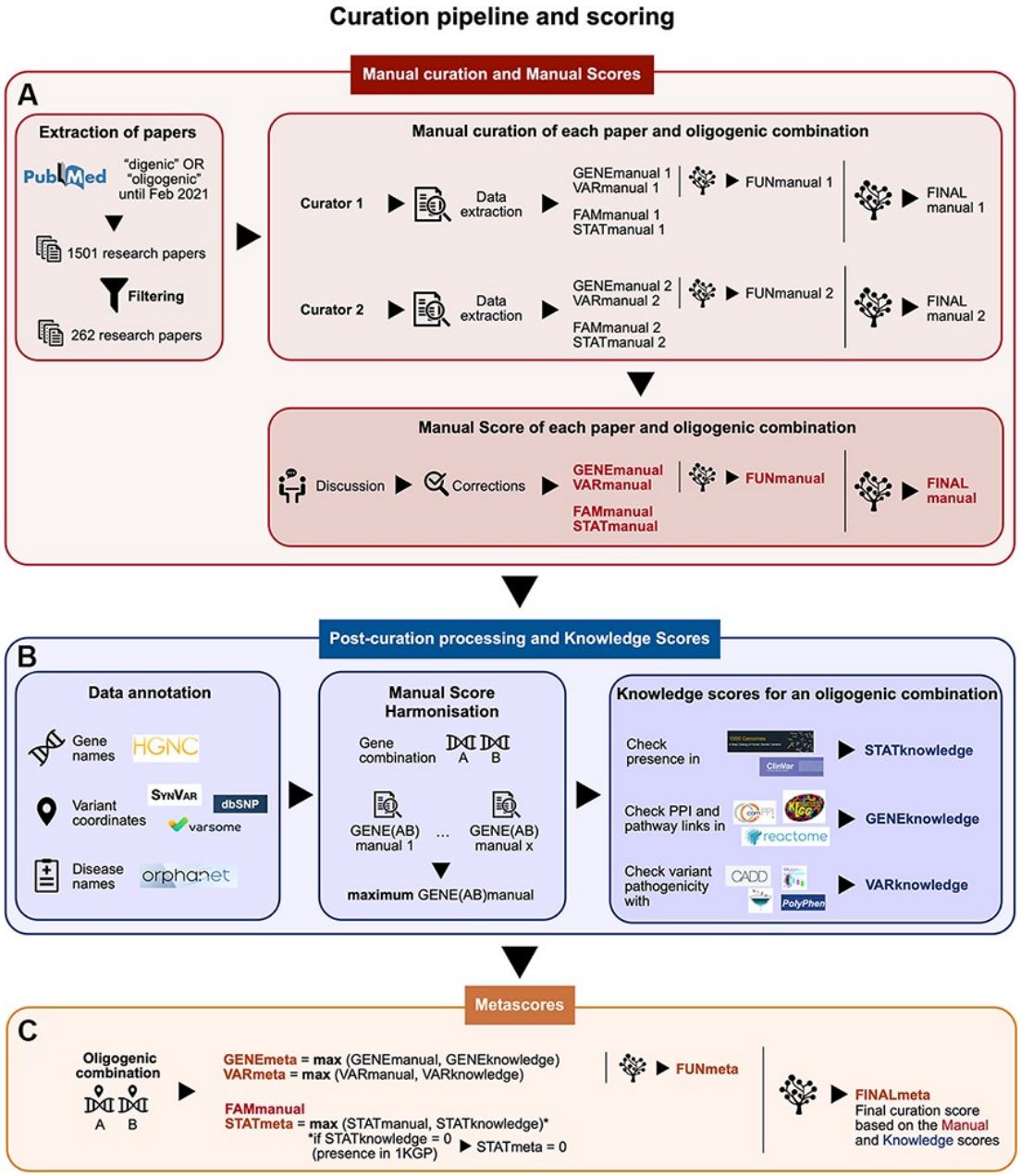


Figure 4.1: Summary of the curation pipeline for the creation of the (A) Manual scores, (B) Knowledge scores and (C) Metascores for each oligogenic combination. (A) Research articles are selected using the keywords 'digenic OR oligogenic' in PubMed and assigned to two different curators who independently extract information of each article. Note that the number of articles shown in this figure correspond to the search performed in January 2021, for OLIDAv1. A discussion then took place to reach a consensus for the Manual Scores. The number of articles correspond to the first version of OLIDA. (B) The data is then processed to formalize the available information using external databases and ressources. To correct for literature bias, the GENEmanual score for each gene pair was harmonized among all articles by assigning the maximum GENEmanual found for that gene pair. To compensate for missing information in the articles due to no prior access to current knowledge, Knowledge scores were assigned per oligogenic combination, based on the obtained annotations. (C) Both Manual and Knowledge scores are combined in order to create the confidence Metascores for each type of evidence by assigning the maximum score found between their corresponding Manual and Knowledge score. The same procedure as in the manual curation is then followed when decision trees are used to define the FUNmeta and FINALmeta scores.

4.3 Manual curation and manual scores

All articles which are found in PubMed[1] using the keywords “digenic” and “oligogenic” are first filtered by removing articles that (i) did not involve humans, (ii) did not report details about the variants, (iii) only conducted statistics at the gene level, (iv) contained large chromosomal rearrangements spanning more than one gene. From the initial 1853 articles found in December 2023 (which corresponds to the last OLIDA update), 399 passed these initial criteria and were manually curated.

Two curators are assigned to each article, extract information, and score the combinations independently according to specific criteria for each type of evidence (see Table 4.1). A discussion then takes place between the curators to assign consensus scores to every combination (Figure 4.1, panel A). The manual curation leads to the creation of manual curation scores, which only scores the information found in the article describing the combination. Four main types of scores were defined: **familial score**, which assesses the quality of the genetic evidence that is obtained from pedigree studies; **statistical score**, which assesses the quality of the genetic evidence provided by cohort analyses,; **gene pair functional score**, which assesses the quality of the functional evidence that implicates the gene combination in the disease of interest; and **variant functional score**, which assesses the quality of the functional evidence that implicates the variants as pathogenic.

Each score ranges from 0 to 3, except for the statistical evidence score which ranges from 0 to 2, and each score is assigned according to specific criteria detailed in Table 4.1 and explained below. The scores obtained from manual curation carry the “Manual” suffix.

4.3.1 Evaluating genetic evidence

The first type of evidence we consider is genetic evidence, which assesses whether the variants reported to be disease-causing are indeed found to be associated with the disease phenotype, and are not just found by chance. This can be done through pedigree studies, or cohort analysis, which we evaluate separately.

Familial evidence

For familial evidence, we evaluate the quality of the information brought forward by pedigree studies. The best type of genetic evidence can be obtained when studying a large pedigree, where the background environment of the involved individuals is controlled as much as possible, providing a clear way of studying the segregation of the suspected variants for the observed phenotype. The score is therefore influenced by the number of direct and indirect relatives that were sequenced and phenotyped, and how well the variants segregated with the

1. https://pubmed.ncbi.nlm.nih.gov/

phenotype. As “perfect segregation” we define the situation where we have information on the phenotypes of individuals in the family that carry the combination of variants in question, each involved variant alone, as well as all possible sub-combinations of the involved variants (in the case of an oligogenic variant combination involving more than two variants). For defining the “healthy” individuals, we take into consideration the fact that the oligogenic combinations can have different effects on the phenotype, the main two being a) the phenotype is only observed when the specific combination of variants is present in an individual and b) the “monogenic plus modifiers” scenario, where one variant can have an effect of the phenotype, which is then modulated by the presence of the extra variant(s) [96, 104]. In the first scenario, healthy individuals are considered as the ones without any disease phenotype, while in the second scenario, we also accept as healthy individuals, those with a less severe phenotype or whose age of onset of disease is later than the individual carrying the oligogenic variant combination (as shown in Figure 4.2).

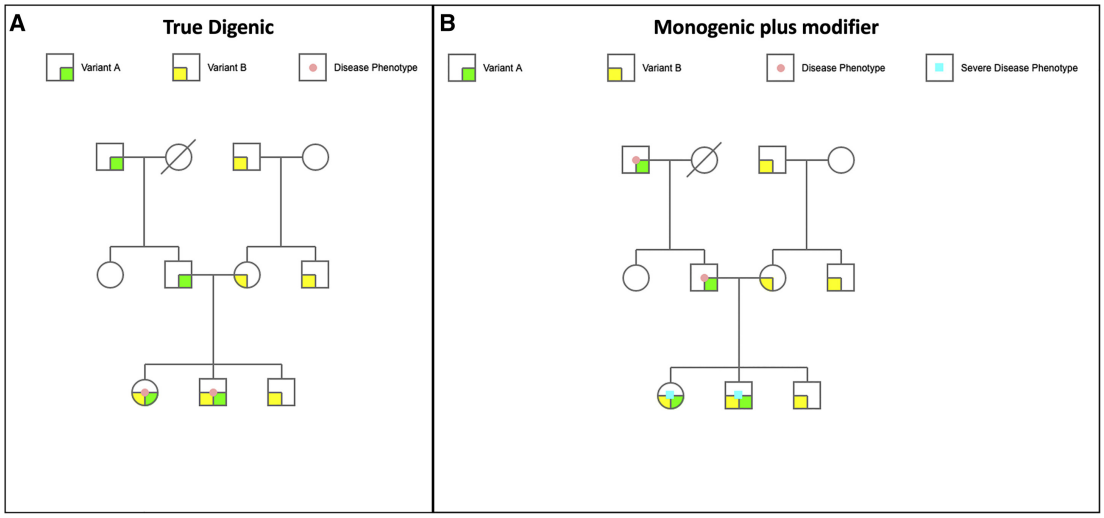


Figure 4.2: Examples of family pedigrees that demonstrate a clear genetic segregation for a digenic variant combination involving gene A and gene B. The digenic variant combination is associated with the disease phenotype under (A) the true digenic model, where the simultaneous presence of both variants in an individual are necessary for the development of the disease phenotype, and (B) the monogenic plus modifier model, where the variant at the second gene modifies the severity or age of onset of the symptoms caused by variant in the first gene. Variant A in the first gene is shown in green, variant B in the second gene is shown in yellow, and the individuals with the disease phenotype are shown with a red or blue dot, the latter representing a more severe disease phenotype.

Statistical evidence

The second type of genetic evidence can be obtained by a statistical study by either using (a) a cohort formed specifically for the study or (b) one (or more) of the available datasets containing genetic data of individuals [37, 273, 381]. The score is determined based on the accurate description of the phenotype of the individuals in the cohort, the number of individuals considered for the control population, considerations of ethnicity biases (which can strongly affect any relevant findings [382]), and the precise reporting of variants frequencies in the patient and control populations. In this case, since the genetic environment can not be perfectly controlled, the maximum score is moderate, which is only obtained when the authors specifically searched for the frequency of the combination of variants in a control population.

4.3.2 Functional evidence

The second type of evidence we consider is functional evidence, which assesses whether there is biologically plausible explanation for the involvement of the reported variants in the phenotype. This is evaluated at the gene level (assessing whether that the genes' functions can be linked to the phenotype) and at the variant level (assessing whether the variants have an impact on the protein product).

Gene level evidence

The gene combination functional evidence evaluates referenced or experimental evidence showing (a) a functional relationship between the genes involved in the oligogenic variant combination and (b) a relevance for the studied disease phenotype. The functional relationship gives information on whether the genes are involved in the same pathway or different pathways related to the same biological process, are being co-expressed together, have their gene products directly or indirectly interacting, are co-localised in the cell or are expressed in the same tissue. The relevance to the phenotype, on the other hand, gives information on whether the involved biological processes and the tissues in which the genes are expressed are relevant for the studied phenotype.

A strong gene combination functional evidence should show a synergy between the genes and an effect on the studied phenotype when these genes are not functional, which is not present or is different when only either single gene is not functional. For combinations involving more than two genes, this means that different sub-combinations of gene knock-outs need to be evaluated independently and compared to the knock-out of all the genes involved, to ensure that all the genes together are associated to the phenotype.

Variant level evidence

The variant combination functional evidence evaluates the effect of the combination of variants on the studied phenotype. This type of evidence is separate from the gene combination evidence, as not all variants can have a pathogenic effect, even if there is a proof of functional synergy between the genes of a combination. A strong variant combination evidence should show a synergistic/additive effect when variants occur together compared to when they occur alone. The variant combination functional evidence is divided here in experimental evidence (i.e. *in vitro* and *in vivo* experiments described or referenced in the paper) and *in silico* evidence (e.g. pathogenic predictors, protein sequence analysis or predictions of disruption of the protein structure). Here, it should be noted that if a clear effect is shown for loss-of-function variants, either those present in the studied combination or for different variants in the same genes, this is also translated as an effect at the gene level. For example, if one of the variants in the oligogenic combination leads to a single gene knockout that is linked to a clear effect on the studied phenotype, this effect is also translated at the gene level as gene knock-out evidence, which is then also used to assign the gene combination manual score.

4.3.3 Defining final scores

The gene combination evidence and the variant combination evidence scores are then combined to obtain the aggregated functional evidence score using decision trees (see Appendix A.1). For this process, we put more importance on the variant level type of evidence, as it is more precise and specific to the studied oligogenic variant combination. Finally, the familial, statistical and functional scores are combined into a final manual score for the oligogenic combination, using a decision tree (see Appendix A.2). This score represents the overall quality of the evidence that associates a particular variant combination to a disease based on the information present in the article only. The creation of the decision trees was done through an iterative trial-and-error process where we first defined decision trees, which were then adapted as we curated cases.

FAMmanual: familial evidence based on the article		
Weak (1)	Moderate (2)	Strong (3)

One of two conditions: a. The genotypic and phenotypic information of only one healthy first-degree relative is described b. Imperfect segregation in a pedigree with information on two (or more) first-degree relatives	One of two conditions: a. Information of two (or more) first-degree relatives, showing a perfect segregation of the variants according to the phenotype b. Imperfect segregation in a pedigree with information on first- and second-degree relatives	Information of healthy first- and second-degree relatives, showing a perfect segregation of the variants according to the phenotype
STATmanual: statistical evidence based on the article		
Weak (1)	Moderate (2)	Strong (3)
Implicit evidence that healthy individuals do not carry the oligogenic combination, based on control cohorts or public databases. Known control phenotypes, sufficient control size and matched ethnicity.	Explicit evidence that healthy individuals do not carry the oligogenic combination, based on control cohorts. Known control phenotypes, sufficient control size, matched ethnicity and (preferably) similar sequencing technology.	N.A.
STATknowledge: Statistical evidence based on databases and cohorts		
Weak (1)	Moderate (2)	Strong (3)
The combination is not found in the 1000 Genomes Project and relevance of all involved variants in Clinvar.	N.A.	N.A.
STATmeta: Maximum of STATmanual and STATknowledge		
Weak (1)	Moderate (2)	Strong (3)
The oligogenic combination is not found in the 1000 Genomes Project. Other implicit evidence of its statistical relevance for the phenotype.	The variant combination is not found in the 1000 Genomes Project. Additional explicit evidence that healthy individuals do not carry the oligogenic combination, based on control cohorts or public databases.	N.A.
GENEmanual: Gene functional evidence based on the article		

Weak (1)	Moderate (2)	Strong (3)
Relevance of involved pathway(s) or expressed tissues on the studied phenotype.	One of two conditions: a) Effect of the gene combination on the observed phenotype using a functional experiment with either only a double knock-out or multiple single gene knockouts b) Direct gene relationship (e.g. common pathway, direct interaction) and relevance for the studied phenotype.	Synergistic or additive effect of the gene combination on the observed phenotype using a functional experiment with single and multiple gene knock-outs.
GENEknowledge: Gene functional evidence based on databases		
Weak (1)	Moderate (2)	Strong (3)
Relevancy of Reactome or KEGG pathways linked with the genes for the observed phenotype.	One of two conditions: a) Gene combination forms a connected PPI network and the comPPI score of each link is >0.8 b) Common Reactome or KEGG pathways, relevant for the observed phenotype.	N.A.
GENEmeta: Maximum of GENEmanual_harmonized and GENEknowledge		
Weak (1)	Moderate (2)	Strong (3)
Relevance of the genes on the studied phenotype using pathway or tissue expression information.	Direct gene relationship or effect of the gene combination on the observed phenotype, without assessing their individual effects.	Synergistic or additive effect of the gene combination on the observed phenotype using a functional experiment with single **and** multiple gene knock-outs.
VARmanual: Variant functional evidence based on the article		
Weak (1)	Moderate (2)	Strong (3)

One of three conditions: a) All variants are predicted as pathogenic b) Functional experiments for some variants and predicted pathogenic effects for the rest c) Functional experiments using single mutants, with a promising but not conclusive effect of the oligogenic combination on the observed phenotype	One of two conditions: a) Effect of the variant combination on the observed phenotype using a functional experiment with **either** only a double mutant **or** multiple single mutants b) Clear pathogenic impact of the variant combination on the observed phenotype in an in silico analysis of the **joint** effect of the variants.	Synergistic or additive effect of the variant combination on the observed phenotype using a functional experiment with single **and** multiple gene mutants.
VARknowledge: Variant functional evidence based on predictors		
Weak (1)	Moderate (2)	Strong (3)
Pathogenicity prediction for all variants by at least one predictor among CADD, SIFT, MutationTester and Polyphen.	N.A.	N.A.
VARmeta: Maximum of VARmanual and VARknowledge		
Weak (1)	Moderate (2)	Strong (3)
Pathogenicity predictions for all involved variants or inconclusive effects of functional experiments	One of two conditions: a) Effect of the oligogenic combination on the observed phenotype using a functional experiment with either only a double mutant or multiple single mutants b) Clear pathogenic impact of the oligogenic combination on the observed phenotype in an in silico analysis of the joint effect of the variants.	Synergistic or additive effect of the variant combination on the observed phenotype using a functional experiment with single and multiple gene mutants
FUNmanual: Functional evidence based on GENEmanual and VARmanual **FUNmeta: Functional evidence based on GENEmeta and VARmeta**		
Weak (1)	Moderate (2)	Strong (3)

Based on a decision tree, not enough evidence to suggest synergy but at least relevance of genes and variants.	Based on a decision tree, evidence of functional synergy and relationship of the involved genes and variants, but the joint pathogenic effect on the studied phenotype is still not confirmed or clear.	Based on a decision tree, strong evidence of the functional synergy of both involved genes and variants on the studied phenotype.
FINALmanual: Overall evidence based only on Manual scores **FINALmeta: Overall evidence based on Manual and Knowledge scores**		
Weak (1)	Moderate (2)	Strong (3)
Based on a decision tree, only evidence of the relevance of the variant combination for the observed phenotype, but not enough to show that the involved variants are the only culprits for the studied phenotype or that the cause is indeed oligogenic.	Based on a decision tree, good genetic and functional evidence of an effect of the oligogenic variant combination on the observed phenotype, but the described information/mechanism is not clear or strong enough to provide proof of oligogenicity.	Based on a decision tree, strong evidence of the synergistic/additive effect of the oligogenic variant combination on the observed phenotype genetically and functionally.

Table 4.1: Summarized descriptions of the curation confidence scores linked to the variant combinations present in OLIDA. Decision trees (see Appendix A) are used to define the FUNmanual, FUNmeta, FINALmanual and FINALmeta scores. For each type of evidence, if the information found for a combination does not fulfil the criteria to provide at least a Weak (1) score, an Absent (0) score is assigned. Decision trees used to assign the FUN and FINAL scores are in Appendix A.

4.4 Post-curation process

The information collected during the manual curation then went through several post-processing step in order to normalize, harmonize and further annotate the data with additional databases (Figure 4.1, panel B). First the variants were converted to genomic coordinates in both Hg19 and Hg38 genome assemblies using SynVar (http://goldorak.hesge.ch/synvar/), Varsome [383] and dbSNP [384], and genes were searched in the HGNC database [385] in order to obtain their standard gene names. Disease names were standardized using Orphanet[2] and OMIM[3].

2. https://www.orpha.net/
3. https://www.omim.org/

Since multiple variant combinations were found in the same gene combinations, we also proceed to a harmonization of the manual scores, by defining the gene combination harmonized score (GENE_harmonized) as the maximum of the gene combination scores found over the course of the manual curation process.

Furthermore, for an oligogenic combination described in multiple papers, we assigned the highest Manual score from each type of evidence found among the papers describing that combination and re-calculated its FINALmanual score, using the decision trees. This harmonization can better depict the fact that different papers may focus their efforts in different aspects of proving the oligogenicity of a combination (e.g. one can focus on the genetic evidence of an enriched pedigree, and another on proving the synergy of the genes and variants with functional experiments).

4.4.1 Knowledge scores

In order to compensate for missing information in some articles (e.g. older articles which did not have access to large biobanks of variants frequencies or missed experimental evidence which appeared later in time), we created additional scoring measures for certain types of evidence that could be obtained from search in public external databases. These knowledge scores have the same strength meaning as the corresponding manual scores.

The statistical knowledge score (STATknowledge) is based on information obtained from the search of variants in databases of variant frequencies and previous reports of the variants in the combination as pathogenic in ClinVar [83]. A variant combination is thus attributed a statistical knowledge score of 1 if all variants involved in the combinations have been reported as pathogenic in ClinVar, and these variants are not found together in an individual of the 1KGP project. Otherwise, the statistical knowledge score is 0 (Table 4.1).

The gene combination knowledge score (GENEknowledge) is based on information about whether the genes are known to interact or are involved in a common pathway that is linked to the phenotype (Table 4.1). Interaction data was obtained from comPPI [313], where only interaction scores of at least 0.8 were considered. Pathway data was obtained from Reactome [314] and KEGG [315], and common pathways between the genes were then manually screened for relevance to the corresponding disease.

The variant combination knowledge score (VARknowledge) is based on the results from in silico prediction tools for single variants (Table 4.1). We predicted each variant with SIFT [158], PolyPhen2 [159], CADD [161] and MutationTaster2 [386]. A variant combination score was attributed a VARknowledge score of 1 if all variants in the combination were predicted as disease-causing by at least one pathogenicity predictor. For CADD, the pathogenicity threshold is defined with the Phred value 15, and for Polyphen2 both “possibly damaging” and “probably damaging” values are accepted as an indication of deleteriousness.

4.4.2 Metascores

Finally, for each type of evidence, metascores were defined by combining the manual and knowledge score (Figure 4.1, Panel C). These scores are given the meta suffix and are computed as the maximum of the manual and knowledge score for each type of evidence. The only exception is the statistical score: if the STATknowledge is 0 due to the presence of the combination in an individual of the 1KGP, then the STATmeta is also 0.

Once the individual metascores (for statistical, gene and variant scores) are defined, the functional and final metascores are computed using the same decision trees as in the manual process, this time based on the metascores (see Appendix A). The FINALmeta scores thus reflect the general confidence level attributed to each combination, using both information present in the article that reports the combination, but also information that has accumulated over time and is present in other articles or public biological databases.

4.5 Maintaining the database up-to-date

Following the publication of OLIDA, and the release of the first version of the database in 2022, we proceeded to yearly updates of the database. At the beginning of each calendar year, we searched for all articles published during the year which could be found using the keywords “Digenic OR oligogenic" in PubMed. These articles were then filtered based on the abstracts, to only include the articles which identified oligogenic combinations in humans. Once again, two curators were appointed per article in order to assign manual scores to the combinations independently.

The collected information then went through the same post-curation process, except that we also included the information present in the previous version of OLIDA.

The GENEmanual_harmonized score was therefore based on the entire database (previous version and update). This thus resulted in updating the scores of some combinations of the previous OLIDA releases, if they involved the same gene combinations as some new OLIDA combinations with a higher GENEmanual score. Scores of all OLIDA combinations were thus recomputed based on the harmonization over all gene combinations in the database.

4.6 Statistics of the database

Each update of the database is tagged with a version number. Since we performed yearly updates of the content, OLIDA is now at version number 4, which contains all oligogenic variant combinations reported in the literature up to December 2023 (included). The original version of OLIDA (containing combinations published up to December 2021) will be referred to as OLIDAv1 for the rest of the thesis, while the subsequent versions of the database are called OLIDAv2 for the 2022 release, OLIDAv3 for the 2023 release and OLIDAv4 for the 2024 release. We here report on statistics of OLIDAv4.

4.6.1 General statistics

OLIDAv4 contains 1808 variant combinations involving 1198 genes, 3777 variants and 219 diseases. As opposed to DIDA, which only included digenic variant combinations, OLIDA now includes 423 combinations involving variants in more than 2 (and up to 17) genes. Furthermore, it contains 129 combinations which involve 70 distinct CNVs, a type of variant not previously included in DIDA. The content of this database therefore represents a 7-fold increase as compared to DIDA, the only existing repository of data on digenic diseases before this work (Figure 4.3).

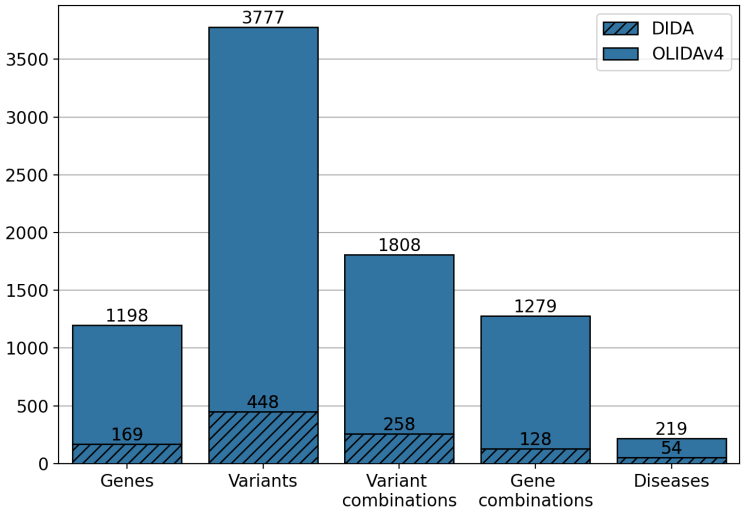


Figure 4.3: Histogram of the count of the number of entities in OLIDAv4 compared to the DIDA database.

OLIDA supplements the information present in the articles identifying the combinations with information from public databases, collecting data on variant effects, variant frequencies and interaction between the genes implicated in a combination. This information reveals that the majority of the variants in OLIDA are rare or absent from public databases of genetic variation,

with 69% and 35% of variants not being reported in the 1KGP and GnomAD, respectively, while 91% of the variants reported in the 1KGP and 95% of the variants reported in GnomAD have a Minor Allele Frequency of less than 1% in these databases. Additionally, annotating the variants with the results of 4 pathogenicity predictors (see Section 4.4.1) reveals that 73% of these variants are predicted as disease-causing by at least one tool.

In total, we encountered twelve different types of relationships between the genes involved in a combination and the relevant disease (Figure 4.4). Genes are "Involved in the same disease" if patients with the same phenotype described or referenced in the manuscript carried mutations in those genes, together or independently. Pathway information for each gene was either described in the article or found in the KEGG [315] or Reactome [314] databases, and were then manually screened to check if the genes belong to "Relevant pathways for the phenotype" or in the "Same pathway". Similarly, genes were labeled as "Affecting the same tissue" only if that tissue is relevant for the phenotype. The "Directly interacting" denotes a protein-protein interaction, either described in the article or retrieved from the comPPI database. It is distinguished from the "Same protein complex" relation where the gene products are considered to only fulfil their function when linked together (e.g. the subunits of a channel). "Indirectly interacting" genes are those whose products indirectly interact with an intermediate protein, or are involved in a gene regulation mechanism with other gene products (e.g. transcription factors). "Similar function" indicates that genes have the same function (e.g. motor proteins). "Co-localization" implies a direct overlap of the location of the gene products in the cell (e.g. shown using immunofluorescence), while "Same organelle" implies that the protein products exercise their function in the same organelle (e.g. cilia proteins). The "Co-expression" relationship implies a positive correlation of the mRNA expression of the genes in a temporal fashion shown or referenced in the article. Finally the "Monogenic experiments only" notes the fact that the experimental evidence and the assessment of their pathogenicity was done on the genes independently (e.g. single knock-outs).

The annotation of combinations with these relationship terms indicate that the vast majority of combinations in OLIDA (74%) have genes that are known to be involved in the same disease or are in pathways that are relevant for the disease phenotype (71%). If we look only at the terms that represent a biological relationship between genes (e.g. same KEGG or Reactome pathway, expression in the same tissue), out of the 1279 distinct gene combinations, 1189 (93%) combinations have at least one type of gene relationship, with a certain amount of combinations (72%) having more than one gene relationship type (Figure 4.4). These results further support the fact that genes involved in similar diseases are also closely related in biological networks.

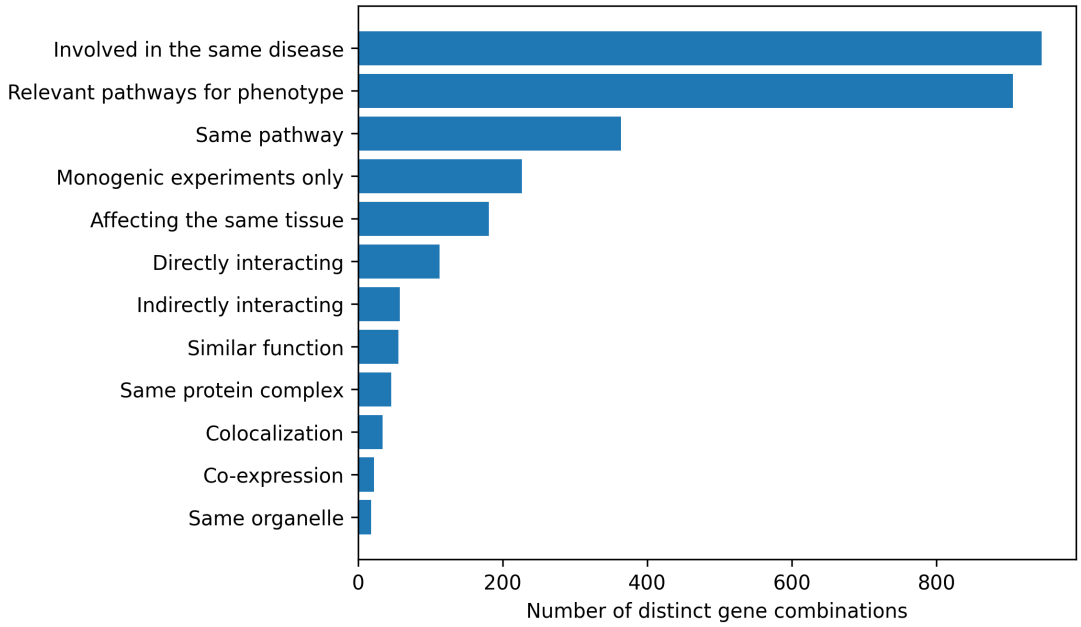


Figure 4.4: Histogram of the different gene relationship types found between the genes involved in an oligogenic variant combination. The types of gene relationship were obtained either directly from the articles or from public databases for the relationships for which such information is available (i.e. the “Relevant pathways for phenotype”, “Same Pathway” and “Directly Interacting” relationships).

OLIDA collects information on 219 genetic diseases, which is four times the amount of diseases (54) reported in DIDA (Figure 4.3). The most represented diseases in DIDA included well-known examples of digenic diseases such as Bardet-Biedl syndrome, familial haemophagocytic lymphohistiocytosis and familial long QT syndrome [104]. New diseases happen to be at the top of this list in OLIDA (Figure 4.5), with Kallmann syndrome and Amyotrophic Lateral Sclerosis (ALS) being linked to 9% of the variant combinations in the database, while congenital hypothyroidism and normosmic congenital hypogonadotropic hypogonadism are each linked to 7% of the total number of combinations. The presence of ALS as one of the top diseases in OLIDA while it was completely absent in DIDA is mainly due to the fact that the oligogenic causes associated with this diseases involve CNVs, a type of variant which was not present in DIDA. OLIDA also includes diseases that were not previously associated with oligogenic inheritance, such as arthrogryposis syndrome, holoprosencephaly, adolescent idiopathic scoliosis and Müllerian aplasia. Finally, it is important to note that more than half of the diseases in OLIDA (59%) are linked with only one or two associated oligogenic combinations.

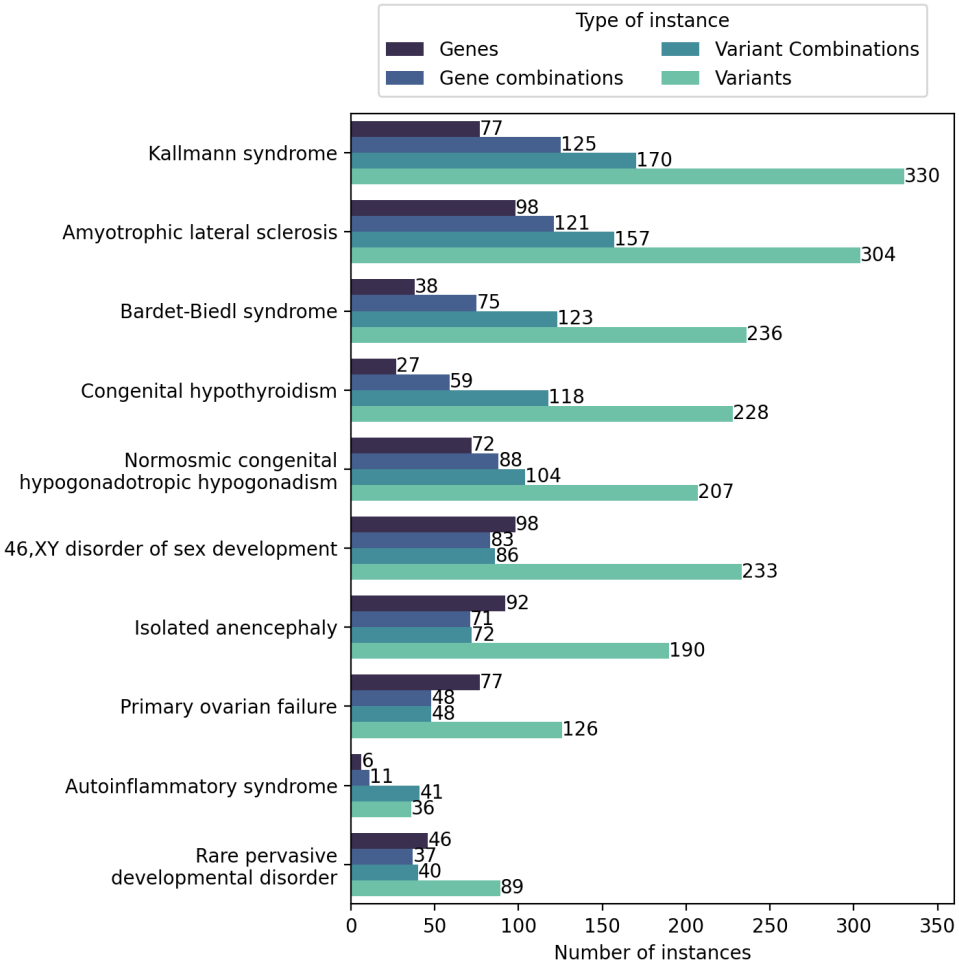


Figure 4.5: Bar chart of the number of instances in OLIDA for the ten diseases linked to the largest number of variant combinations. The diseases are ordered by the number of variant combinations they are associated with.

4.6.2 Statistics on the confidence scores of the oligogenic combinations

An important innovation of OLIDA is the attribution of confidence scores to every variant combination reported in the database, which are derived from our detailed curation protocol (see Table 4.1). This protocol assigns scores for the different types of evidence that link a combination to a particular disease: familial evidence (FAM), statistical evidence (STAT), gene functional evidence (GENE) and variant functional evidence (VAR). These confidence levels are assigned based on information retrieved from the articles identifying the combination (scores with the 'manual' suffix), information from external databases ('knowledge' suffix) and a combination of both sources of information ('meta' suffix). These individual scores are then combined into a final score (FINAL) which reflects the overall strength of the evidence that links a combination to a particular disease.

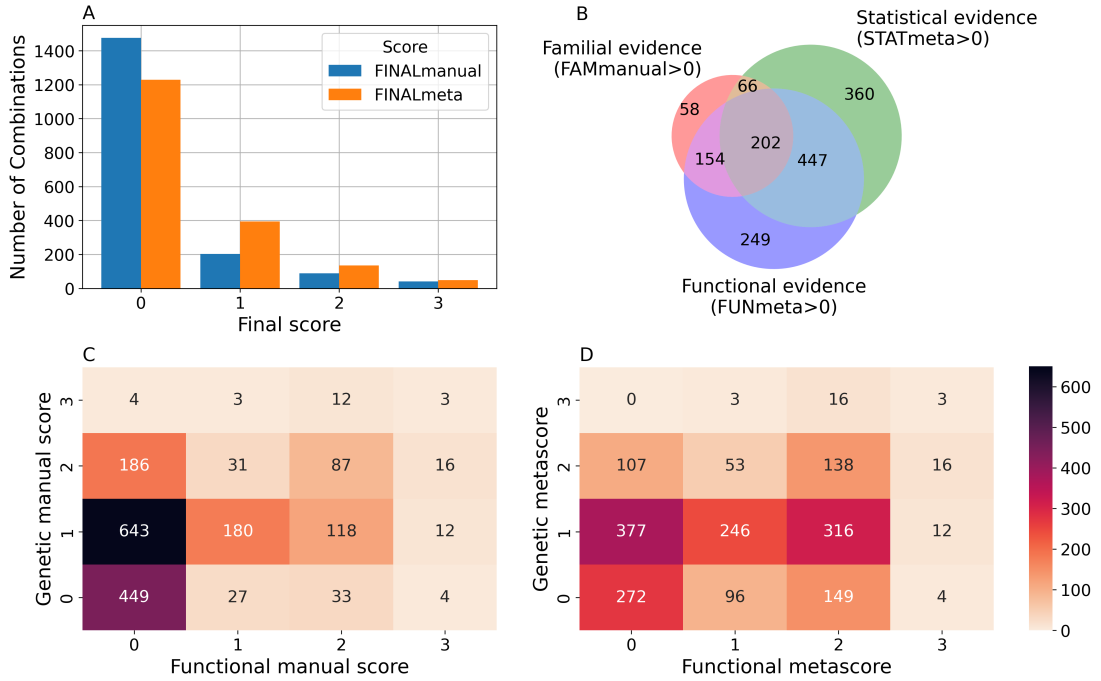


Figure 4.6: Confidence scores and types of evidence associated with the combinations of OLIDAv4. (A) Distribution of the FINALmanual and FINALmeta scores. (B) Venn diagram of the number of combinations associated with each type of evidence (Familial, Statistical and Functional). Heatmap of the number of combinations associated with the different levels of genetic and functional evidence based on the article information only (manual scores, panel C) and based on both article and database information (metascore, panel D).

Out of the 1808 combinations in OLIDA, 332 (18%) initially had a FINALmanual confidence score larger than 0 (i.e. based exclusively on information present in their corresponding article). The addition of information from external databases to supplement any missing information from the articles resulted in 579 (32%) oligogenic combinations with a FINALmeta of 1 or higher (Figure 4.6). The majority of these combinations have a FINALmeta score of 1, which corresponds, according to our criteria, to the minimum amount of evidence required to link a combination with a disease.

Based on both article and knowledge information, 1075 (59%) of the combinations are supported by at least weak statistical evidence, 1052 (58%) are supported by functional evidence, while only 480 (27%) of the combinations have at least weak familial evidence (Figure 4.6B). This is partly due to the fact that familial evidence could not be obtained in the post-curation process, while the statistical and functional evidence were increased with the knowledge scores. Overall, 202 (11.1%) combinations were supported by all three types of evidence. Only three combinations, all digenic, are associated with strong oligogenicity confidence (score of 3) for both genetic and functional evidence, while 138 cases are supported by both moderate genetic and functional evidence (Figure 4.6D).

The main reason why combinations had final scores of 0 based on either article information only (FINALmanual) or article information supplemented by external knowledge (FINALmeta) is the lack of functional evidence (Figure 4.6). Indeed, 833 (46%) combinations presented with sufficient manual genetic evidence but had a functional score of 0, while only 64 combinations had a FUNmanual higher than 0 but absent genetic evidence. Finally, there were 449 (25%) combinations which lacked both functional and genetic evidence considering manual curation, highlighting an overall problem with the reporting of pathogenic variant combinations in the literature.

The oligogenic combinations involving more than two genes, which were absent in DIDA, overall present lower confidence scores than the digenic ones (Figure 4.7). The large majority (89.4%) of these combinations have a FINALmeta score equal to 0, with 39 (9.2%), 4 (0.9%) and 2 (0.5%) combinations being assigned a score of 1, 2 and 3, respectively. The distribution of FINALmeta scores for these combinations is more skewed than the one for the combinations with variants in two genes only, where 851 (61.4%) combinations have a FINALmeta score of 0, 356 (25.7%) of 1, 131 (9.5%) of 2 and 47 (3.4%) of 3.

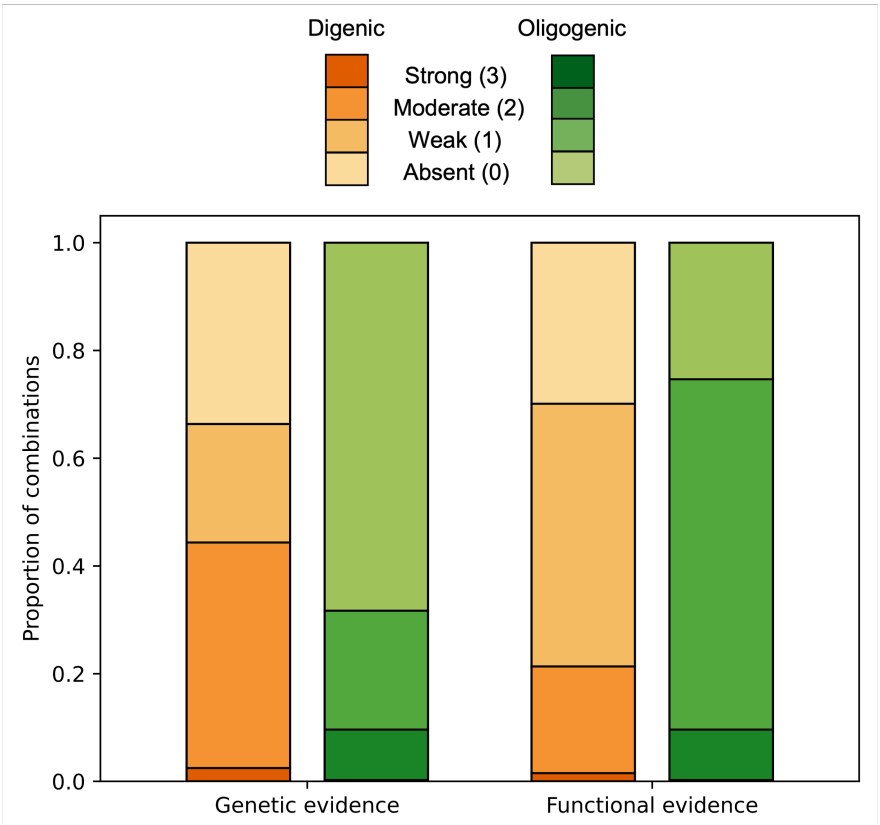


Figure 4.7: Proportion of the associated genetic and functional evidence scores for the digenic (in orange) and higher order (in green) oligogenic variant combinations present in OLIDA The color gradient represents the confidence score (as “absent”, “weak”, “moderate”, or “strong” and the corresponding score number from 0 to 3 in parentheses).

4.6.3 Evolution of the content of the database

Since its first release in 2022, we proceeded to 4 udpates of the content of OLIDA, adding combinations reported between the years 2021 and 2024. All versions are available in Zenodo at https://doi.org/10.5281/zenodo.10732286.

Investigating the evolution of the content of the different entities of the database allow to notice certain biases in the content of the repository. First, we notice that the number of combinations have almost doubled in the past 4 years, indicating a strong increase in the reporting of oligogenic combinations in the literature and thus highlighting the need for such a public repository. However, this huge increase in the number of combinations is not reflected in the number of diseases the combinations are associated with, which only increased from 158 to 218 (Figure 4.8). This means that the majority of the newly identified variant combinations are found in association with genetic diseases that were previously shown to display oligogenic inheritance, and indicates that researchers are probably more prone to search for oligogenic variants in specific diseases, therefore demonstrating a bias towards specific diseases in the field of oligogenic research. Similarly, the number of genes in which variant combinations are found only increased by 1.5 fold, implying that combinations are also increasingly found in the same genes, probably through their associations to the same diseases.

Second, we observe that the increase in the number of articles reporting on oligogenic variant combinations also does not follow the increase in the number of variant combinations (Figure 4.8). This means that more variant combinations are identified per article, illustrating a shift toward a new era where more genetic causes are found through large cohort analyses.

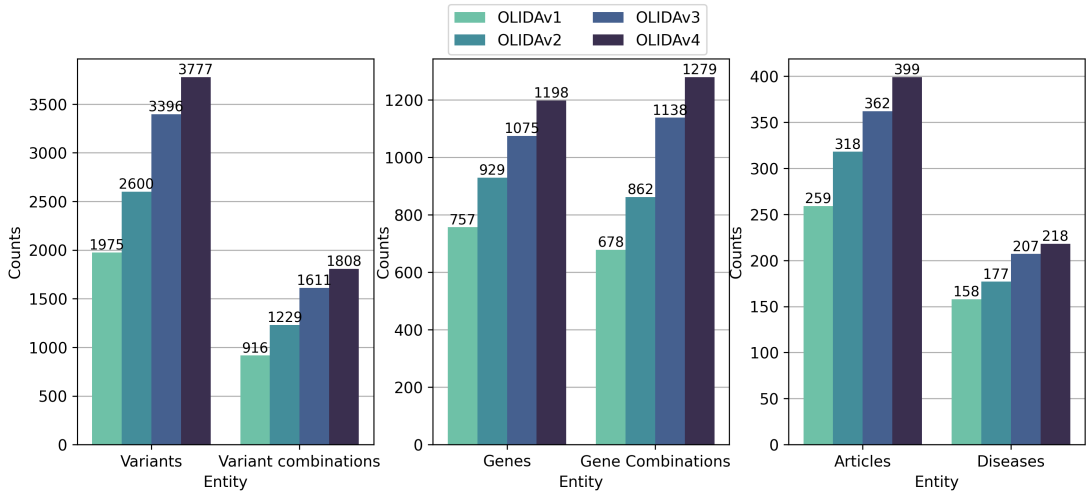


Figure 4.8: Histograms of the number of variants, variant combinations, genes, gene combinations, articles and diseases over the different versions of OLIDA.

This change in the analyses performed to identify genetic causes to disease is also characterized with a decrease in the proportion of combinations in the database associated with familial evidence over the years, and an increase in the proportion of combinations associated with at least weak statistical evidence (Figure 4.9).

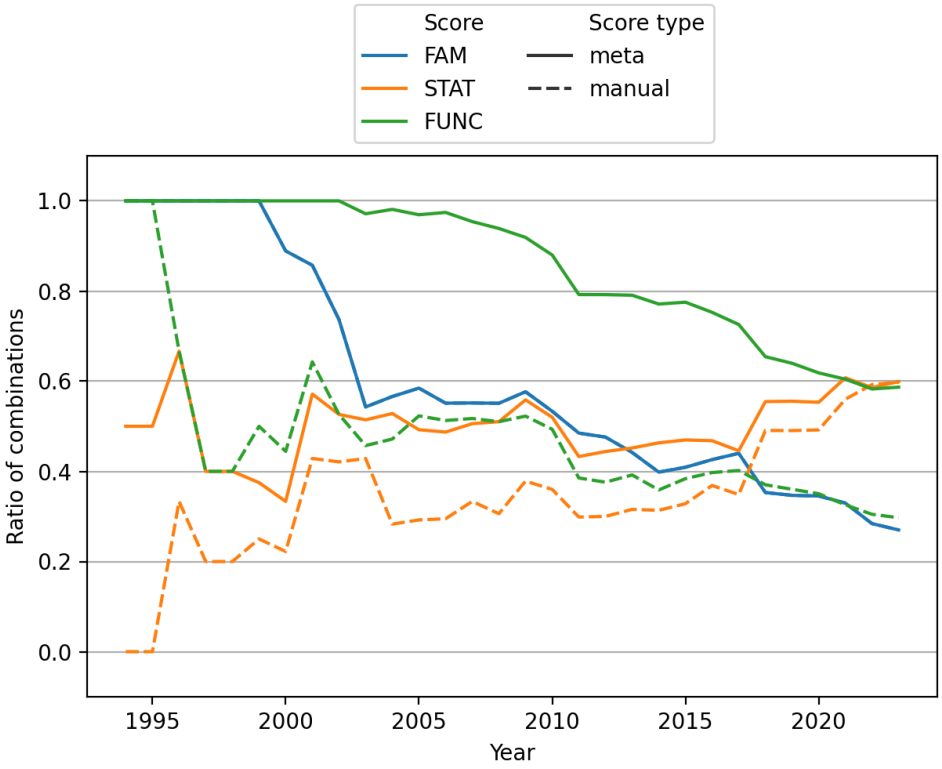


Figure 4.9: Cumulative proportion of combinations with a familial (FAM), statistical (STAT) or functional (FUNC) score of at least 1, from the first report of oligogenic combination in 1994 to the last update of the database in 2024.

Finally, it is important to notice that overall, the proportion of combinations which are associated with sufficient evidence of oligogenicity is not increasing over the years (Figure 4.9). We can observe that the ratio of combinations supported by at least weak level of evidence for both familial and functional evidence decreases over the years, even when considering the integration of knowledge from external databases and pathogenicity predictors for Functional evidence. This observation led us to define a first set of standards and recommendations for the reporting and assessment of pathogenicity of oligogenic combinations, which are summarized in the following section.

4.7 Defining a first set of standards for the reporting of oligogenic combinations

One important goal with the creation of the curation protocol of OLIDA is to initiate the discussion about standards in assessing the pathogenicity of oligogenic variant combinations. Indeed, while such guidelines exist for single variants, with the ACMG criteria [78] and other recommendations [121], very little has been done about oligogenic variant combinations, leading, as we noticed with the confidence scores in OLIDA, to very low levels of evidence associated with such combinations. Since the number of articles reporting such causes to disease is continuously increasing, setting such standards is becoming imperative.

In order to insist on the importance of these standards, we published a commentary article that identifies the key points that medical researchers and biologists should look at when assessing the pathogenicity of variants in combinations. These are based on the same two premises as the curation protocol of OLIDA:

- The articles should show evidence that rejects monogenicity, i.e. show that it is the joint effect of the variants that lead to the observed phenotype and that each variant alone has a smaller or no effect.
- The rejection of monogenicity should be supported by both genetic and functional evidence, i.e. there should be proof that the segregation of the variants identified as linked to the phenotype is not due to chance and that the variants have a synergistic effect when tested using *in vitro*, *in vivo* or *in silico* experiments.

As shown in Figure 4.10 the definition of such standards is essential as most studies only present weak or absent evidence when reporting oligogenic combinations in relation to disease. The proportion of combinations lacking evidence even seems to increase with time (see Section 4.6.3, although more resources are now becoming available such as large databases of variant frequencies and prediction tools for variant pathogenicity.

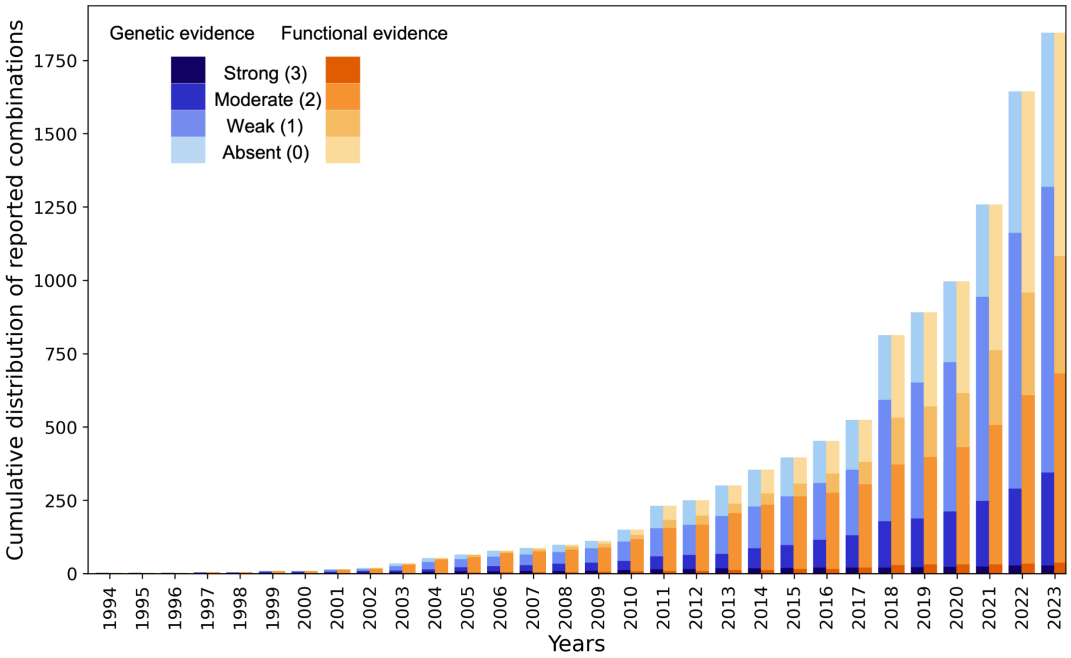


Figure 4.10: Number of reported oligogenic combinations per year and distribution of the associated genetic and functional evidence, based on the data collected in OLIDA, extracted from articles published between January 1994 until December 2023. The colour gradient represents the confidence score (as “absent”, “weak”, “moderate” or “strong” and the corresponding score number from 0 to 3 in parenthesis) for the associated genetic (in blue) and functional (in orange) evidence linked to each oligogenic variant combination in OLIDA.

The main points that we aimed to address in this commentary include the need to better report on variants in general, including a clear description of the variant coordinates and the genes they are found in, as well as proper description of the phenotype of the patients analysed but also of the control individuals, whether they are relatives in the case of a pedigree study or unrelated individuals in the case of cohort studies. We also point out the different considerations that should be taken into account when assessing the pathogenicity of variant combinations related to disease using pedigree studies, statistical studies and functional analyses. These different points are illustrated by examples of reports of oligogenic variant combinations associated with strong levels of evidence in OLIDA.

The aim is therefore to improve the quality of the evidence that links oligogenic variant combinations to disease, by providing researchers with an overview of the necessary considerations for associating such variants to disease, but also to highlight the need for new resources to assist researchers obtain such information, such as pathogenicity prediction methods for combinations of variants. In particular, we highlight the need to be able to search for the frequency of combinations of variants in large population databases such as GnomAD, as opposed to the current availability of monogenic frequencies of variants.

4.8 Database structure, architecture and FAIR implementation

The database is built around six entities: the oligogenic variant combinations, its components which are the variants, the genes, the gene combinations and the diseases, as well as the references which were curated to extract the variant combinations. Each table contains properties and attributes attached to the entity, gathered during the curation and the post-curation process by automatic annotations with several biological databases and tools.

The development of OLIDA consisted of three main aspects. First, a new web portal was built from the ground up with the Django web framework [4] for the Python programming language. Second, the database structure was updated and migrated from MySQL to PostgreSQL. Finally, a REST API was added with open data access standards. The web portal was built in Django to allow for an improved and more flexible long-term maintenance of the website. This also comes with added security and improved performance for data access. We updated the interface of the web portal to use Django's built-in admin interface to make a custom curator interface and add user profiles. In combination with a new submission procedure, this should create a good foundation to make OLIDA into a community-curated data repository.

A new database schema was designed to support oligogenic combinations involving more than two genes, together with other changes that allow for more flexibility to extend the design. An overview of this schema can be found in Appendix B.

The newly created web portal for OLIDA was built to make it Findable, Accessible, Interoperable and Reusable (FAIR) [279]. To make it more findable, metadata was added to the site pages to ensure that keywords such as oligogenic would bring users to OLIDA upon search queries through a search engine. To improve accessibility, the new web portal allows downloading the data in multiple formats and exploration of the data in different ways. This also contributes towards making OLIDA interoperable and ensuring that the data is reusable. The newly added REST API also aims to ensure this. The REST API was built using the Django REST Framework [5] software library. The specification of the API follows the OpenAPI [6] standards. The structure of returned data is in JSON and formatted according to the JSONAPI[7] standard. The API can be explored through a Swagger-UI [8] or Redoc [9] interface. To maximize interoperability, extensive metadata describes the different items and their relationships.

OLIDA is available online at https://olida.ibsquare.be/.

4. https://www.djangoproject.com/
5. https://www.django-rest-framework.org/
6. https://swagger.io/specification/
7. https://jsonapi.org/
8. https://olida.ibsquare.be/api/swagger/
9. https://olida.ibsquare.be/api/redoc/

The different entities that represent a combination are organized in 6 different tables: variant combinations, variants, genes, gene combinations, disease and reference tables. Each table can be accessed through the 'Browse' page of the website, and collects different types of information on the different instances, such as associations with other entities of the database (e.g. associated genes and variants for oligogenic variant combinations), or annotations from external tools and databases (e.g. pathogenic predictions for variants or different gene identifiers for each gene). The content of each column is described in the documentation of the website, and the columns of interest can be selected to be shown or hidden using the 'toggle' function (Figure 4.11B). The full tables with the selected columns can be downloaded directly in the browse page (Figure 4.11C), with the data sorted in ascending or descending order of a particular column (Figure 4.11D) according to what was pre-selected by the user.

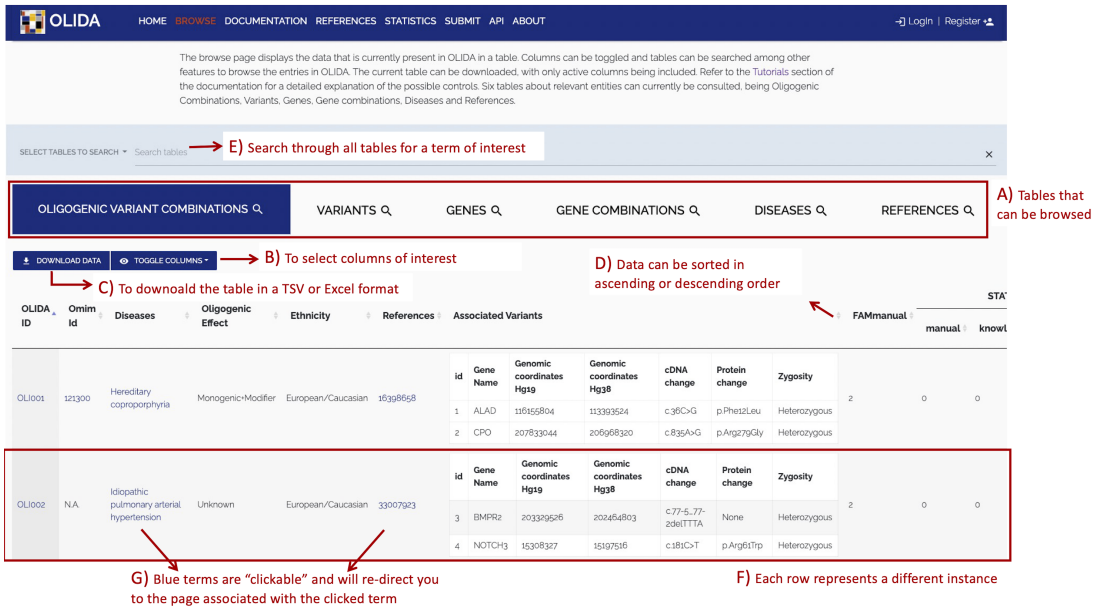


Figure 4.11: Screenshot of the Browse page of OLIDA with the Oligogenic Variant Combinations selected showing the different possibilities that the database offers. Six different tables can be browsed (A) with the currently selected one shown in blue. (B) The user can then select the columns of interest to be displayed in the table and (C) download the table with the selected columns. (D) In a particular table, data can be sorted in ascending or descending order based on a particular column's data. (E) A specific term (e.g. gene name and disease name) can be used to search all tables. (F) Each row represents a specific instance and (G) clicking on specific terms in blue will bring the user to the detail page for that specific instance.

Each row represents a separate instance, which is also attributed a specific page, reached by clicking on the instance identifier (Figure 4.11F). This page provides more details on the variant combination, variant, gene, gene combination, disease or reference of interest (Figure 4.12). Finally, blue terms in the table are clickable (Figure 4.11G) and will redirect the user to the page for that instance (if it is an OLIDA entity) or to the corresponding external ressource (e.g. Reactome for pathways, Orphanet for diseases, etc).

OLIDA follows the FAIR principles on data management, by allowing easy access to data through its API and table downloads, using unique identifiers, enabling links to other other ontologies and providing a thorough documentation. It further allows for user submission through its submission interface.

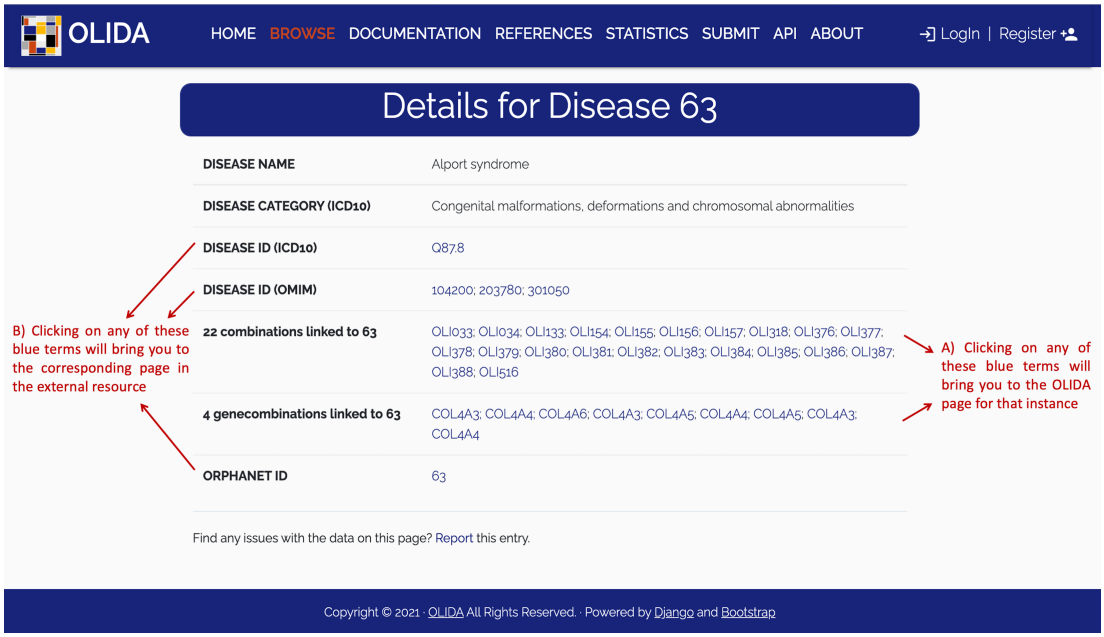


Figure 4.12: Screenshot of the detailed page for Alport syndrome. This page allows the user to visualize in more detail any instance of the database. It provides (A) links between this instance and the other entities of the database, as well as (B) clickable links towards corresponding pages in external databases where information about this entity was retrieved.

4.9 Usage of OLIDA

OLIDA is used actively around the world and the related publication has already been cited more than 20 times according to google scholar's statistics. Figure 4.13 highlights the number of users per country based on google analytics statistics, with the top user countries including the United States of America, Italy, China and Spain. The datasets available in zenodo have also been downloaded several times.

In the citing literature, the database has been mostly used as a reference for showcasing the existence of oligogenic inheritance models for specific diseases [227, 387–389]. More recently, some combinations of the database were used to test new base editor systems able to modify several DNA bases at the same time [390]. These systems are promising for the modelling of digenic combinations in live cells, which will help the functional validation of digenic combinations.

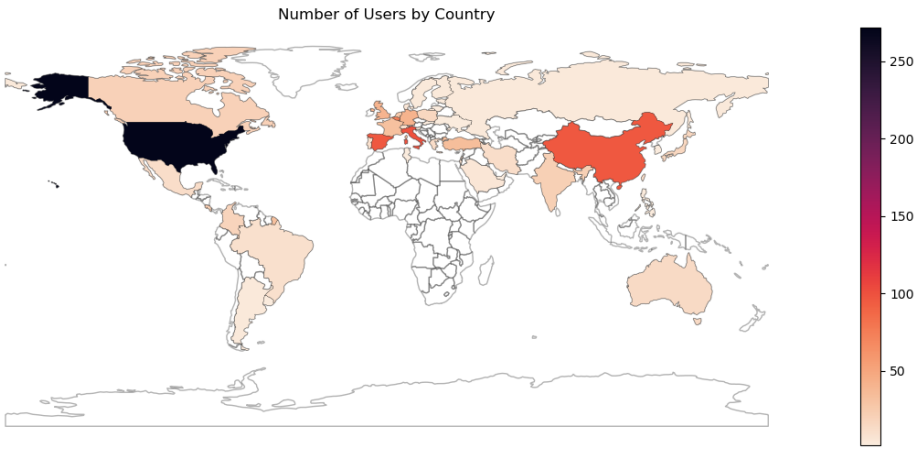


Figure 4.13: Number of cumulated OLIDA users per country between April 2023 and May 2024.

Finally, the database has been used to develop novel computational methods in our research group, including two predictors which are the subject of the next chapter of this thesis [254, 255]. The data has been integrated in context with large biological networks in the BOCK knowledge graph, which has led to the development of a novel approach to predict and explain digenic mechanisms [374]. In an attempt to facilitate the curation process, it has also been used to develop a dataset that can help with automatic detection of such combinations in text [391].

Our commentary article has also received interesting attention, and the recommendations are being used by different researchers reporting on novel digenic combinations [227, 392, 393].

4.10 Conclusion

OLIDA is now the largest and most comprehensive repository of data regarding variant combinations causative of oligogenic diseases. This collection of data regarding variant combinations involving two genes or more paves the way to novel discoveries in the field of genetic diseases. Indeed, as a first generation of digenic predictive tools were developed based on the data collected in DIDA, the step forward into the oligogenic spectrum provided by OLIDA might lead to new predictors able to detect more complex genetic architectures underlying human diseases.

The new database is built on a completely redesigned schema and website, allowing for an easier maintenance and improved access to data, closely following the FAIR principles of data management. With the indexing of the pages in search engines, the existence of an API, the open availability of the data, the use of unique identifiers and links to existing ontologies, as well as an explicit data licence, OLIDA now contributes to open science and the study of oligogenic diseases in the most possible FAIR way. Furthermore, the addition of an interface enabling submission of data by external users provides a first step towards transforming this repository into a community maintained resource for oligogenic diseases. The increase of data and the involvement of the scientific community in their assessment are crucial to advance our knowledge on the synergistic mechanisms and the genetic components behind oligogenic diseases.

One of the main novelty of OLIDA is the introduction of a precise curation protocol for evaluating the evidence associating an oligogenic variant combination to a disease. This curation protocol relies heavily on manual curation, which is then supplemented with external knowledge from biological and clinical databases, and results in the attribution of confidence scores to every combination in the database. These scores are created using structured and clearly defined criteria reflecting the level of evidence supporting the causality of the combination for its associated disease. The confidence metascores can be particularly helpful in allowing the user to assess how confidently a particular combination found in OLIDA is actually linked with its associated disease based on the existence of adequate genetic and functional evidence. Since the number of publications identifying oligogenic causes to disease is increasing, the establishment of clear specialized standards to identify a variant combination as causative of a particular disease, such as the ones described in our protocol [253] and the related commentary article [256], is becoming essential.

This curation protocol assigned 579 out of 1808 (38%) oligogenic combinations with a FINALmeta score of 1 or higher, meaning that these combinations have the minimum required genetic and functional evidence — according to our criteria — to make them at least relevant to their associated disease. The fact that the majority of oligogenic combinations present in OLIDA are not assigned a confidence score above 0 can be attributed in most cases to the absence of sufficient functional evidence, since the description of the functional synergistic

consequences of the involved variants on the disease phenotype and the role of the genes they are located in is lacking or unclear. This is an important aspect of our curation protocol, as it is based on the premise that both genetic and functional evidence showing the synergistic effect of the involved variants and genes should be present to prove causality for a particular disease [121]. The lack of such evidence can be mostly explained by the fact that the vast majority of articles in OLIDA are cohort studies or reports on clinical cases, and, thus, the authors of these articles were mostly focused on obtaining genetic evidence on the variant combinations. Furthermore, during the manual curation, we observed discrepancies in the functional evidence among the different articles describing similar gene combinations as some lacked certain methodologies that were not available at the time of writing, such as the use of variant pathogenicity predictors. Moreover, functional knowledge on gene relationships builds up over time as, for example, most Bardet–Biedl syndrome genes are now known to be involved in the same protein complex [394], whereas this was not the case when the first articles suggesting oligogenic inheritance in this disease were published. This observation motivated our implementation of knowledge and metascores for the combinations, which helped to objectively increase the functional scores of a significant number of instances. Nevertheless, it is evident that important functional knowledge on the synergistic mechanisms among genes, even for those previously reported to be involved in the same disease, is still missing. This is also reflected in the observation that, when looking at the evolution of the reports of oligogenic combinations, the proportion of combinations associated with at least weak levels of evidence is decreasing in all categories except for statistical evidence. As we opened the discussion about new standards in the reporting of oligogenic combinations with our two articles, we hope that this limitation in functional evidence will be addressed in future studies reporting oligogenic cases. It is unfortunately too early to assess the impact of these recommendations.

The curation process is, at the moment, semi-automatic. This means that, with a good team of data curators, maintaining the database up-to-date has been relatively easy. However, this also presents with certain limitation. Indeed, although the annotation and attribution of the knowledge scores are mostly automated, the initial extraction of information is done by curators who must read and discuss each article. Certain annotation parts also require manual input, such as the processing of variant coordinates, for which information is often missing, leading to an overall time-consuming curation process. The development of data mining tools specifically for the collection of data on oligogenic diseases, such as the DUVEL dataset [391], is therefore essential to decrease the time spent extracting information from articles.

Furthermore, the statistics of OLIDA have highlighted the presence of several biases which need to be addressed, or at the very least taken into account when using this database for developing predictive tools. First, the vast majority of combinations in OLIDA are found to be in genes that are known to be involved in the same disease or in relevant pathways for

phenotype (Figure 4.4). Indeed, it appears from the articles that most studies reporting oligogenic combinations performed gene-panel analysis, which means that they only considered variants in genes that are known to be involved in the diseases, or at least in related pathways. Secondly, we observe that in the last 3 years the number of oligogenic combinations reported have nearly doubled, while the number of diseases associated with these combinations has not increased at the same rate. This seems to indicate that the majority of the combinations reported remains associated to the same diseases which are known to be oligogenic, and that not as many discoveries of oligogenic inheritance models are made in other diseases. Both these observations seem to indicate the presence of selection, historical and confirmation bias in oligogenic studies, which have already been noticed in other areas of biomedical research [395–397].

We acknowledge that the criteria used to attribute the confidence scores could also, in itself, introduce a bias towards more closely related genes. For example, a moderate GENEmeta score is assigned to combinations whose genes are involved in the same pathway or are directly interacting without other clear synergistic experiments for the studied phenotype and a weak GENEmeta score is assigned to combinations whose genes are involved in different, but relevant for the phenotype, pathways, without further functional evidence. Nevertheless, it is important to note that this choice of scoring does not imply that only genes that are biologically very closely related can be involved in oligogenic diseases. Indeed, phenomena such as indirect epistasis (i.e. genes being involved in different pathways but that could impact general important metabolic processes, such as signalling or developmental pathways) show that understanding the biological mechanisms behind gene interactions is a complex problem. This choice of scoring rather depicts the fact that the functional evidence described for genes that are more closely related is usually more direct and clearer compared to the evidence for genes suggested to be involved in indirect epistasis. For indirect epistasis, additional functional analyses need to be conducted to demonstrate and clarify the mechanisms involved. If this is the case, as shown with our curation criteria, the GENEmeta score can be strong. This observation further depicts the need for improved functional assays to detect epistasis for such more complex cases.

Despite these limitations, OLIDA has already been widely used across the world, assisting researchers in obtaining detailed information on variant combinations underlying human diseases, and also enabling the development of novel prediction methods, including two models which are the topic of the next chapter of this thesis.

Chapter 5

Investigation into the prioritization of oligogenic combinations

In this chapter, we investigate how to construct a prioritization tool to detect oligogenic causes to diseases. The chapter introduces two novel methods for the pathogenicity prediction of oligogenic combinations. The first method is an improved version of the VarCoPP predictor, which is called VarCoPP2.0 and presents with a higher performance in terms of both prediction quality and computational time. The second method is called High-throughput oligogenic prioritizer (Hop), and is the first prioritization tool directly targeting the detection oligogenic combinations in Whole Exome Sequencing (WES) data. It does so by integrating the predictions of VarCoPP2.0 with a disease-relevance score in order to effectively rank variant combinations based on how likely they are to be causative of a patient's disease.

We first present an overview of the developed method, which was inspired by monogenic variant prioritization tools. In a second step, we detail how we generated synthetic exomes to assess the performance of the prioritization framework, and present statistics on this dataset. We then present the two scoring systems integrated our prioritization approach: first the pathogenicity score, which is obtained by training a new VarCoPP model on an improved dataset and using a simpler model structure; and second the disease-relevance score, which is computed using a propagation algorithm from a set of different seeds in a knowledge graph. Following the description of these two scores, we analyze how to integrate them together to create a prioritization tool, and investigate several characteristics of the ranking produced by the individual scores and the combined ranking score. Finally, we compare the performance of our novel approach with existing prioritization tools, highlighting the necessity of developing a tool specifically targeting variant combinations.

5.1 Motivation and objectives

Over the past few years, there have been great advances in the development of algorithms for understanding the oligogenic architecture underlying genetic diseases. Bioinformatics methods such as VarCoPP and OligoPVP have shown that the detection of oligogenic causes to diseases is now possible, and pave the way towards identifying oligogenic combinations in previously undiagnosed cases. Nevertheless, these methods present limitations, with VarCoPP exhibiting a high false-positive rate, limiting its application to small gene panels, and OligoPVP being restricted by the missing knowledge in PPI data, and thus not able to predict a certain proportion of known oligogenic combinations (see Section 3.5).

In this chapter, we investigate how we can create a novel prioritization method for oligogenic variants, and thus make the detection of bi-locus variant combination in exomes possible. Following from the creation of OLIDA, which provides a larger and more confident ground-truth dataset for digenic combinations, we first develop an improved version of the VarCoPP model. We then investigate how we can use this novel predictor, in combination with phenotype information, in order to create a phenotype-driven prioritizer for variant combinations.

The results presented in this chapter have been published in:

- "Faster and more accurate pathogenic combination predictions with VarCoPP2.0" Nassim Versbraegen, **Barbara Gravel**, Charlotte Nachtegael, Alexandre Renaux, Emma Verkinderen, Ann Nowé, Tom Lenaerts and Sofia Papadimitriou, BMC Bioinformatics 24, 179 (2023). https://doi.org/10.1186/s12859-023-05291-3
- "Prioritization of oligogenic variant combinations in whole exomes", **Barbara Gravel**, Alexandre Renaux, Sofia Papadimitriou, Guillaume Smits, Ann Nowé and Tom Lenaerts, Bioinformatics, 2024, btae184, https://doi.org/10.1093/bioinformatics/btae184

The work on the development of VarCoPP2.0 (Section 5.4) was done in collaboration with Nassim Versbraegen and Sofia Papadimitriou, with my contribution mostly focusing on the literature feature search and annotation, as well as providing feedbacks on the model training and evaluation. The model was re-evaluated on a larger independent set for the work presented in this thesis, which is why the results on that set differ from the results presented in the article.

5.2 General framework of the method

The development of High-throughput oligogenic prioritizer (Hop) is inspired by what is done in the most successful monogenic variant prioritization methods, i.e. integrating disease-related information together with pathogenicity predictions to rank which instances are more likely to cause the patients' disease. In our case, the novelty is that we aim to rank combinations of variants instead of single variants, and we thus want to score the pathogenicity of *variant combinations* as well as the relevance of *gene pairs* to disease.

In order to assess the performance of our predictor, we create synthetic exomes, by inserting known oligogenic variant combinations from OLIDA in exome data from publically available datasets (the 1KGP and UK10K projects described in Section 3.1).

For the variant pathogenicity prediction, we decided to develop a new version of the VarCoPP predictor, which was already very successful at predicting pathogenicity of variants in gene pairs. Now that we have access to a larger and more confident training dataset for digenic combinations, which is provided by OLIDA (see Chapter 4), we can build a novel predictor with improved performance in terms of both predictive power and computational time, which will be essential to predict whole exome data. This predictor will be used to assign a value between 0 and 1 to all combinations in an exome, which will represent the probability that the combination is pathogenic and will be referred to as *Pathogenicity Score (PS)*.

For disease-relevance scoring, we compute the proximity of genes to a set of user-defined seeds, which represent knowledge about the disease carried by the patient. This score is computed using a Random-Walk-with-Restart (RWR) algorithm in a knowledge graph integrating different biological networks. We investigate several types of RWR algorithms and seeds to assess how to best score gene pairs using this approach. This analysis led to the creation of the *Disease-relevance Score (DS)*, which represents the probability that a gene pair is involved in a particular disease.

Finally, we investigate how to integrate the PS and DS to generate a final ranking and analyse the performance of this method to retrieve known oligogenic combinations in synthetic exomes.

5.3 Creating synthetic exomes for performance evaluation

The method takes as input genotype data, in the form of a Variant Call Format (VCF) file (See Section 3.1.3), and information about the patient's disease, in the form of a combination of Human Phenotype Ontology (HPO) terms, describing the patient's phenotype, and a gene panel associated with the disease, or either one of them if only one is available. Since WES data of patients with an oligogenic disease for which the combination underlying the disease

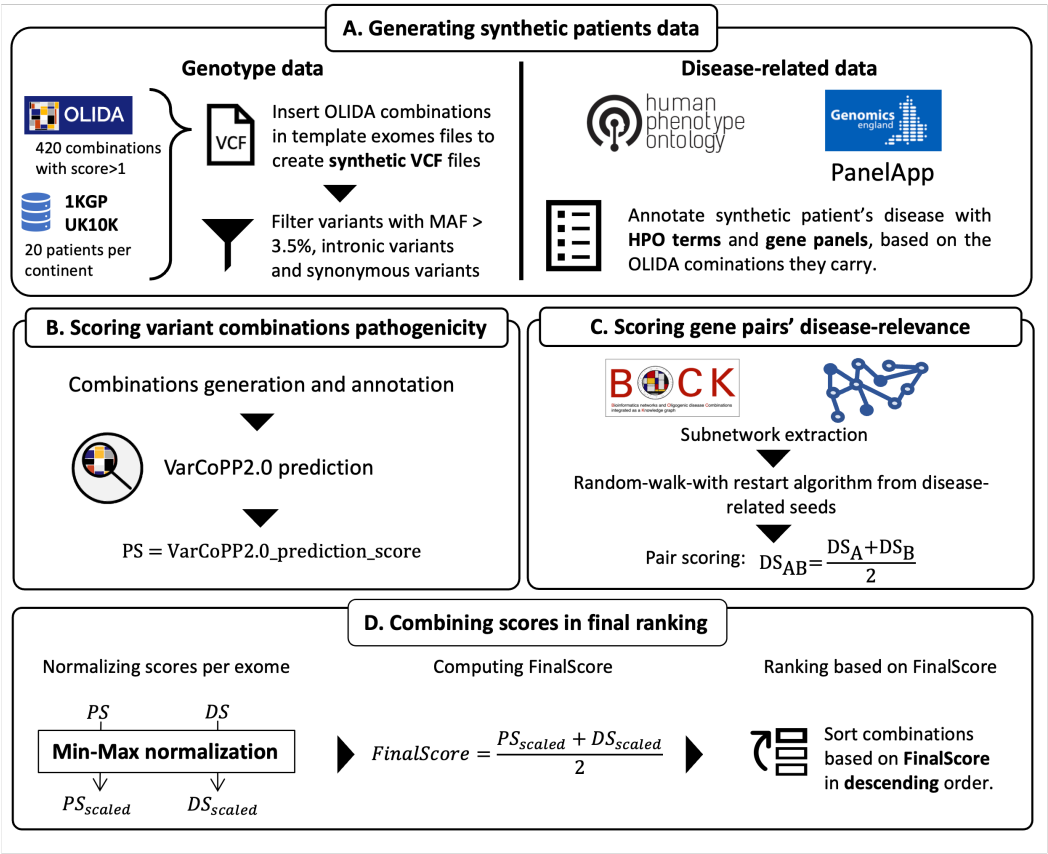


Figure 5.1: General framework of the creation of the Hop predictor. (A) Synthetic exomes are generated to assess the performance of the model, by inserting known pathogenic combinations from the OLIDA database in exomes from the 1KGP and UK10K datasets. Each synthetic patients is associated with data about the corresponding diseases in the form of HPO terms and a gene panel. (B) All possible variant combinations in gene pairs are generated and assigned a pathogenicity prediction score by the VarCoPP2.0 predictor. (C) Each combination is also assigned a disease-relevance score, computed through a propagation algorithm of the disease-related information in a background network. (D) Pathogenicity and disease-relevance scores are normalized across the exome to have equal weight and then combined into a FinalScore that is used to rank all combinations in the exome.

was known are not publicly available, we decided to generate synthetic data, by inserting known oligogenic combinations into exomes of healthy patients and then investigate where the combination is ranked by the predictor. This is a commonly used approach for testing the performance of prioritization methods [171, 194, 398].

5.3.1 Genotype data

The VCF files are generated by inserting oligogenic combinations from OLIDA [253] in exomes from the 1000 genomes project (1KGP) [37] and the UK10K project [264].

OLIDA (v3 May 2023) combinations involving variants in two genes were selected and further filtered based on their confidence scores, considering only combinations with FINALmeta score of at least 1. This resulted in the inclusion of 420 combinations to insert in 1KGP (and UK10K) exomes.

Since some of these combinations (the combinations of OLIDAv1) are used as training instances for the VarCoPP2.0 model used for pathogenicity scoring (see Section 5.4), we divided the 420 OLIDA combinations in a "training set", which comprises only the OLIDAv1 combinations that are used as training instances for VarCoPP2.0 (301 combinations), and a "testing set", which consists of the remaining (unseen) combinations that are used for independent validation (119 combinations).

We selected 100 individuals at random from the 1KGP project, with 20 individuals from each continent, as exome templates for our synthetic exomes. Since VarCoPP2.0 is partly trained on 1KGP data (see Section 5.4), we also included 20 individuals selected at random from the UK10K ALSPAC cohort as additional, completely independent, exome templates. In total, our prioritization method is thus evaluated on 36,120 synthetic exomes containing the training instances (120 templates x 301 combinations) and 14,280 synthetic exomes containing the testing instances (120 templates x 119 combinations). The identifiers of the OLIDA combinations, and of the 1KGP and UK10K individuals are available in the Hop github repository [1] and annotated data for these instances are available in zenodo [2].

The variants of these synthetic exomes were then filtered based on Minor Allele Frequency (MAF) and variant position criteria, following what was done for the training of the VarCoPP2.0 model (see Section 5.4). These variant filters are based on the characteristics of the variants found to be causative of an oligogenic disease as reported in OLIDA. More specifically, we remove variants with MAF $> 3.5\%$, synonymous variants that are further than 195 nucleotides from exon edges, and intronic variants.

1. https://github.com/oligogenic/HOP/tree/main/data/rwr_seeds
2. https://zenodo.org/records/8121283

5.3.2 Disease-related data

When the information was available, each synthetic patient was annotated with both phenotypic data, describing the patient's symptoms, and a gene panel, associated with the patient's disease. The contribution of both types of information for prioritization is analysed in this work.

Phenotypic data is encoded as terms from the HPO [71]. In order to obtain accurate phenotypic annotation that would mimic case descriptions, the HPO terms are extracted from the articles that described the OLIDA combinations. For each OLIDA combination in our training and test sets, we went back to the article that identified the combinations, and annotated each described case manually and with the help of the Text Annotator tool from the Monarch Initiative [297, 298]. Each OLIDA combination, and subsequent synthetic patient carrying that combination, is thus annotated with HPO terms corresponding to the phenotype caused by that combination. The annotations are available in the Hop github repository [3].

Each synthetic patient is assigned a disease based on the disease linked to the OLIDA combination carried by that patient. OLIDA links each combination to diseases from the Orphanet and/or OMIM databases (see Section 3.2.2). To annotate synthetic patients with the relevant gene panels, an extensive search was conducted in the Genomics England PanelApp [80] for Orphanet and/or OMIM diseases associated with the combinations in OLIDA. The association between diseases and gene panel were manually reviewed to ensure accuracy in matching. The gene panels associated with each OLIDA combination are also made available in the Hop github repository [4].

5.3.3 Statistics on the synthetic exomes

We here first present statistics on the number of variants and the annotations associated with the generated set of synthetic exomes.

Statistics on the number of variants in the synthetic exomes

We investigate a few statistics on the variants present in the exome templates used to generate our synthetic exomes, by collecting data on the number of variants, genes and gene-pairs in these exomes, across the different samples and depending on their continental distribution in the 1KGP.

After applying the different filtering steps, each exome template comprised on average of 3,320 variants, located in 2,555 genes, making up 4,540,129 gene pairs and yielding 7,589,030 variant combinations to predict and rank. However, these numbers varied quite a lot across the exomes, with the smallest exome containing 1,619 variants, found in 1,392 genes making

3. https://github.com/oligogenic/HOP/tree/main/data/rwr_seeds
4. https://github.com/oligogenic/HOP/tree/main/data/rwr_seeds

up 1,399,858 gene pairs and 1,718,173 variant combinations (Figure 5.2). On the other hand, sample HG03291, a sample from the 1KGP project of African origin, carried 7,330 variants, located in 5,046 genes and leading to 13,819,024 gene pairs and 32,825,878 variant combinations.

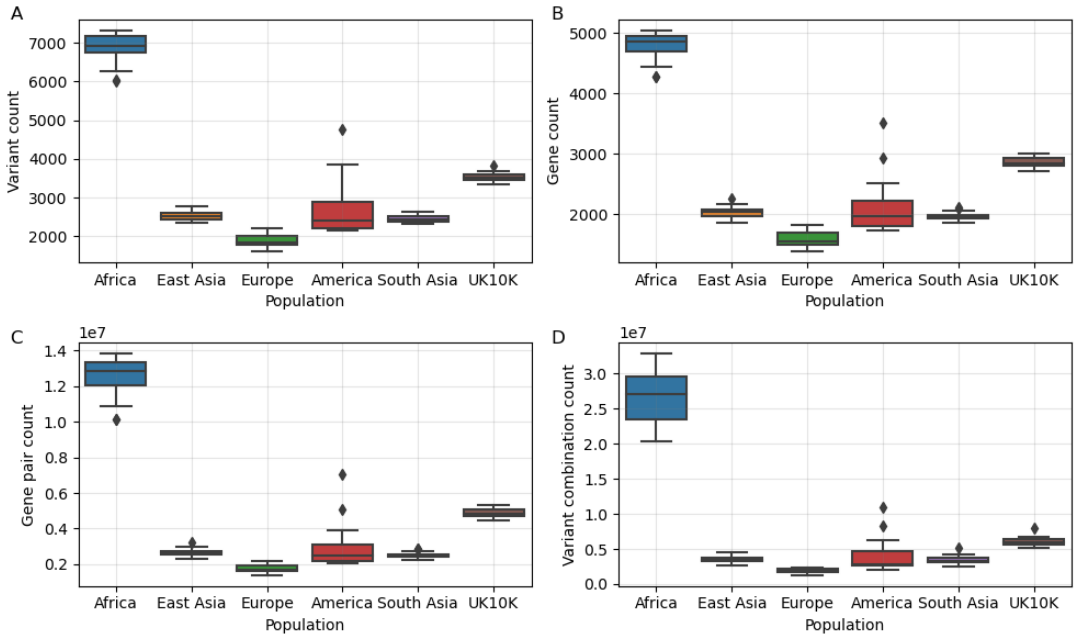


Figure 5.2: Statistics of the number of instances across the different exome templates used to generate the synthetic exomes. Number of variants (A), genes (B), gene pairs (C) and variant combinations (D) for the exomes of the samples from different populations used as templates. Each boxplot represents the distribution over the 20 samples from that population. All continental populations come from the 1KGP samples, while samples from the UK10K project are considered as one population.

Overall, African samples from the 1KGP carried a higher number of variants than the other groups (see Figure 5.2). This led to huge differences in the number of variant combinations to be predicted and ranked in these exomes, due to combinatorial effects, with the median number of variant combinations to be predicted among African exomes being more than ~ 25 million, while it is less ~ 0.5 million in European individuals.

The samples from the UK10K dataset also appeared to have a higher number of variants than the European individuals from the 1KGP dataset, which is probably due to the fact this dataset has been sequenced using a different technology and has been less “polished” over the years than the 1KGP dataset which has been extensively re-analysed [259].

Differential disease-related annotation between the "training" and "testing" exomes

The generated synthetic patients were separated in two groups: "training" synthetic patients, for which the inserted OLIDA combinations belonged to the training set of the VarCoPP2.0 predictor, and "testing" synthetic patients for which the inserted OLIDA combination was independent from the training set of the VarCoPP2.0 predictor. These synthetic patients were then annotated with HPO terms and gene panels to describe the disease associated with the OLIDA combination. Since not all diseases have an associated gene panel, this resulted in a differential annotation between the "training" synthetic patients and "testing" synthetic patients which led to slight differences in the performance of the tool when it was evaluated in cross-validation or in independent validation (Section 5.6. This difference in annotation between these two sets is quantified in Table 5.1.

	HPO terms	Gene Panel	HPO + Panel
	Cross-validation exomes		
Percentage of annotated combinations	100%	72%	100%
Average number of seeds per combination	4.80 ($\sigma = 3.81$)	112.29 ($\sigma = 185.73$)	116.17 ($\sigma = 184.93$
	Independent validation exomes		
Percentage of annotated combinations	100%	87%	100%
Average number of seeds per combination	4.14 ($\sigma = 3.15$)	66.70 ($\sigma = 141.26$)	70.84 ($\sigma = 140.67$)

Table 5.1: Percentage of combinations that are annotated by HPO terms and gene panels and average number of terms/genes associated with each combination for the three types of annotations investigated.

While the HPO terms annotation was similar for the two sets, only 72% of the synthetic "cross-validation" patients could be annotated with a gene panel and 87% of the independent synthetic patients were annotated with a gene panel.

5.4 Pathogenicity scoring: the VarCoPP2.0 predictor

Since we now have access to a larger and more confident dataset on variant combinations causative of oligogenic diseases, we have the possibility to create a new version of the VarCoPP predictor introduced in Section 3.5.2. This new model comprises a new training dataset – including high-quality digenic OLIDA combinations and similar combinations generated from the 1KGP dataset, new features – selected from the literature and then through a feature reduction algorithm, and a novel model structure – chosen from a selection of interpretable model structures able to predict imbalanced datasets. The performance of the new predictor is evaluated in a cross-validation setting and on an independent set of combinations

which appeared in the latest version of the OLIDA database. With its improved predictive performance and decreased computational time, VarCoPP2.0 is perfectly suited to compute the pathogenicity score for prioritizing digenic combinations in whole exomes in the context of Hop.

5.4.1 Creation of novel training and testing sets based on OLIDA

Variant combinations from OLIDA (see Chapter 4) were used as "positive" or "disease-causing" instances and variant combinations found in individuals from the 1KGP were used as "negative" or "neutral" instances for both the training and testing of the model.

Training set

In order to create the VarCoPP training set, we used variant combinations from the first version of OLIDA (OLIDAv1) as positive instances. We only selected the variant combinations involving two genes (i.e. bi-locus combinations) and which had a FINALmeta score of at least 1. This initial filtering resulted in the inclusion of 301 bi-locus combinations as positive instances.

In order to obtain neutral combinations, we used data from the 1KGP. We first filtered the variants so that they had similar characteristics as the OLIDA variants by removing variants which had a MAF $> 3.5\%$ as well as intronic variants and synonymous variants that were not within 195 nucleotides from the exons edges. Even after this filtering, billions of combinations can be selected from 1KGP individuals, leading to a class imbalanced problem. We decided to keep the same class imbalance ratio (1:500) as in the training set of the original VarCoPP predictor, and selected 150,500 variant combinations from individuals of the 1KGP as follows:

1. 301 individuals from the 1KGP were selected randomly but ensuring equal continental distribution (i.e. 60 individuals were selected at random from each continent, based on the project population data).
2. For each individual two mutated genes were selected at random, and for each gene one or two variants were selected at random (allowing for two heterozygous variants, one homozygous or one heterozygous variant to be selected from each gene).
3. This process was repeated 500 times to generate 500 sets of variant combinations with the same number of variant combinations as the OLIDA set, making sure that the same neutral variant combination did not occur twice.

Independent validation sets

In order to validate the performance of the model, we use two different independent sets: a disease-causing independent set, which was used to assess the true and false positive rates of the predictor and a neutral independent set which was used to assess the true and false negative rates of the predictor.

The independent disease-causing set include combinations which were obtained from the latest version of OLIDA at the time of development of the method (OLIDAv3), i.e. variant combinations which were collected during the first two yearly updates of the database. These combinations were filtered in the same way as the training set, keeping only the ones involving two genes, with a FINALmeta score of 1, and which were not present in the training set. This resulted in the inclusion of 119 disease-causing variant combinations in the independent set.

The independent neutral set was constructed in the same manner as the neutral training set. 10,000 variant combinations were selected at random from 1KGP individuals, following an equal continental distribution and selecting only variants which had similar characteristics as OLIDA variants.

5.4.2 Annotation with novel features and feature reduction

The original VarCoPP model used features from the different biological levels to distinguish between pathogenic and neutral combinations. However, some of these features were no longer up-to-date and new variant, gene or pair features could be explored. We therefore first performed a literature search of what computational characteristics could be used and then applied a feature selection procedure to determine which of the pre-selected features had the greater predictive power. Features were selected by searching the literature for variant pathogenicity scores, gene intolerance measures, and measures of genes relatedness. We also looked for features that were proven successful in other methods of oligogenic prediction. We initially annotated the training sets of the new VarCoPP with 20 different features at the variant, gene and gene pair levels. A summary of these features and what they represent can be found in Table 5.2.

Feature	Feature abbreviation	Feature description	PMID	Version
CADD raw score	**CADD1**	CADD score variant 1 of gene A	24487276	CADDv1.6
	CADD2	CADD score variant 2 of gene A		
	CADD3	CADD score variant 1 of gene B		
	CADD4	CADD score variant 2 of gene B		
Haploinsufficiency prediction	**HIPred_A**	HIPred of gene A	28137713	N.A.
	HIPred_B	HIPred of gene B		
Inheritance mode specific pathogenicity prediction	**ISPP_AD_A**	ISPP prediction for AD mode of inheritance for gene A	27354691	N.A.
	ISPP_AD_B	ISPP prediction for AD mode of inheritance for gene B		
	ISPP_AR_A	ISPP prediction for AR mode of inheritance for gene A		
	ISPP_AR_B	ISPP prediction for AR mode of inheritance for gene B		
	ISPP_XL_A	ISPP prediction for XL mode of inheritance for gene B		
	ISPP_XL_B	ISPP prediction for XL mode of inheritance for gene B		
Selection Pressure (dN/dS)	**dN_dS_A**	Selection pressure for gene A	26896847	v99
	dN_dS_B	Selection pressure for gene B		
Biological distance	**Biol_Dist**	Biological distance between gene A and gene B	24694260	v12.2015
Coexpression	Coexp	Coexpression value between gene A and gene B	30462320	CoexpresDBv8.0
Gene ontology similarity	**BP_sim**	Biological Process similarity	10802651 18460186	GO release 2021-12-15
	MF_sim	Molecular Function similarity		
	CC_sim	Cellular Component similarity		
Knowledge Graph distance	**KG_distance**	Distance between gene A and gene B in an in-house developed knowledge graph		N.A.

Table 5.2: Features that were considered for the creation of VarCoPP2.0. Abbreviations as used in the data files for respective features can be found in the feature abbreviation column. The PMID column provides the PMID for the associated relevant literature for each feature. Additionally, the version used in this work are listed in the Version column. Features that are used in the presented model are marked in bold.

Similarly to what was done in VarCoPP, variant combinations are represented by a feature vector assigning values to the different variants, genes and gene pair that compose the combination. For consistency, the genes within a combination are ordered by their tolerance to variation, which is measured using the Residual Variation Intolerance Score (RVIS) [399], with the gene that is least tolerant to variation (and thus with the highest RVIS) ordered first as gene A while the most tolerant gene is second as gene B. Variants within a gene are also ordered based on their impact on the gene, measured using the Combined Annotation Dependent Depletion (CADD) score [161]. The first variant in each gene is the variant with highest CADD score, i.e. the variant perdicted to have the largest impact on the gene.

Variant level features

The variants are annotated with the CADD score as a measure of deleteriousness. CADD is one of the only pathogenicity predictors to score different types of variants including indels (which are present in a large amount of OLIDA combinations), integrates different single variant prediction tools in a meta-prediction and is regularly updated [161]. Variants are scored using CADD version 1.6 GRCh37.

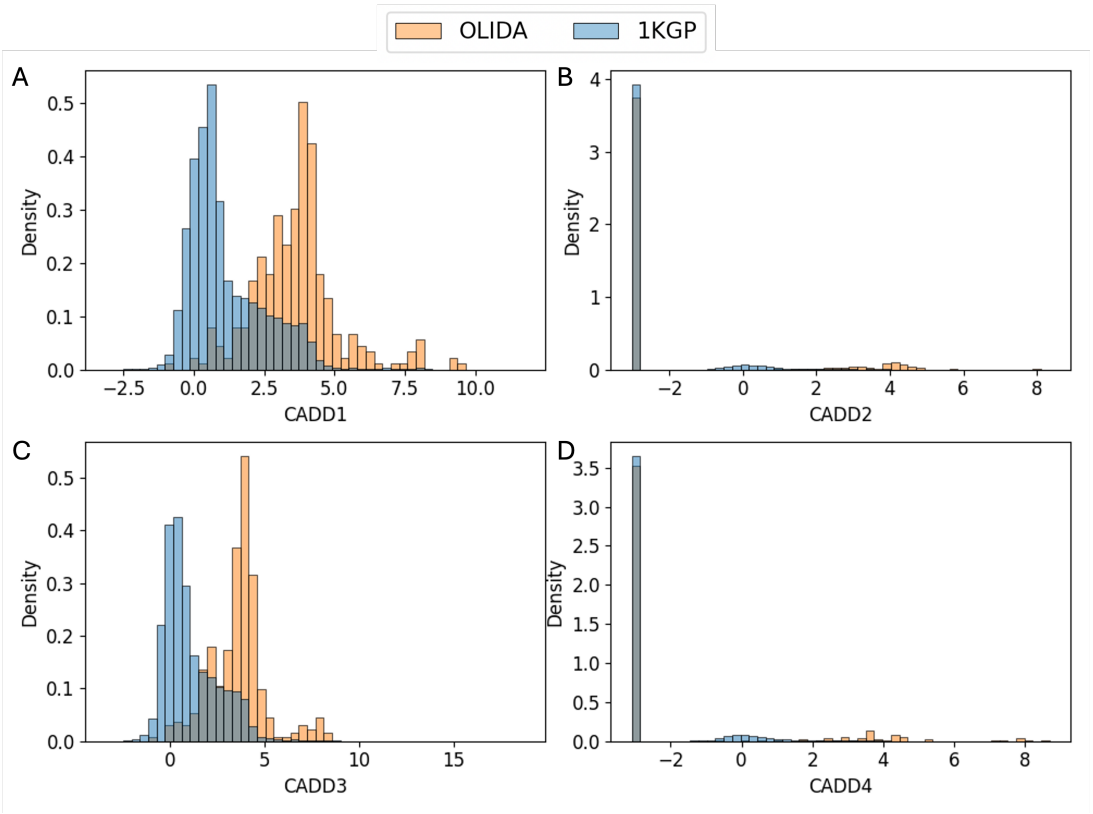


Figure 5.3: Distribution of the CADD1 (A), CADD2 (B), CADD3 (C) and CADD4 (D) feature values in the OLIDA and 1KGP training sets. Each CADD feature corresponds to the raw CADD score of a different variant in each combination, with CADD1 and CADD2 representing the CADD scores of the first and second variants in Gene A, while CADD3 and CADD4 represent the scores of the first and second variants in gene B. The genes of a combination are ordered as gene A and gene B based on their RVIS scores, see Table 5.2.

The distribution of CADD values in the OLIDA and 1KGP sets show that this feature appears to split well the data (Figure 5.3). The CADD values in the OLIDA set appear to be overall higher than in the 1KGP set.

Gene level features

The HIPred [400] predictor is used to annotate genes with a measure of haploinsufficiency. This ML predictor integrates different types of features to predict haploinsufficiency. The scores were downloaded from https://github.com/HAShihab/HIPred.

The genes are also annotated with an evolutionary feature: the dN/dS ratio. This measure quantifies selection pressure by comparing the rate of non-synonymous mutations (dN), which possibly experience selection with the rate of synonymous mutations (dS), which are presumed neutral [401]. dN and dS values were downloaded for each gene in human and 9 different organisms from Ensembl Biomart version 99 and the final value was computed as the mean of the dN/dS ratios in the different organisms.

Finally, the Inheritance Specific Pathogenicity Predictor (ISPP) [402] is used to annotate genes with measures of pathogenicity. This method uses different features and genes that are known to be involved in diseases under specific patterns of inheritance to train a ML model that predicts the probability for each gene to be involved in a disease, under a specific inheritance mode. The three types of inheritance scores available - Autosomal Dominant (AD), Autosomal Recessive (AR) and X-Linked (XL) - are used to annotate the genes. Since genes that are not located on the X chromosome can not be attributed an XL score, their XL score was set to a negative value of -0.5 as this falls outside of the informative feature range (0-1).

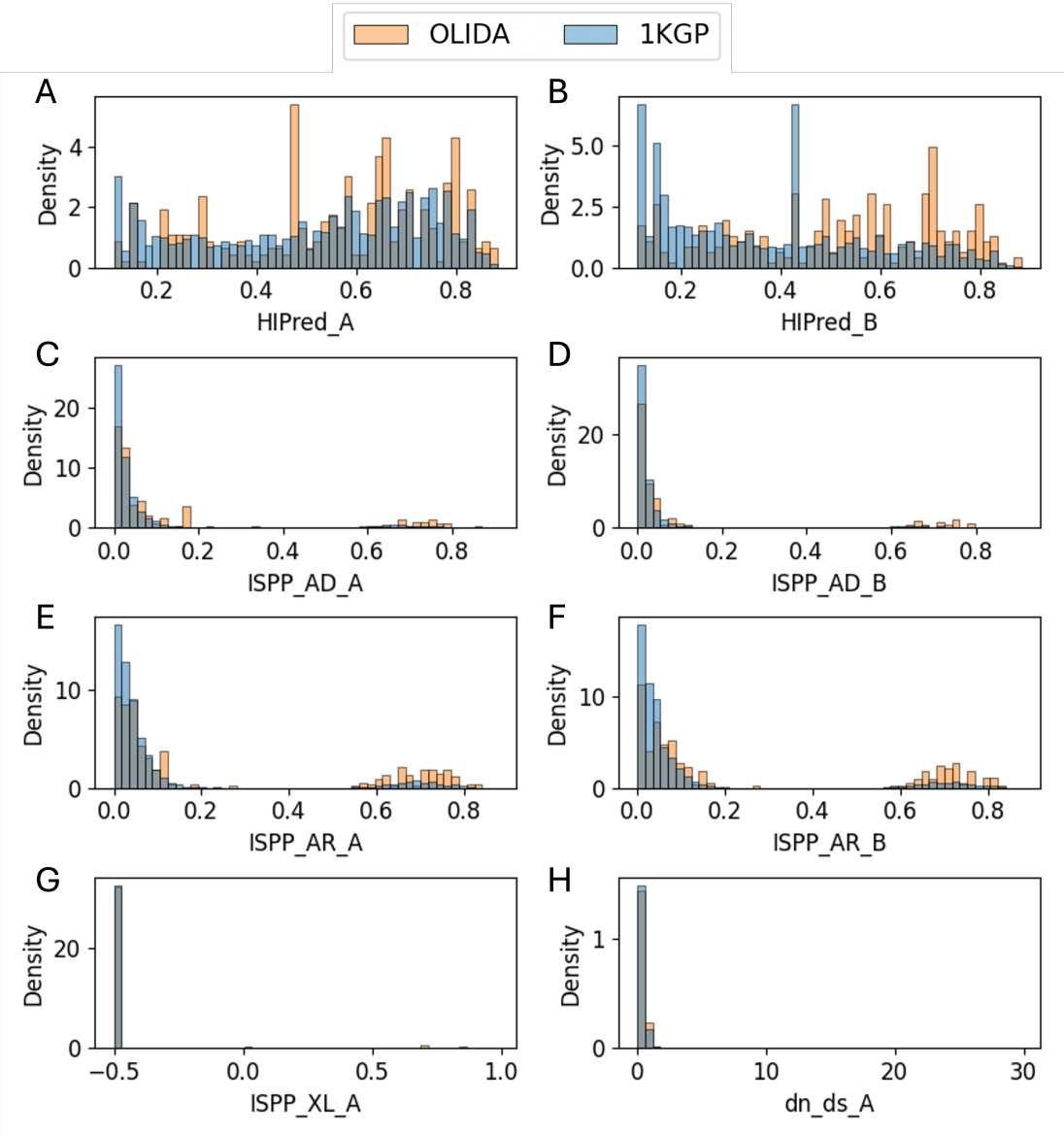


Figure 5.4: Distribution of HIPred_A (A), HIPred_B2 (B), ISPP_AD_A (C), ISPP_AD_B (D) , ISPP_AR_A (E), ISPP_AR_B (F), ISPP_XL_A (G) and dn_ds_A (H) feature values in the OLIDA and 1KGP training sets.

The gene features do not appear to split as strongly as the CADD scores (Figure 5.4), at least on a linear scale. Some differences can be visualised for the ISPP scores, where the proportion of OLIDA combinations with a high ISPP_AR score seems higher.

Gene pair features

Gene pairs are annotated with the biological distance, a feature which was already used in VarCoPP and is a measure of distance between the two genes in a protein-protein interaction network (See Section 3.3.3 and [322]).

A measure of the co-expression between the genes is also used as gene pair feature. Co-expression values were obtained from the COeXPRESsed gene DataBase (CoexpresDB), which integrates data from 23 co-expression platforms and computes the mutual rank as coexpression measure between two genes [308]. Values for human genes were downloaded from CoexpresDB version 8.0.

The Gene Ontology (GO) similarity between two genes was computed for each subgraph of the gene ontology (Biological Process, Molecular Function and Cellular component) independently, leading to three scores of GO similarity. This was done using a software to compute HPO similarity measures[5], which was adapted to use the GO Directed Acyclic Graph (DAG), instead of the HPO one. This software computes different semantic similarity scores (Resnik, SimGIC and Lin), based on the maximum Information Content (IC) among all pairwise comparison of GO terms. The SimGIC measure was chosen as it was shown to outperform other semantic similarity scores [403]. For genes that were annotated with several gene ontology terms, the Best-Match Average (BMA) method, which takes the average of the maximum similarity score for each pairwise comparison, was used as it is shown to provide better results [403]. The simGIC function of the software was thus adapted to compute a simGIC_BMA score.

Finally, we annotated the genes with a distance measured in BOCK [374], which aggregates information about relationships between genes, pathways, gene ontology terms, protein domains and protein families (Section 3.5.5). The distance between two genes is computed as the length of the shortest path between the genes in the graph using the Dijkstra algorithm [323] and then divided by the number of different nodes that are part of the path, in order to take into account the heterogeneity of the graph: e.g. if two genes are part of the same biological process (i.e. they are both connected to the same biological process node in the graph), the length of the path between these genes is 2 and the Knowledge Graph (KG) distance value is 1 since there are two types of nodes (Gene and Pathway nodes) in the path. For two genes that are directly interacting through PPI, the KG distance is also 1, since the length of the shortest path between these genes is 1 and there is only one type of node involved.

5. https://github.com/jeremymcrae/hpo_similarity

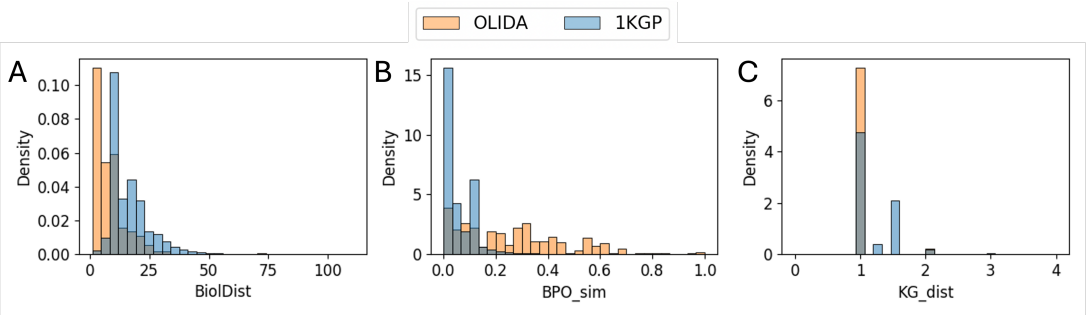


Figure 5.5: Distribution of BiolDist (A), BPO_sim (B), and KG_dist (C) feature values in the OLIDA and 1KGP training sets.

The pair features appear to split the data quite well, with OLIDA combinations showing lower Biological distance and KG_distance, but higher BPO similarity, in comparison the 1KGP combinations (Figure 5.5).

Feature selection procedure

Following this initial selection of feature through literature search, we performed a computational feature reduction procedure to select the most relevant features for the model.

Feature selection among the set of 20 potential features (see Table 5.2) was translated to a heuristic optimisation problem using a wrapper approach (see e.g. [404]). The search was formulated as a relaxed version of the full VarCoPP classification problem, by only considering 301 positive and 301 negative instances to train one random forest (with 100 trees). The performance, assessed by the mean F1 score of a 5-fold cross validation, was the objective function to maximise and all possible subsets of features composed the search space. The step function consisted of inverting the state of a feature (i.e. include it in the set of features used for training if it was excluded and vice versa). At each iteration, 10 random neighbors (i.e. one step removed from the current set) were generated and evaluated. To avoid redundant computation, memoization was used to store and retrieve the result of the performance computation for each considered set. The search was restarted 10000 times. In each, a different random starting point is selected in the search to avoid local maxima. Each search continued until 100 successive search steps did not lead to an improvement.

5.4.3 Model structure selection

An important objective when creating this new model was for it to be less complex than the original VarCoPP predictor, which comprised 500 random forests. We therefore investigated different types of random forests predictors, since they are generally favoured in biological application and are not completely "black-box" predictors. First, different ensembles of RFs (which was used by the original VarCoPP model) were investigated by varying the number of forests and trees within the forests. Second, the use of a single random forest model was also investigated by varying the number of trees in the forest. Finally, we tested the Balanced Random Forest (BRF) [405], a type of random forest specially suited for imbalanced datasets.

For each type of model, the performance was assessed in a 5-fold cross-validation setting using the Average Precision Score (APS). The best model was then selected heuristically, based on its performance as well as relative complexity of the model (for similar performances, a simpler model was preferred).

The performance of the different models investigated is shown in Figure 5.6.

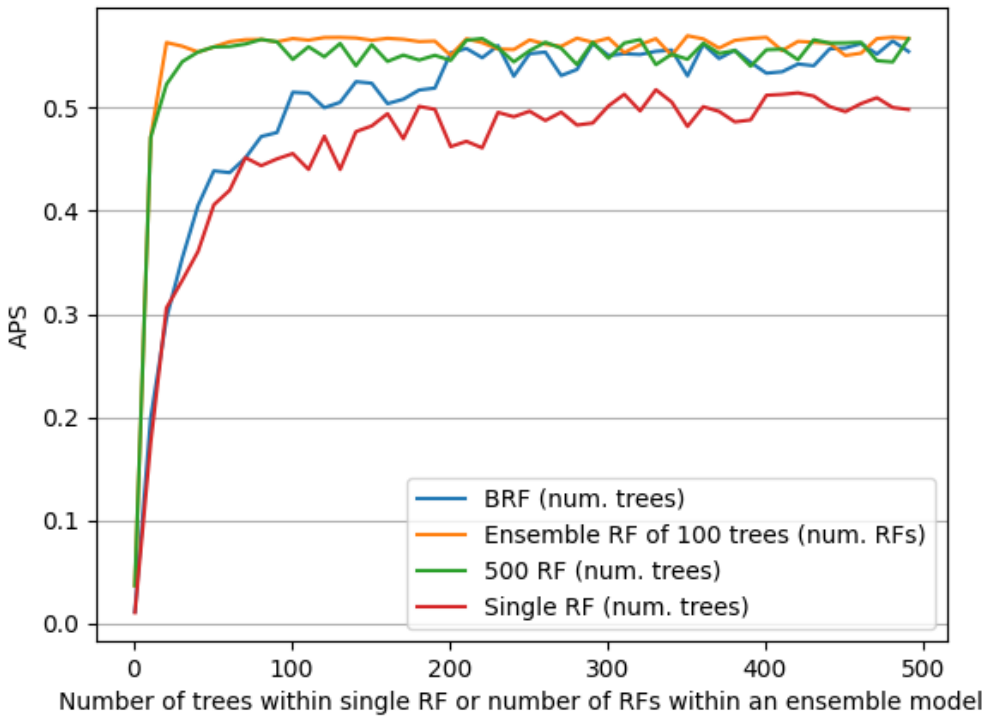


Figure 5.6: Average Precision Score of the different model structures evaluated for the VarCoPP2.0 model. For each model structure, the number of trees (for the Balanced Random Forests, 500 Random Forests, and single Random Forests) or the number of Forests (for the ensemble of Random Forests) is varied between 1 and 500 by increments of 10.

The results show that increasing the complexity of the model (i.e. adding more random forests in the ensemble or more trees in the random forests) does not lead to a big increase in performance. Though the ensembles of RFs perform relatively well with a moderate number of forests or trees, the BRF matches their performance in terms of APS given sufficient trees composing it, i.e. 400 trees (Figure 5.6). After this point, its performance reaches a plateau. On the other hand, the single RF consistently under-performs compared to the other models.

Based on these results, we opted to work with a BRF model with 400 trees, since it leads to a similar or better performance than the other models, while being considerably less complex.

5.4.4 Performance evaluation

The selected model structure for VarCoPP2.0, a BRF with 400 trees, is then evaluated to obtain an accurate assessment of its performance in terms of prediction and computational time. This performance evaluation is used to create confidence zones for the prediction as well as compare the performance of the new predictor to the original VarCoPP model, demonstrating significant improvements, and thus paving the way for its usage in our prioritization setting.

Evaluation in cross-validation and independent test sets

In order to accurately assess the performance of the selected model in the training set, we first perform a leave-one group out (LOGO) cross-validation procedure (see Section 3.4), with gene pairs as stratification groups. At each fold of the cross-validation, a gene pair is selected and all variant combinations which involve that gene pair are removed from the training data and used as test set. In a second step, we also evaluate the performance of the model on an independent validation set consisting of combinations added to OLIDA in the most recent updates of the database (up to OLIDAv3).

In both cases, we compute the Receiver Operating Characteristic Curve (ROC) and Precision-Recall (PR) curves as well as several performance metrics based on the confusion matrices which are shown in Table 5.3 (see Section 3.4.4 for a description of the metrics).

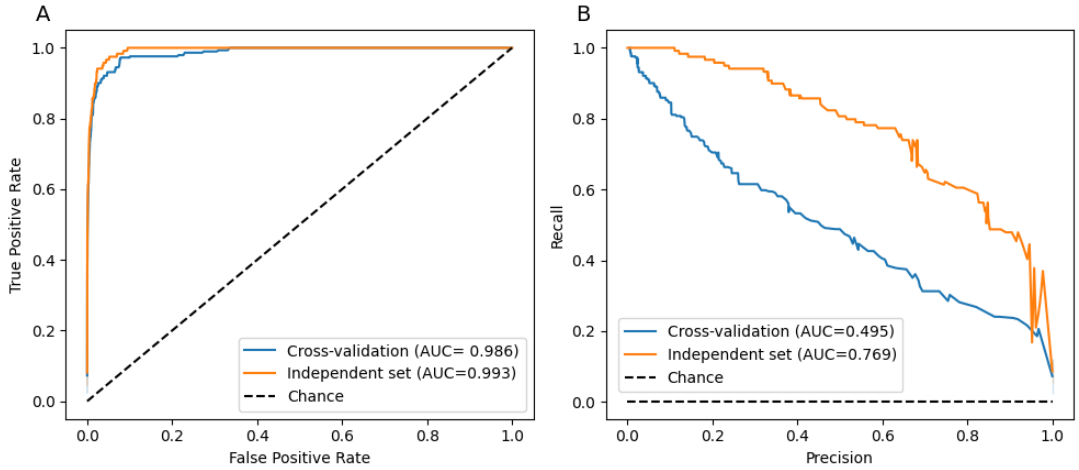


Figure 5.7: ROC curves (A) and PR curves (B) evaluating the performance of the VarCoPP2.0 predictor in a LOGO cross-validation (blue curves) and independent test set (orange curves). The black dotted line represent the performance of a random classifier. The area-under-the-curve measure is shown in the legend.

		Predicted	
		Disease	**Neutral**
Actual	**Disease**	268	23
	Neutral	7003	143506

(a) Cross-Validation

		Predicted	
		Disease	**Neutral**
Actual	**Disease**	115	4
	Neutral	483	9525

(b) Independent Validation

Table 5.3: Confusion Matrices for Disease-causing and Neutral classes in (a) Cross-validation and (b) Independent validation sets.

The cross-validation results (LOGO) indicate that VarCoPP2.0 classifies bi-locus variant combinations with high accuracy, achieving a TP rate of 0.95 and a FP rate of 0.05 (see Table 5.3a). The computed ROC-AUC of 0.986 and the PR-AUC of 0.495 (Figure 5.7) demonstrate that the predictor performs very well in this cross-validation setting.

As can be observed in Table 5.3b and Figure 5.7, VarCoPP2.0 also shows great sensitivity on the independent set, being able to classify 115 out of 119 OLIDA combinations as disease-causing, and mis-predicting only 4.83% of the neutral combinations, similarly to its performance during the cross-validation. The sensitivity of VarCoPP in this setting is particularly important, as in medical genetics it is crucial to be able to detect the pathogenic variants if these are present, and indeed causative, in the patient. A ROC-AUC of 0.993 and a PR-AUC of 0.769 is obtained in the independent validation setting (Figure 5.7), showing superior performance to the cross-validation setting. This difference in performance between the cross-

validation and independent setting is probably due to the fact that data leakage was more controlled in the cross-validation setting since the combinations in the test set did not involve the same gene pairs, while this is not necessarily the case for the independent test set. Indeed, the combinations included in the test set were simply selected for being reported in a later version of OLIDA, and as we have seen from that there is bias present in OLIDA (See Chapter 4), it is very likely that the combinations in the independent set involve some common gene pairs with the training set, and are thus easier to predict.

Comparison with the performance of VarCoPP

The performance of the VarCoPP2.0 model was compared with the performance of the original VarCoPP predictor (trained on the training set containing DIDA combinations) and to a predictor with the original VarCoPP structure, thus including the initial set of features, but trained on the new training set including OLIDA combinations as positives (referred to as VarCoPP OLIDA). The performance of these 3 models was evaluated in the independent validation set. The results of the comparison (Table 5.4), show that although the inclusion of the new OLIDA data leads to a slight improvement in performance, it is the combination of the new features and model structure of VarCoPP2.0 that enables the model to achieve a much higher performance.

The results indicate that the new model outperforms the original VarCoPP predictor, accurately classifying all but four of the independent pathogenic variant combinations and achieving a FP rate of about 5% in both cross-validation and independent sets. In comparison, the first version of the VarCoPP predictor misclassified 13 out of 119 independent combinations, and presented a FP rate of 7 to 8%.

		Predicted					
		VarCoPP		**VarCoPP OLIDA**		**VarCoPP2.0**	
		Disease	**Neutral**	**Disease**	**Neutral**	**Disease**	**Neutral**
Actual	**Disease**	106	13	109	10	115	4
	Neutral	770	9230	473	9527	475	9525

Table 5.4: Confusion matrices showing the performance of the original VarCoPP, the original VarCoPP trained using OLIDA instances (VarCoPP OLIDA) and the VarCoPP2.0 predictor in the independent validation set.

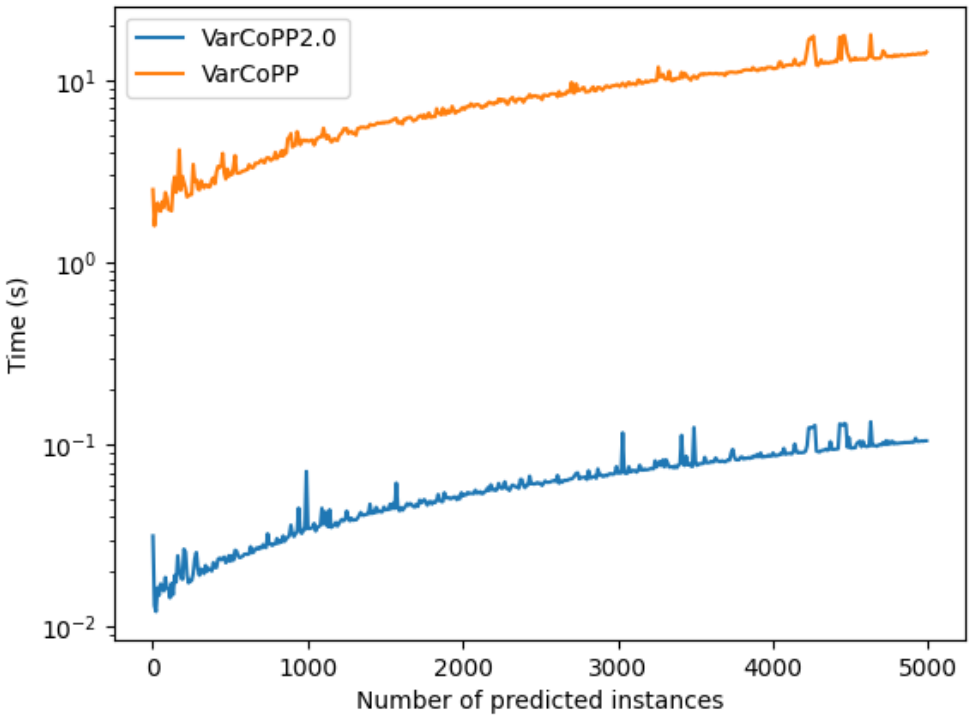


Figure 5.8: Execution time required to predict a certain number of instances (varied between 1 and 5000) using the VarCoPP and VarCoPP2.0 predictors. The time is measured in seconds and is shown in log scale for easier comparison of the two models.

An important improvement of the VarCoPP2.0 model is the reduction of complexity of the model which results in a strong decrease in prediction time, especially when a large number of bilocus combinations are being predicted. This was measured by computing the prediction time for both VarCoPP and VarCoPP2.0 on the same set of 10,000 variant combinations. For the original VarCoPP, the time for predicting 10,000 digenic combinations is 47.5 ($\sigma = .49$) seconds, while VarCoPP2.0 predicts the same samples in 0.08 ($\sigma = .005$) seconds.

This difference in computational time is especially important for the use of the predictor with exome sequencing data, which can include millions of variant combinations to predict.

5.4.5 Creation of confidence zones for predictions

Similarly to what was done in the original VarCoPP, we define confidence zones based on the performance of the predictor on the neutral independent set. These confidence zones are used to identify which combinations predicted as disease-causing are less likely to be false positives. 99% and 99.9% confidence zones were defined, by selecting the minimum VarCoPP2.0 score such that the amount of variant combinations in the neutral independent with a higher VarCoPP2.0 score is less than the confidence level. A threshold of 0.743 is used for the 99% zone and a classification threshold of 0.891 is used for the 99.9% zone, in which only 0.1% variant combinations are expected to be FP.

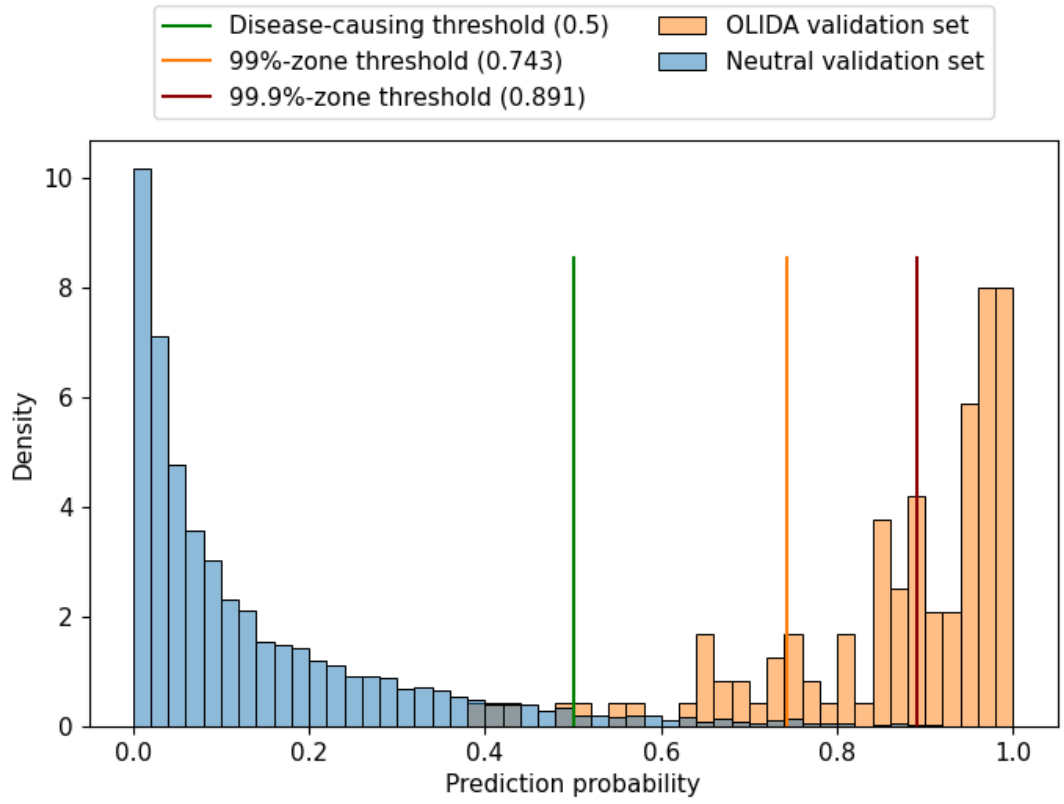


Figure 5.9: Distribution of the VarCoPP score in the OLIDA validation set and the Neutral validation set with the thresholds for the different confidence zones shown in green (threshold for pathogenic prediction), orange (threshold for the 99% confidence zone, where there is only 1% of expected FPs) and red (threshold for the 99.9% confidence zone where there is only 0.1% of expected FPs).

Application of these confidence zones in the independent positive set shows that 67 out of 119 (56%) of the disease-causing combinations are in the 99.9% zone, while 98 out of 119 (82%) are in the 99% confidence zone, indicating that the known disease-causing variant combinations are generally predicted with high scores.

5.4.6 Feature importance analysis

In order to further examine the contribution of the chosen features for predictions and the potential existence of any bias, the individual feature importance of the trained model was examined using the Mean Decrease in Impurity (MDI) measure. A Gini importance boxplot is presented for each feature over all trees of the BRF in Figure 5.10, providing a global model interpretation. This interpretation reveals that CADD1 (i.e. the CADD score of the first variant allele of gene A), CADD3 (i.e. the CADD score of the first variant allele of gene B), Biol_Dist and BPO_sim are the most important features for prediction.

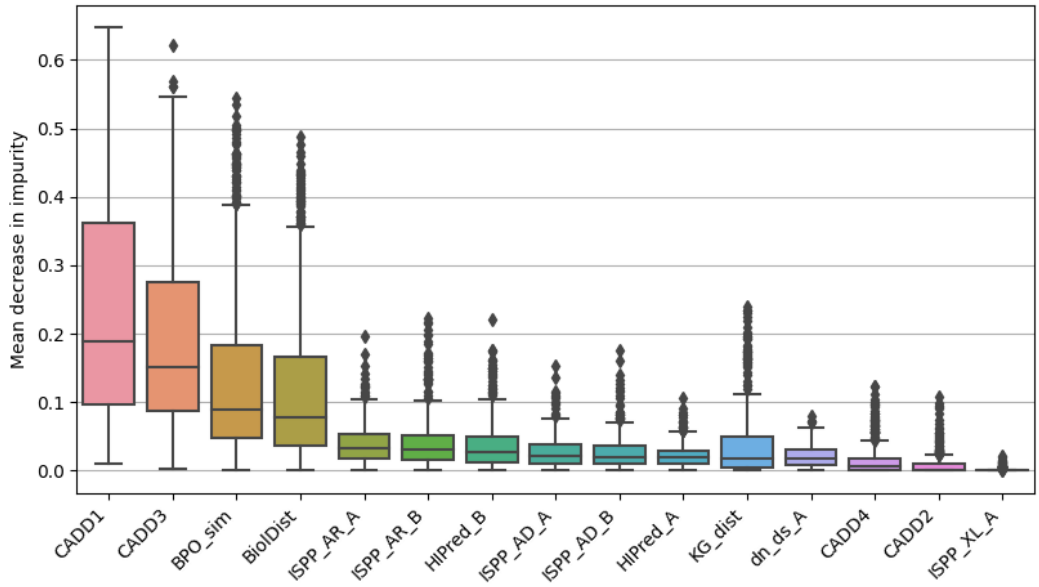


Figure 5.10: Mean decrease in impurity measure of feature importance across the 400 trees of the Balanced Random Forest model. Features are ordered based on the median feature importance across all trees.

Compared to the original version of VarCoPP, we observe that although variant-related features, such as the CADD1 and CADD3, remain two of the most important drivers for predictions, all gene pair features in VarCoPP2.0 (Biol_Dist, BP_similarity and KG_dist) emerge as the second most important predictive feature group. This demonstrates that VarCoPP2.0, compared to the original VarCoPP, uses more the biological relatedness information between genes and that these features contribute to the better performance of the model. This difference is caused by the fact that the neutral training set of the original VarCoPP model was stratified using the degrees of separation between genes (i.e. the number of proteins connecting the protein products of the genes in a pair inside a protein-protein network), a decision that had made VarCoPP less sensitive to gene-pair related information differences between the neutral and the disease-causing set.

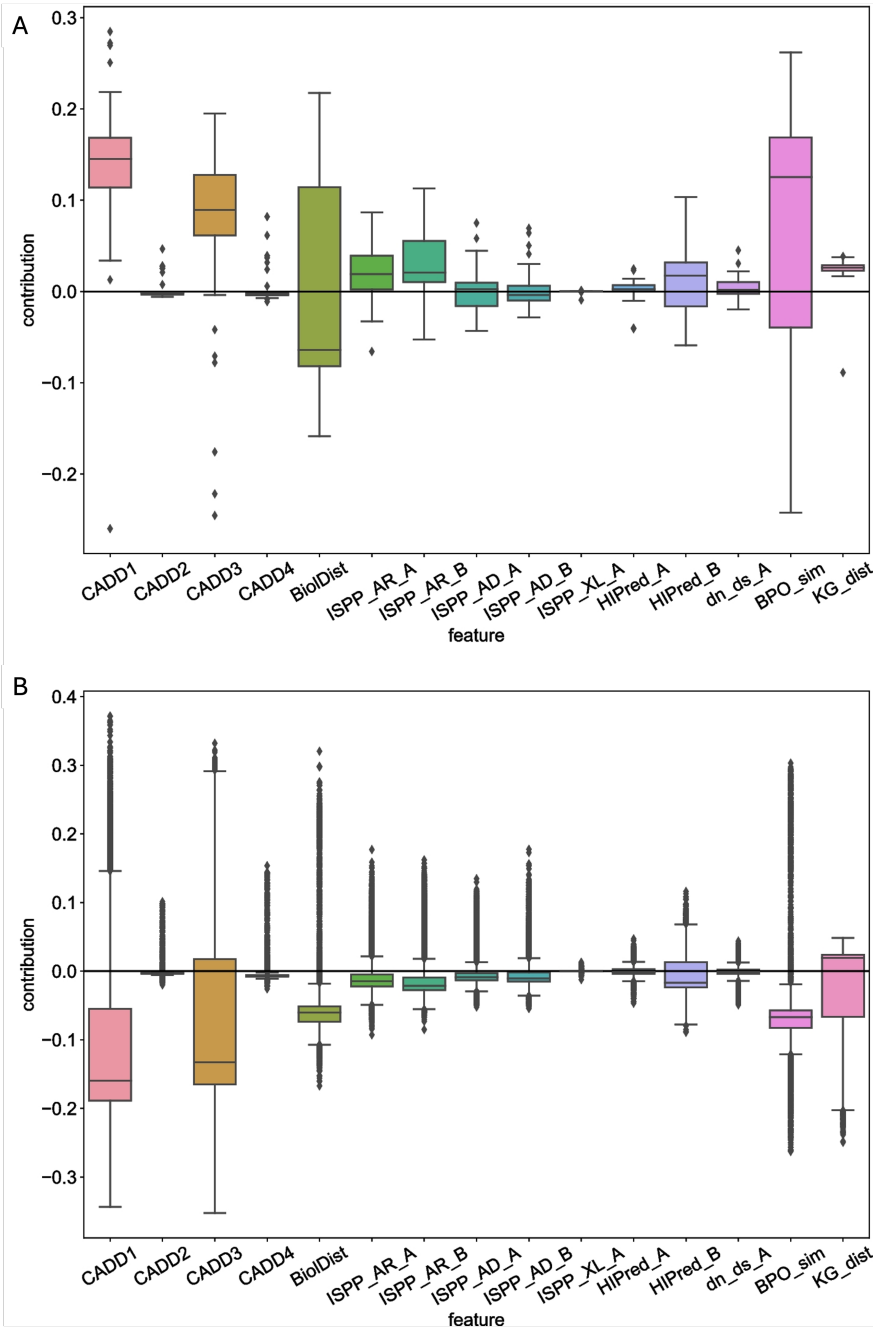


Figure 5.11: Boxplot of the feature contributions the disease-causing or the neutral class, among all positive instances (A) and all negative instances (B) in the validation set inferred using treeinterpreter. A feature contribution value above 0 indicates a vote for a positive prediction (i.e. towards the disease-causing class), while a value below 0 indicates a vote for negative prediction (i.e. towards the neutral class). The more the feature contribution value deviates from 0, the stronger the vote is for either class

Moreover, we examine local model interpretations inside the model (i.e. on specific instances), using *treeinterpreter*[6]. This package separates each prediction into a fixed bias value and contribution values linked to each of the used features, specific to a particular instance being examined. We observe that the local model interpretations are consistent with the global model interpretation for both the positive and the negative data set, in the sense that CADD1, CADD3, Biol_Dist and BPO_sim seem to drive the predictions (or, vote) the most in the negative direction (i.e. towards the neutral class) for the negative instances and towards the positive direction (i.e. towards the disease-causing class) for the positive set. Noteworthy is that variation seems relatively high in both BP_similarity and Biol_Dist contributions for the positive instances (Figure 5.11).

The distribution of the feature values in the positive and neutral sets, as shown in Section 5.4.2, gives us visual insights into why these features are found to have higher importance. Indeed, the CADD1 and CADD3 scores seem to exhibit a quite different distribution between the disease-causing and the neutral sets, with the disease causing variants having higher CADD scores (Figure 5.3). This difference in distribution is also observed, although less strong, in the Biol_Dist and BPO_sim features (Figure 5.5), with the gene pairs from the disease-causing set having smaller distances and higher similarity in comparison to the neutral instances.

In this section, we presented a new ML model for the prediction of the pathogenicity of variant combinations within gene pairs: the VarCoPP2.0 predictor [254]. We investigate in Section 5.6, how the score outputed by this predictor, which we refer to as *Pathogenicity Score (PS)*, can be best combined with disease-relevance scoring to prioritize oligogenic variants.

5.5 Disease-relevance scoring via network propagation

In this section, we investigate how to score the disease-relevance of gene pairs, using the BOCK knowledge graph as background knowledge. More specifically, we investigate the performance of the Random-Walk-with-Restart (RWR) algorithm (see Section 3.3.3), for scoring the proximity of gene pairs to a set of seeds, which can be provided by the user as either HPO terms, describing the patients phenotype, or a set of genes which are known to be involved in the disease of interest, i.e. a gene panel. For the research here, we remove from BOCK the oligogenic interaction nodes and their links to the diseases and genes nodes (as this is what the method aims to detect). Moreover, as random walks will be performed on this KG we do not take into account the edges weights, which results in considering the KG as an unweighted, multiplex-heterogeneous network with 8 layers (see Section 3.3.1 for definition of this network terminology).

6. https://github.com/andosa/treeinterpreter

5.5.1 Defining a disease-relevance score for gene pairs

The RWR algorithm is a network-based method that simulates a random walk starting from a source node, with a probability of returning to the source at each step, to prioritize or rank nodes based on their proximity to the source node (see Section 3.3.3). The output of the RWR algorithm is an approximation of the vector p^{∞}, which contains the probability of reaching each node in the network, when performing a RWR from a set of pre-defined seed nodes. These probabilities can be interpreted as a proximity measure to the set of seeds.

By performing a RWR using HPO terms or a gene panel as seeds, we therefore obtain the proximity of each node in the graph to the set of seeds linked to the disease of interest. This is used for example in Exomiser [170] to score the phenotype-relevance of genes. For all gene nodes, this value is defined as the phenotypic relevance score of the gene.

In order to obtain a measure of gene-pair disease-relevance, we investigated 2 ways of combining the individual gene phenotype relevance scores: multiplying and summing the individual gene scores. Taking the sum of the scores seemed to lead to slightly better performance, and the Disease-relevance Score (DS) for gene-pair score was thus defined as follows:

$$DS_{AB} = p_A^{\infty} + p_B^{\infty}$$

where p_A^{∞} and p_B^{∞} are the values of vector p^{∞} at the nodes corresponding to gene_A and gene_B respectively, with p^{∞} being the output vector of the RWR algorithm run using the seeds defining the disease of the patient.

In the following sections, we investigate how different RWR algorithms, parameter settings, and seeds are efficient at ranking OLIDA gene-pairs based on disease relevance, using BOCK as network. This was done using the training synthetic exomes described in Section 5.3, and only assessing the ranks of the gene pairs: all possible gene-pairs of the synthetic exome gene list are scored based on the different phenotypic-relevance scoring measures, and the rank of the OLIDA gene-pair is computed. The results are visualized as Cumulative Density Function (CDF) curves (see Section 3.4.4), which plot the proportion of synthetic exomes for which the OLIDA gene pair is found in the top K, for varying values of K.

5.5.2 Investigation of different graphs and types of propagation

The subnetwork from BOCK is a multiplex-heterogeneous network (genes are connected in a PPI network, a coexpression network and a sequence similarity network, and other types of nodes such as GO terms are also present), and we therefore investigate 3 different types of RWR algorithms: a classic RWR algorithm, for which the adjacency matrix of the full “flattened”

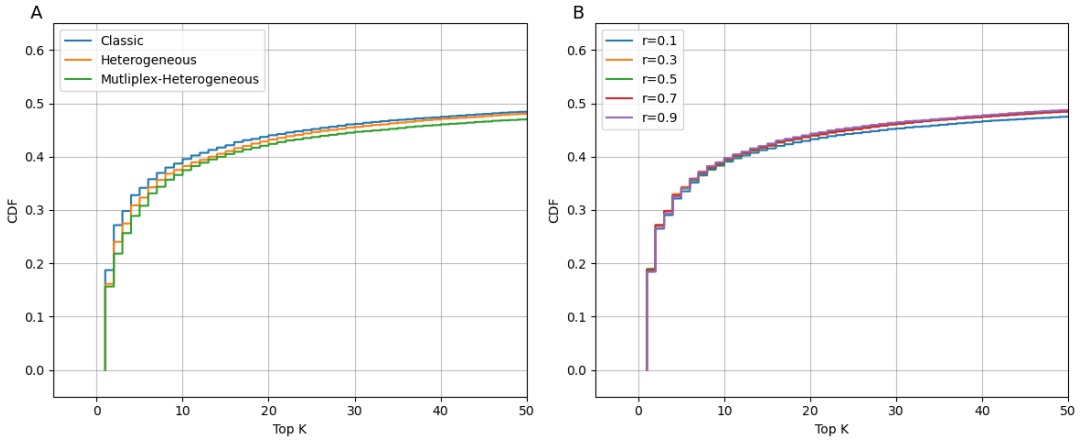


Figure 5.12: CDF curves of the rank of OLIDA combinations in training synthetic exomes, based on the three different types of RWR algorithms for scoring the proximity of gene pairs to the phenotypic annotations (A), and based on different parameter values for the "classic" RWR algorithm with HPO terms as seeds (B).

graph was used as transition matrix, a heterogeneous RWR, for which the transition matrix reflects the heterogeneous nature of the graph but the multiplex layers of the gene-gene network are flattened, and a multiplex-heterogeneous RWR, for which the transition matrix reflects the heterogeneous and multiplex nature of the graph (see Section 3.3.3).

In this comparison to investigate the best RWR algorithm, we only used the HPO terms associated with the synthetic exome data as seeds. Equation 3.4 was computed iteratively until change between p^t and p^{t+1} fell below 10^{-9}.

The results, shown in Figure 5.12 A, indicate that the different RWR algorithms yield very similar ranking of the OLIDA combinations, with the classic RWR algorithm performed by flattening the different layers of BOCK actually showing the best performance when compared to the other algorithms.

In order to further assess the potential differences between the rankings performed by the different types of RWR, we computed the Kendall-Tau correlation on the whole gene-pairs ranking between the different types of RWR. The correlation coefficient is 0.78 ($\sigma = 0.03$) between the RWR classic and the RWR heterogeneous rankings, 0.71 ($\sigma = 0.03$) between the RWR classic and the RWR multiplex-heterogeneous rankings and 0.71 ($\sigma = 0.04$) between the RWR heterogeneous and the RWR multiplex-heterogeneous rankings. This confirms that although the rankings are different, they are also strongly correlated and that no single algorithm appears to outperform the others.

Based on these results, we decided to pursue the work using the classic RWR algorithm, since it required a simpler matrix and seemed to show slightly improved performance (Figure 5.12A).

5.5.3 Investigation of the effect of the restart parameter

The RWR algorithm relies on only one parameter, r, which controls the spread of the walk in the network. We therefore evaluated the influence of this parameter on our results by reproducing our results with the restart probability varying between 0.1 and 0.9 with intervals of 0.2.

The results, presented in Figure 5.12B, show that this parameter has very little effect on the percentage of exomes for which the known OLIDA combination is ranked in the top K when K is lower than 50. This is consistent with the conclusions of other works on RWR algorithms in biological networks, which have shown that except for extreme values, this parameter did not have a strong effect on the output ranking [174, 332, 333]. Except for a restart of 0.1, which seems to lead to slightly decreased performance, the curves based on all other restart parameters seem to overlap. Based on these results, we decided to set the restart to 0.7, as done in most biological applications of RWR algorithms, and then further investigate the effect of the parameter on the full ranking (integrating both pathogenicity score and disease-relevance score) in Section 5.6.

5.5.4 Investigation of different seeds and seeds quality

In order to assess whether using other information sources as seeds for the algorithm can be useful in prioritizing the combinations, we test three different types of seeds, as well as combining different seed types as the starting point of the algorithm. We therefore investigate the use of HPO terms, gene panels and GO terms as seeds for the algorithm. Gene panel and HPO annotations are described in Section 5.3. For GO annotation, the HPO2GO predictor was used [406], based on the seed HPO terms, to obtain GO annotations for each gene pair. It is important to note that not all OLIDA gene pairs could be annotated with all types of seeds, since for example some diseases do not have associated gene panels, or some diseases are not associated with any pathway.

We also investigate using the combination of HPO terms and gene panels as source since they were the seeds that lead to the best rankings individually (see Figure 5.13A).In that case, all HPO terms and genes in the panel are used as seeds, and if no panel is associated to a particular disease, the propagation is based only on the HPO terms.

The results indicate that among individual seed types, HPO terms are the best type of seeds to use as starting point for the RWR algorithm, but integrating both HPO terms and gene panel information of the disease lead to improved performance (Figure 5.13A). The extremely low performance of GO terms as seeds is likely to be due to the fact that this source of information was indirect (the GO terms were not associated to a disease or phenotype but predicted as linked to HPO terms), leading to imprecise prior information and thus to poor performance.

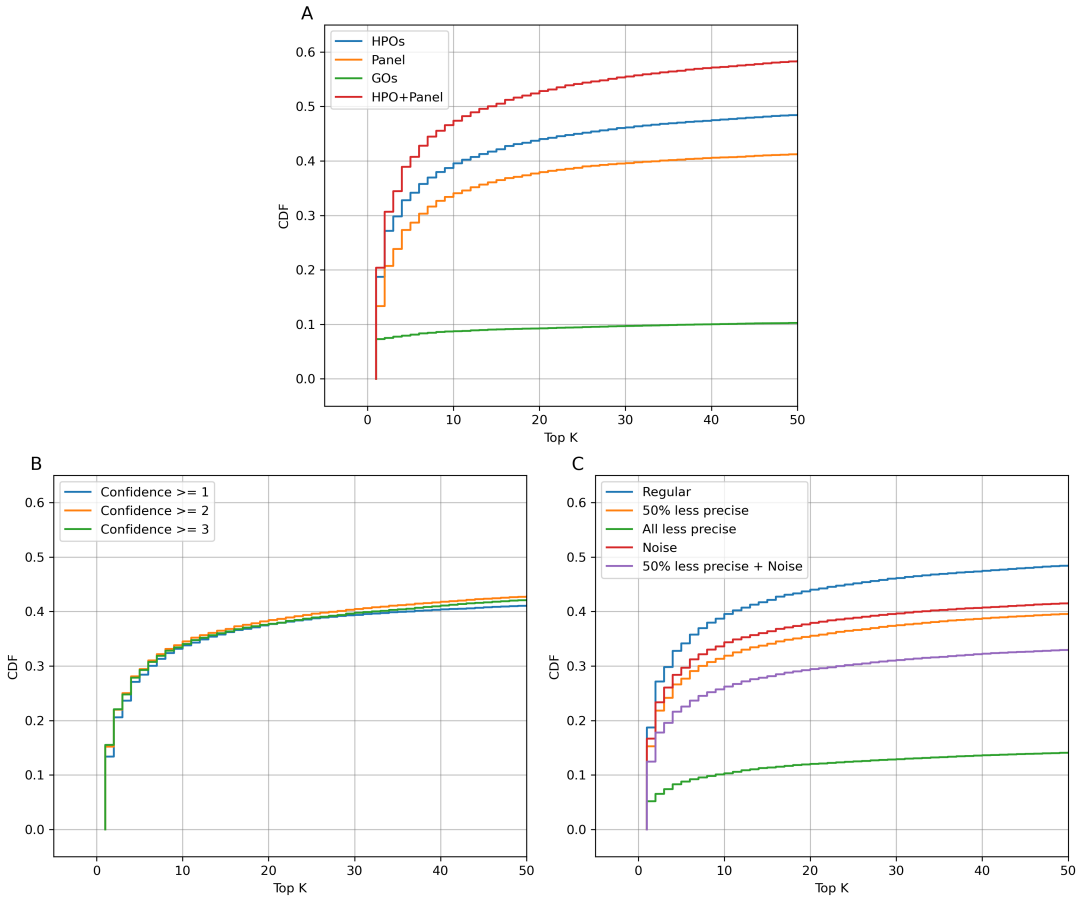


Figure 5.13: Cumulative Density Function (CDF) of the ranking of OLIDA gene pairs in the synthetic training exomes using different types of seeds as the starting point for the RWR algorithm.

Having identified gene panels and HPO terms as the most efficient seed types for scoring disease relevance, we investigated whether using annotations of different quality for both these seeds might have an effect on the ranking.

For the gene panel annotations, we looked at the different levels of confidence associated with the panels in the Genomics England PanelApp, and compared the use of genes associated with a confidence level of at least 1, at least 2 and at least 3. For the HPO terms, we compared the use of the manually curated HPO annotations that we generated from the papers describing the phenotype of the patients to 4 different sets of HPOs:

- A 50% less precise set, where half of the HPO terms of each combinations are replaced by the parent HPO term in the DAG of the ontology.
- A set where all HPO terms are less precise, where all HPO terms from the original set are replaced by their parent term.

- A set with noise, where random HPO terms were added to the patient HPO terms (half the number of real HPO terms are added).
- A set with noise and imprecision, where half the HPO terms are replaced by their parent terms and a random set of half the number of original HPOs are added.

The results for the performance of the algorithm depending on the quality in seed annotation are shown in Figure 5.13B for the gene panels and Figure 5.13C for the HPO terms. We notice that the difference in quality of the gene panel does not seem to strongly affect performance (Figure 5.13B).The best performance is obtained when using gene panels associated with a confidence of at least 2. This is probably due to the fact that the panels with confidence of at least 1 contain "noise" in the genes that are not associated with sufficient confidence with the disease, while the gene panels with a confidence of 3 are too restrained to only genes with high levels of confidence.

On the other hand, the ranks based on the HPO annotations are highly sensitive to imprecision and noise. The method appears to be particularly sensitive to imprecision, as we observe a 4-fold reduction in performance when replacing all the seed HPO terms by their parent term. On the other hand, we only observe a ~ 1.3 fold decrease in performance in the top 10 when comparing the ranking based on regular annotations and the ranking based on annotations with extra noise (Figure 5.13C).

The fact that the ranking based on gene panel annotations is less sensitive to annotation quality is probably due to the fact that the number of seeds in the panel annotation is much larger than HPO annotations (see Table 5.1) and that small changes in these sets are therefore less likely to affect the output of the RWR algorithm. We conclude that keeping both sources of information since although the HPO terms appear to lead to better performance, this source of information is also less robust.

5.5.5 Comparison with simpler measures of similarity

Finally, we compare the results obtained with our RWR-based disease relevance measure with simpler methods of similarity scoring (which do not involve graph propagation). We thus test three different measures of similarity, each based on the Jaccard index, to rank the gene pairs based on their similarity to the set of seed HPO terms. The Jaccard index is defined in Equation 3.3 in Section 3.3.3.

In our case we want to assess the similarity between the HPO terms associated with a gene pair AB and a set of HPO terms describing the patient's disease D. We therefore investigate three measures of similarity based on the Jaccard index:

$$JaccSim_1 = J(A,D) + J(B,D)$$

$$JaccSim_2 = J(A \cap B, D)$$

$$JaccSim_2 = J(A \cup B, D)$$

where A and B are the set of HPO terms associated with the genes A and B respectively, based on the HPO annotations, D is the set of HPOs used to annotate the patient's symptoms and $J(A,B)$ is the Jaccard index between sets A and B.

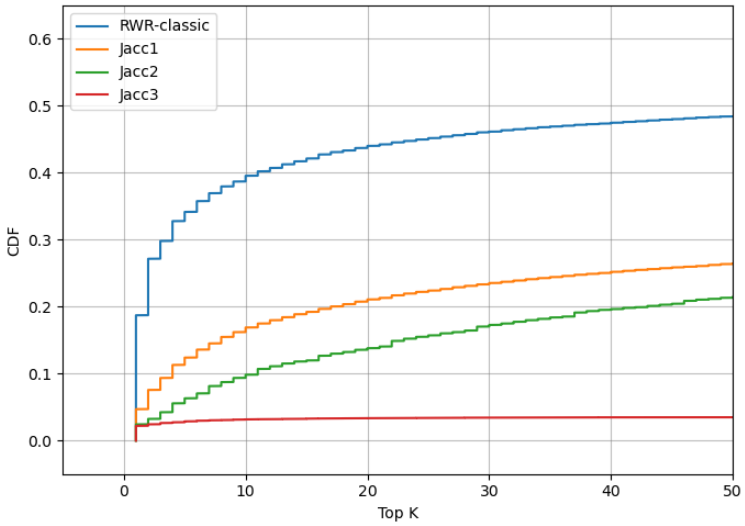


Figure 5.14: Cumulative Density Function (CDF) of the rank of OLIDA gene pairs in synthetic training exomes based on different types of similarity measures for ranking. The RWR similarity measure is based on a "classic" random walk with restart algorithm in the full graph, while the three other measures are based on different ways to integrate the Jaccard similarity measure to assess the similarity between the HPO annotation of a gene pair and the seed HPOs describing the patient's symptoms.

The CDF for the different Jaccard similarity measures is shown in Figure 5.14 with the CDF obtained when using the "classic" RWR in the whole network for comparison. The results indicate that the RWR scoring clearly outperforms the scoring based on Jaccard index, which is expected, since the RWR relies on all prior-knowledge about gene-gene interactions, phenotypic similarities and gene-HPO associations, while the Jaccard index based measures only make use of the HPO annotations of the gene of interests. These results thus justify the use of the RWR algorithm for computing the disease-relevance score and confirms that the different networks are indeed used by the RWR algorithm, and that the propagation is not solely limited between the HPO terms and gene-HPO associations.

Following from these investigation into disease-relevance scoring for gene pairs, we defined the *Disease-relevance Score (DS)* as follows: it is computed using the "classic" RWR algorithm with HPOs, Gene panel and the combinations of both as seeds, using a restart parameter of 0.7 and is computed as the sum of the individual gene scores. In the next section, we investigate how integrating this score together with the PS can effectively rank variant combinations.

5.6 Combining pathogenicity and disease-relevance scores into a final ranking

In the third part of the development of Hop, we investigate the performance of the whole predictor, integrating the disease-relevance score for gene-pairs based on the RWR algorithm and the pathogenicity score from VarCoPP2.0.

Different operators were tested to combine the DS and the PS in a *Final Score (FS)*. We evaluate the use of the minimum of the two scores, the maximum of the two scores, the average of the two scores and the multiplication of the scores (Figure 5.15A).

Except for the maximum of the two scores, all other operators appeared to perform similarly well. The average of the PS and the DS seemed to slightly outperform the other operators and was thus used to compute the FS .

The fact that the maximum of the two scores did not perform as well as the other operators is probably due to the fact that this operator does not put any minimum condition on each of the two scores. This leads to any combination with any one of the two scores as high to be positioned highly in the ranking (even combinations that are predicted as highly pathogenic but in genes that are not relevant to the disease of interest or vice-versa) and therefore will not put emphasis on ranking highly the combinations that are both “pathogenic” and relevant to the disease.

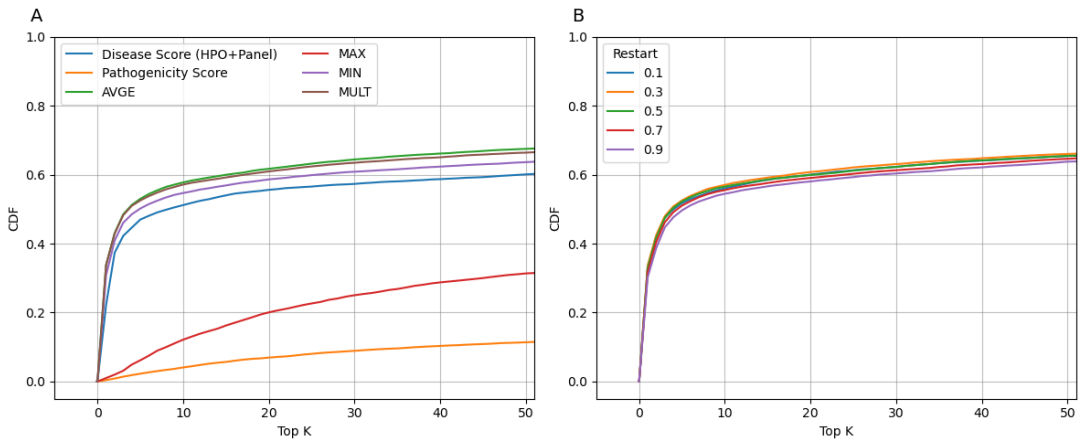


Figure 5.15: Cumulative density function plot of the rankings obtained in the cross-validation exomes, with the HPO and gene panel seeds for the pathogenicity and disease scores and for the four different operators tested to combine the two scores (A), and using the "average" operator for 5 different values of the restart parameter (B). The CDF plots illustrate the proportion of exomes for which the OLIDA combination is ranked in the top k for k varied between 1 and 50 (inclusive).

For the final predictor, we therefore opted for the average of the two scores as the final score used for ranking.

We further re-analyse the effect of the RWR restart parameter on the final ranking (although it was already done on the computation of the DS in Section 5.5) since it is the only parameter to tune in the full model. As before, we do not observe much difference in the performance for the different values of the restart (Figure 5.15B). We therefore set the restart parameter to 0.3 for the rest of the analyses as it seems to perform slightly better.

5.6.1 Combining VarCoPP2.0 with disease-information is essential for efficient prioritization

We first assess how efficient each type of score is for ranking combinations in exomes. To visualize the performance, we plot the Cumulative Density Function (CDF), which shows the proportion of exomes for which the inserted oligogenic combination is in the top K for varying values of K. We present results for K ranging up to 50, as we believe this is the maximum number of combinations a user would realistically consider. The results over the whole range of rank values are analysed in Section 5.6.4. We compare the ranking performance of using the PS alone, i.e. the VarCoPP2.0 predictor, the DS alone, i.e. the disease-related information, and the combination of the two scores (FS). The results of this analysis are shown in Figure 5.16, for the cross-validation on the training exomes, and in Figure 5.17 for the validation on an independent test set of exomes.

Using, on the one hand, only PS, less than 13% of synthetic exomes had the known pathogenic OLIDA combination in the top 50 of prioritized variant combinations (Figure 5.16 and Figure 5.17, green), indicating, as expected, that the VarCoPP2.0 predictor alone is not sufficient for prioritizing digenic combinations in exomes. On the other hand, using only the DS based on HPO terms, 52% of the exomes used in cross-validation and 57% of the exomes in the independent set had the OLIDA combination ranked in the top 50 prioritized combinations (Figure 5.16 and Figure 5.17, orange). By using FS, combining both PS and DS with HPO terms, 61% of the cross-validation exomes and 70% of the validation exomes contain the known OLIDA combinations in the top 50 (Figure 5.16 and Figure 5.17, blue). These results indicate that both PS and DS must be combined to achieve the highest performance.

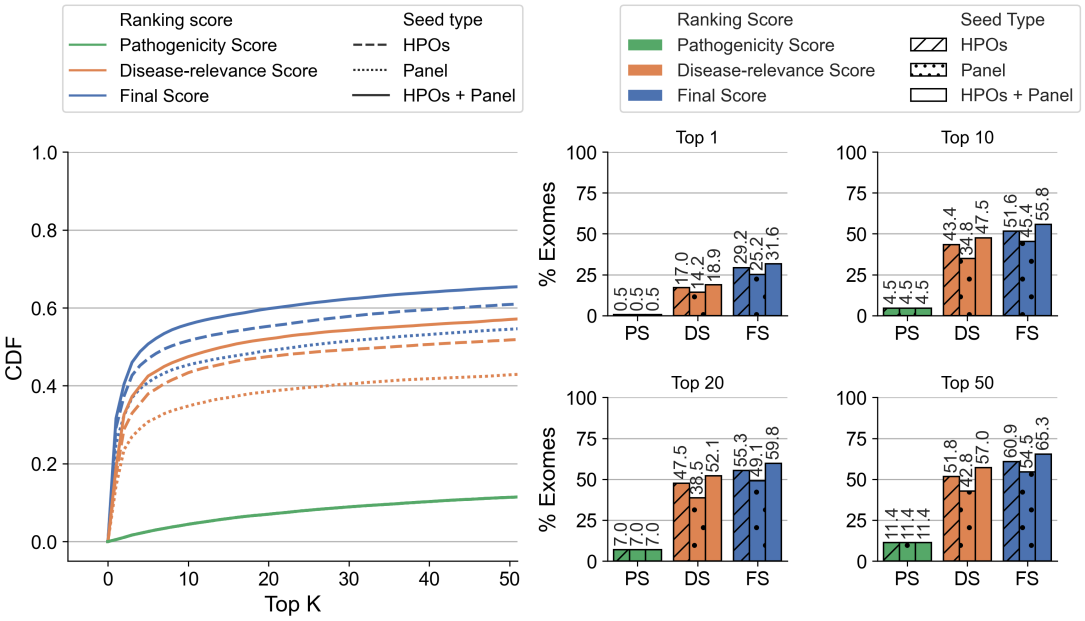


Figure 5.16: Performance of Hop in the cross-validation exomes. (A) Cumulative Density Function (CDF) plot of the rankings obtained in the cross-validation exomes, by using the FS (blue), DS (Orange), and PS (green) as ranking scores, with HPO terms as seeds (dashed line), genes from a gene panel as seeds (dotted line) and both HPOs and gene panel as seeds (solid line). The CDF plots illustrate the percentage of exomes for which the OLIDA combination is ranked in the top K by each method, with K varying between 1 and 50 (inclusive). (B) Percentage of exomes for which the known OLIDA combination is ranked in the top 1, top 10, top 20 and top 50 of the cross-validation exomes based on the different types of scores and seeds for ranking.

5.6.2 Different sources of disease information improve ranking

We investigate here how the different sources of prior knowledge on the disease can lead to different ranking performances by comparing the use of different seed types for the computation of the DS score. HPO terms have been commonly used as prior information in state-of-the-art prioritization methods, providing granularity in the description of a patient's disease and enabling phenotypic similarity profiling to several diseases, which can lead to the discovery of new gene-disease associations. Notwithstanding their success, other sources of prior information could be useful. Here, we investigate whether gene panels, which are readily used in clinics to assess the possibility of the genetic origin of a disease, provide useful prior information for ranking.

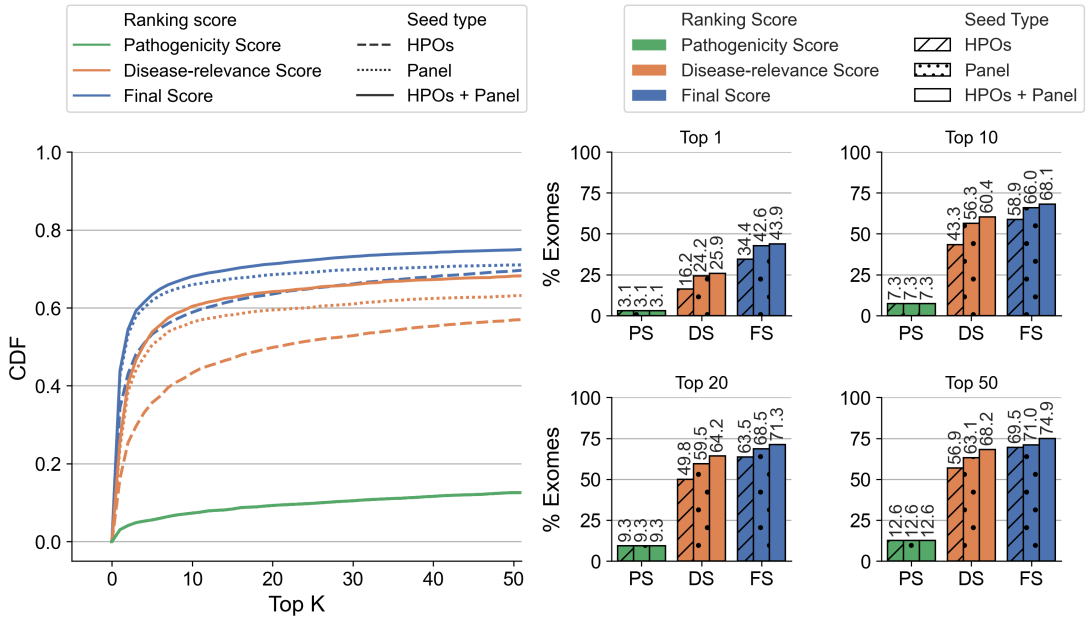


Figure 5.17: Performance Hop in the independent validation exomes. (A) Cumulative Density Function (CDF) plot of the rankings obtained in the independent validation exomes, by using the Final Score (FS), Disease-relevance Score (DS), and Pathogenicity Score (PS) as ranking scores, with HPO terms as seeds (dashed line), genes from a gene panel as seeds (dotted line) and both HPOs and gene panel as seeds (solid line). The Cumulative Density Function (CDF) plots illustrate the percentage of exomes for which the OLIDA combination is ranked in the top K by each method, with K varied between 1 and 50 (inclusive). (B) Percentage of exomes for which the known OLIDA combination is ranked in the top 1, top 10, top 20, and top 50 of the independent validation exomes based on the different types of scores and seeds for ranking.

Our results show that, consistently with what was investigated in the creation of the DS, using both gene panels and HPO terms as seeds for the prioritization algorithm always leads to an improvement in performance when compared to the use of either type of information alone. In cross-validation, 64.6% of the known pathogenic combination are identified in the training exomes in the top 50 when both HPO terms and gene panels are used as starting points for the algorithm, as opposed to 60% when only HPO terms are used and 53.6% when only gene panel information is provided (see Figure 5.16). In the independent validation, the OLIDA combinations are found in the top 50 combinations in 74.3% of the tested exomes when both types of information are used while this percentage reduces to 68.9% when HPO terms are used alone and of 70.7% when gene panels are used alone (see Figure 5.17).

In the cross-validation exomes, using HPO terms as prior-information seems to perform better than using gene panels (Figure 5.16), while the converse is true in independent validation exomes (Figure 5.17). This difference can partially be explained by the difference in annotation between the combinations in the training set and in the independent validation set. Only

72% of the combinations belonging to the training set could be annotated with a gene panel compared to 87% of the combinations in the test set, which contributed to making this type of information less useful as a seeding strategy in the first case (see Table 5.1). On the other hand, since HPO terms are used to describe patients' symptoms, all combinations can be annotated with this type of information, which can explain why the performance when using HPO terms appears to be more consistent between the cross-validation and independent exomes.

5.6.3 Analysis of the contribution of each score to the ranking

We further investigate the influence of the two individual scores, DS and PS, on the selection of top-ranked combinations, which suggest that the DS score provides a slightly stronger contribution to the prioritization of combinations than the PS score; but that the top-ranked combinations are chosen based on a combination of the scores rather than relying solely on one of the scores alone.

In order to do this, we decide to focus on the combinations that are ranked at the top, since these are the combinations that are considered to be the most relevant and which will require further investigation by the clinician. The objective is to determine whether one of the individual scores has a stronger impact on the selection of the top-ranked combinations. We thus analyze how many of the top-ranked combinations (in 4 different tops) are also highly ranked by the individual scores. This is done for the rankings using both HPO terms and gene panels as prior information since these seeds lead to the highest performance.

The results show that very few combinations with the highest FS are also highly-ranked by both individual scores (Figure 5.18). When comparing the overlap between the highly prioritized combinations based on the FS and the DS score, we find that there is a larger overlap than when compared to the PS score. This suggests that the DS score provides a slightly stronger contribution to the prioritization of combinations than the PS score. However, it's important to note that the number of overlapping combinations between any individual score and the FS is quite low, indicating that the top-ranked combinations are chosen based on a combination of the scores rather than relying solely on one of the scores alone.

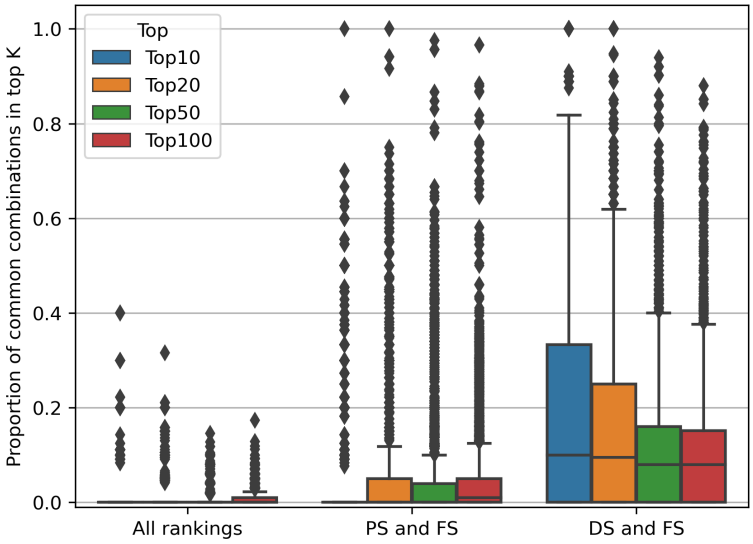


Figure 5.18: Proportion of common combinations at the top K of the rankings generated by the different scores for K values of (10, 20, 50 and 100). All rankings represent the proportion of the combinations in the top K of the ranking based on the FS that are also in the top K of the combinations based on the PS and in the top K of the combinations based on the DS. PS and FS represent the proportion of the combinations in the top K of the ranking based on the FS that are also in the top K of the combinations based on the PS. PS and FS represent the proportion of the combinations in the top K of the ranking based on the FS that are also in the top K of the combinations based on the PS. Each data point comes from one of the cross-validation synthetic exomes. In this analysis, both HPOs and Panel were used as seeds.

However, despite this observation, the number of overlapping combinations between any individual score and the FS remains relatively low (the median of the percentage of common top combinations between the DS and FS is less than 20%). This implies that the top-ranked combinations were primarily selected based on a combination of the scores rather than relying solely on one of the scores alone.

5.6.4 Analysis of the performance of Hop across the whole range of ranks

In this section, we analyze the performance of the tool over the full range of possible ranks. The results are shown in Figure 5.19, for the cross-validation and independent exomes combined. We show both the results in terms of absolute rank (5.19A) and in terms of percentile rank (5.19B), since it is important to take into account that the different exomes contain a diverse range of numbers of combinations that needed to be prioritized (see Section 5.3).

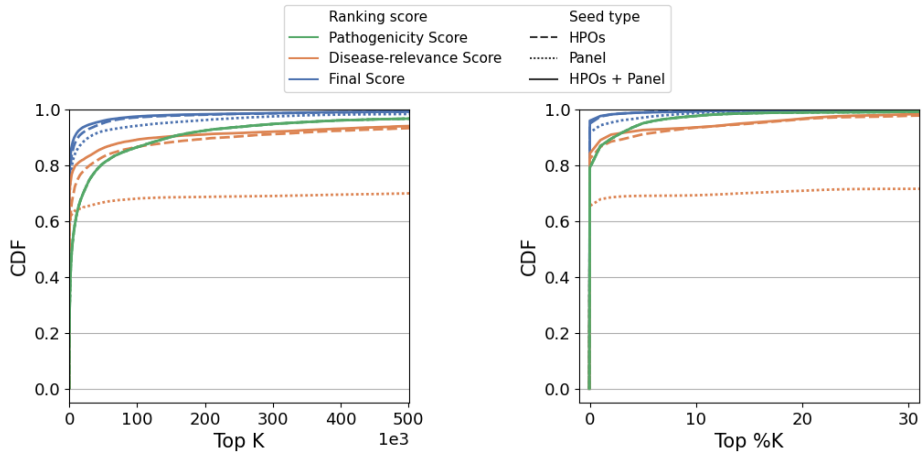


Figure 5.19: Performance of Hop in the cross-validation and independent exomes over a large range of ranks. (A) Cumulative Density Function (CDF) plot of the rankings, by using the FS (blue), DS (Orange), and PS (green) as ranking scores, with HPO terms as seeds (dashed line), genes from a gene panel as seeds (dotted line) and both HPOs and gene panel as seeds (solid line). The CDF plots illustrate the percentage of exomes for which the OLIDA combination is ranked in the top K by each method, with K varying between 1 and $5001e^3$ (inclusive). (B) Cumulative Density Function (CDF) plot of the rankings, by using the FS (blue), DS (Orange), and PS (green) as ranking scores, with HPO terms as seeds (dashed line), genes from a gene panel as seeds (dotted line) and both HPOs and gene panel as seeds (solid line). The CDF plots illustrate the percentage of exomes for which the OLIDA combination is ranked in the top K percentile by each method, with K varying between 1 and 30% (inclusive).

Overall, in more than 95% of the exomes, the pathogenic combinations are ranked in the top 1% combinations of the exomes, even though this percentage can represent a large number of combinations for some exomes (the maximum absolute rank of a combination using Hop is 9841606, for OLI694 inserted in an exome containing over 29 million combination).

The plot shows that for the large majority of the exomes (over 99%), the known pathogenic combinations were ranked in the top 6% of the exome, although this can represent up to 372000 in absolute rank, due to some exomes containing a very large number of combinations.

5.6.5 Influence of the exome template on the performance of Hop

The performance of Hop appears to be consistent across the different templates used to generate the synthetic exomes. We observe a slightly lower performance when the predictor is used in synthetic exomes generated from individuals of African descent (see Figure 5.20A). This can be explained by the fact that these individuals typically present with a much larger number of variants, making the problem of prioritization more difficult.

The performance of the tool is similar and even slightly better in the exomes that were generated by using data from the UK10K project as templates (Figure 5.20B). This shows that the performance is not biased by the fact that part of the training set of VarCoPP2.0 was obtained from individuals of the 1KGP.

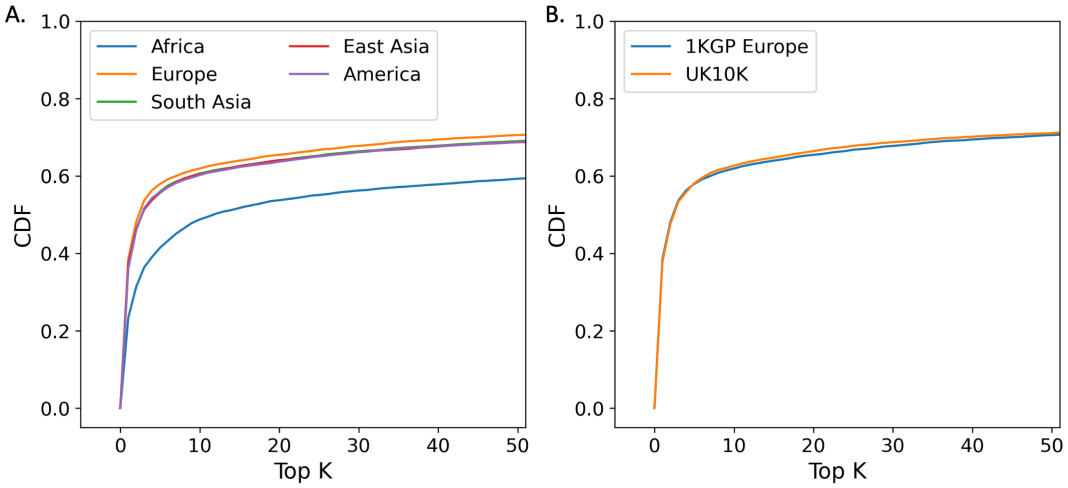


Figure 5.20: Cumulative density function plot of the rankings obtained by using the FS in all synthetic exomes (cross-validation and independent) comparing based on the ethnicity of the exome template in the 1KGP (A) and comparing based on the dataset used (1KGP or UK10K) for Europeans (B). Each line represent the rankings of the 420 OLIDA combination inserted in 20 different exome templates. The CDF plots illustrate the percentage of exomes for which the OLIDA combination is ranked in the top k by each method, with k varied between 1 and 50 (inclusive).

5.7 Hop outperforms current methods for the task of prioritizing oligogenic variants

In order to evaluate the relevance of digenic prioritization and to compare the performance of our method to state-of-the-art tools, we assessed the performance of three alternative prioritization methods on our independent validation dataset: (i) CADD (v1.6), which is a single variant prioritizer using only genotype information (ii) Exomiser (v13.1), which is a single-variant prioritizer using both phenotype and genotype information, and (iii) OligoPVP, which integrates monogenic prioritization together with PPI information, making it a first attempt to oligogenic phenotype-based prioritization. Whereas (i) and (ii) serve to show the added value of prioritization of variant combinations directly, (iii) is used to compare the performance. These comparisons are made against the different versions of our tool (i.e., using HPO terms, gene panel and both types of information), with the results shown in Figure 5.21.

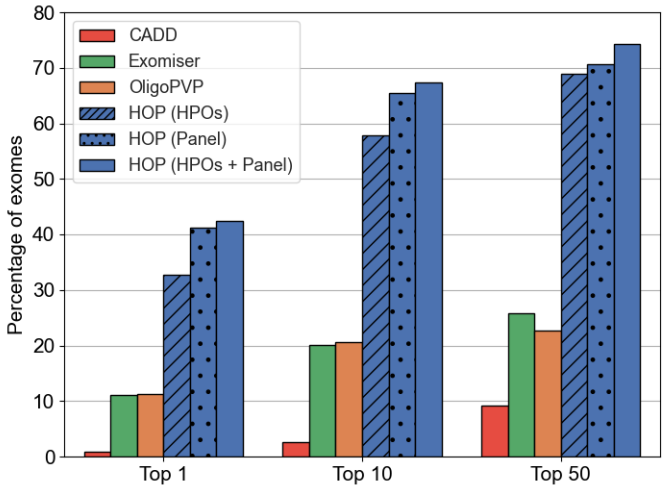


Figure 5.21: Comparison of the performance of the predictor with other prioritization tools. Percentage of exomes for which the OLIDA combinations is ranked in the Top 1, Top 10 and Top 50 instances, when using CADD, Exomiser, OligoPVP, and Hop for prioritization.

The results show that Hop, even using only HPO terms as seeds, is more adequate for identifying the (potential) digenic origins of a disease as it ranks the relevant variants earlier (and in combination) than monogenic prioritizers: Hop ranks the known pathogenic combination in the top 50 in 69.5% of the synthetic exomes, while CADD and Exomiser only identified the relevant variants involved in the combination in the top 50 in 6.4% and 22.7% of the exomes respectively. OligoPVP, the only attempt so far to perform digenic ranking, only identified the variants involved in the combination in the top 50 in 19.2% of the exomes, demonstrating the usefulness of our Hop approach for digenic prioritization.

In terms of time of execution, Hop is comparable to Exomiser for the majority of tested exomes, taking less than 10 minutes to prioritize a whole exome. However, the execution time can get considerably higher for exomes containing a large number of variants, since the number of combinations to be predicted then grows quadratically.

It is here important to note that the aim of this comparison is not to diminish the relevance of the monogenic prioritization tools for exome analysis, but rather to show that they are not tailored for the identification of oligogenic variants, and that we thus need complementary approaches, such as the Hop predictor introduced here, to detect such variants in WES data.

To further demonstrate the relevance of the monogenic prioritization approaches, we also investigate the percentage of exomes for which at least one variant of the OLIDA combinations in the test set were found in the top 1, top 10 and top 50 (Figure 5.22). These results, which compute the percentage of exomes for which the minimum rank of a variant involved in a combination is in the top K variants, indicate that Exomiser has a similar performance as

Hop to discover at least one of the oligogenic variants in the top 50. In particular, Exomiser performs slightly better than Hop using only HPO terms at identifying the first causative variant of a combination, although this predictor only relies on HPOs for disease-relevance scoring. This is probably due to the more complex disease-relevance scoring used by Exomiser, which takes advantage of other species phenotypic ontologies, which highlights that this could be improved in future versions of Hop.

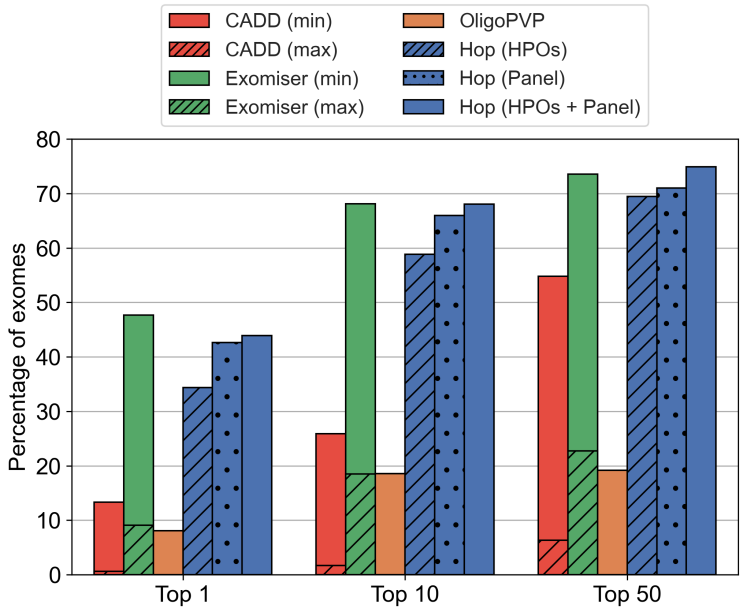


Figure 5.22: Comparison of the performance of Hop with other prioritization tools. Percentage of exomes for which the OLIDA combinations is ranked in the Top 1, Top 10 and Top 50 instances, when using CADD, Exomiser, OligoPVP, and Hop for prioritization. For Exomiser and CADD percentage of the exomes for which at least one variant in the combination is found (minimum rank, full bars) as well as the proportion for which all variants of the combination are found (maximum rank, striped bars).

It is important to note that this comparison does not diminish the relevance of Hop since we introduce this predictor as a tool to discover causative variants following the oligogenic inheritance model. This highlights the need for complementary approaches when diagnosing patients, with tools such as Exomiser being very useful to rank and identify monogenic causes, while our approach performs better in the task of identifying oligogenic causes. This comparison also highlights a certain assymetry in the OLIDA combinations present in the test set, with a large proportion of oligogenic combinations comprising at least one variant with a strong monogenic effect that can be easily detected by monogenic tools, while the second variant is more complex to detect using monogenic approaches.

5.8 Comparing ranking and classification metrics for performance assessment

In this section, we investigate how both classification and ranking metrics can lead to different assessments of the performance of computational tools. As described in Section 3.4.4, tools that aim to classify variants as pathogenic or neutral (i.e. classification tools) and tools that aim to rank variants based on how likely they are to cause a patient's phenotype (i.e. prioritization tools) are evaluated using different metrics to reflect their performance in a scenario resembling their "real-world" intended usage.

Evaluating the performance of Hop using classification metrics

First, we investigate the use of classification metrics to compare the performance of the different scores used by Hop to classify instances. To do so, we use as test set a subset of the independent test exomes by using all OLIDA independent set combinations as positive instances and two different sampling strategies for the neutral combinations: a random sample of 50 combinations from the entire exome and a random sample of 50 combinations from the top 500 prioritized combinations.

Since the scores are normalized per exome, the PS, DS and FS of the OLIDA combinations depend on the exome template they are inserted in. This testing set therefore resulted in 14,280 positive combinations (119 OLIDA combinations x 120 exomes) and 714,000 negative combinations (50 random combinations x 119 OLIDA combinations x 120 exomes). This performance evaluation analysis was done using the HPO and panel seeds to compute the DS and FS.

Since we do not have a classification threshold for the Hop predictions, we assess the performances using the ROC and PR curves for the PS, the DS and the $FinalScore$ (Figure 5.23).

We observe that the performance using classification metrics matches what was observed using the CDF curve. Although the PS and DS scores appear to have similar performance based on the ROC curve in the random set (AUC of 0.96 and 0.97 respectively), the DS shows a much higher PR-AUC of 0.82, while the PR-AUC is 0.6 for the PS. This difference in performance is increased when using the neutral combinations selected from the top ranked combinations, with the PS having a ROC-AUC of 0.66 and a PR-AUC of 0.08 in that set, while the DS has values of 0.78 and 0.31 for the same metrics. The evaluation of the performance of VarCoPP2.0 differs from the evaluation in Section 5.4.4 for three reasons: (i) the negative set is generated differently, as we here select combinations from synthetic exomes generated with the inserted OLIDA combinations, which means that there might be neutral combination

with one of the OLIDA variants, (ii) the imbalance level is different as we here select 50 neutral combinations for each pathogenic combination, (iii) the VarCoPP2.0 scores have been min-max normalized over the whole exome which means that the scores are not the same as the true predicted scores.

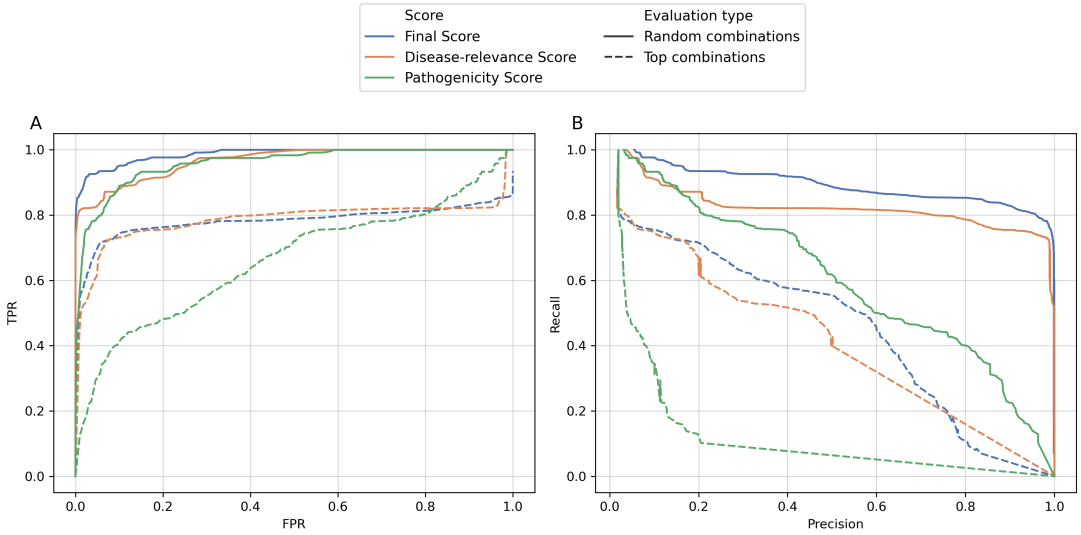


Figure 5.23: Evaluation of the different scores of the Hop predictor using the ROC (A) and PR (B) curves. The evaluation of each score was assessed in OLIDA combinations from the independent set as positive instances and sampled negative instances from the synthetic exomes using two sampling strategies: a random sample from the whole exome (full lines) and a random sample from the top 500 combinaitons (dotted lines).

Furthermore, we here again observe that it is the combination of both scores in the FS that achieves optimal classification performance, with the FS reaching a ROC-AUC of 0.99 and 0.78 in the random and top combinations sets respectively and a PR-AUC of 0.89 and 0.45 in the same sets.

This performance comparison highlights the fact that classification metrics reflect well the performance of a tool for the ranking of instances, since the order of scores from most to least performant is the same as the one obtained when assessing the CDF curves. Nevertheless, when only assessing the performance using the ROC curves, we observe that the DS and PS have a similar performance, while the DS shows much higher recall and ranking performance, highlighting the need to assess performance using both types of curves.

Finally, it is important to note that this comparison does not diminish the importance and relevance of the VarCoPP2.0 predictor alone. Although, Hop appears to show higher performance also for the classification of variant combinations as either pathogenic or neutrals, it does not have the same scope of application. Indeed, Hop requires the input of seeds to guide the search, and ranks all combinations based on the scores, but is therefore not useful to assess the pathogenicity of a specific variant combination outside of an exome context.

Comparing the performance of Hop with other tools using ranking metrics

In this second part of the comparison between classification and ranking metrics, we analyze the performance of Hop, VarCoPP2.0, and the three predictors used for comparison in Section 5.7 (CADD, Exomiser and OligoPVP) using the ranking metrics described in Section 3.4.4. For the monogenic prioritizers, these metrics were computed by considering each variant of the known OLIDA combination as a relevant instance in the ranking, while for the oligogenic prioritizers, only the full OLIDA combination is considered as a relevant instance. For the computation of the nDCG metric, relevant instances were assigned a relevance score of 1 and irrelevant instances a score of 0. Each metric was evaluated in the top 50, the top 100 and the full ranking (Figure 5.24).

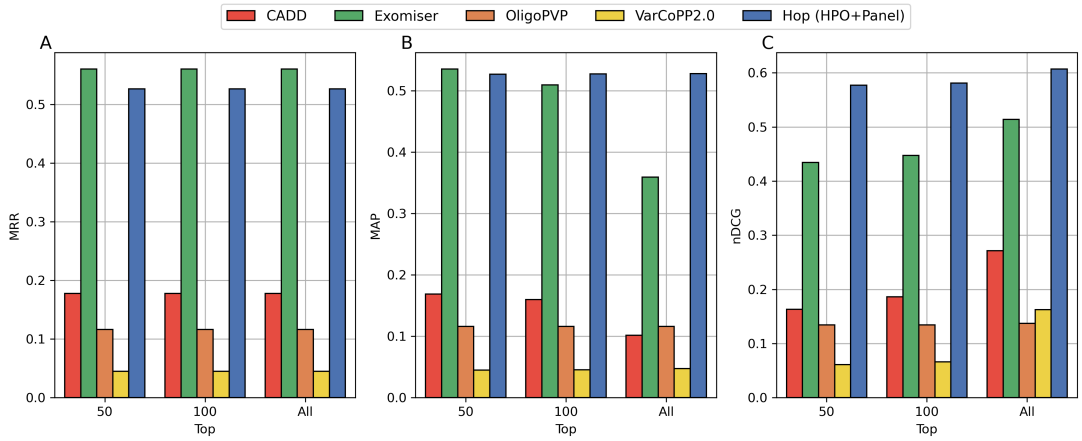


Figure 5.24: Performance evaluation of CADD, Exomiser, OligoPVP, VarCoPP2.0 and Hop in the independent set synthetic exomes. The performance is assessed using different ranking metrics: the Mean Reciprocal Rank (MRR) (A), the Mean Average Precision (MAP) (B) and the normalized Discounted Cumulative Gain (nDCG) (C).

We observe that using the MRR metric, which is computed as the inverse of the rank of the first relevant instance in a ranking, Exomiser performs better than Hop (Figure 5.24A). This is linked to the observation that Exomiser ranks well one of the variants in the combination (see Section 5.7). Since only the first relevant instance is taken into account by this metric, this predictor therefore appears to have better performance although only one of the relevant

items is identified in most cases. This is reflected when using the MAP and nDCG metrics to assess performance (Figure 5.24B,C), especially when these metrics are computed over the whole list of ranked items and not the top 50 or top 100 instances. In this case, Hop appears to outperform both monogenic and other oligogenic tools. The high performance of monogenic tools using these ranking metrics is due to the fact that these metrics put a lot of emphasis on having at least one relevant instance at the top of the ranking, and less importance on the lower ranked relevant instances. This bias might be important for tasks such as information retrieval or recommender systems, but can be misleading in the task of variant prioritization, since all relevant variants (in the case of an oligogenic inheritance model) need to be retrieved for the patient to obtain an accurate diagnosis.

In conclusion, it appears that using such ranking metrics to compare the performance of Hop to other prioritizers is misleading, since the definition of a relevant instance differs based on the scope of the used prioritizer. The comparison between Hop and Exomiser, for example, is strongly affected by the fact that in the top 50 and top 100, Exomiser only recalls one of the relevant variants and still shows high performance according to these metrics, since they do not take into account recall. However, these metrics appear to be relevant for the comparison between Hop and VarCoPP2.0, as it seems that the difference in performance matches the one observed using the CDF curve.

5.9 Integration in ORVAL and usage of the tools

In order to make our tools easily accessible for other researchers and clinicians, we integrated both the VarCoPP2.0 and the Hop predictor in the ORVAL platform, as well as its recently developed python package version [188]. This work was done with the help of Emma Verkinderen and Nassim Versbraegen.

Not many changes were required for the integration of the VarCoPP2.0 model except that the results section had to be adapted to reflect the fact that the model now outputs a single pathogenicity score instead of the CS and SS scores outputted by VarCoPP. The S-plot of the ORVAL results page was thus replaced by a swarm plot showing the distribution of predicted combinations in each of the possible predicted categories (Neutral, Disease-Causing, Disease-Causing 99% and Disease-Causing 99.9% confidence). In addition to this integration, the annotation database used by ORVAL has also been updated to contain data using both the Hg19 and Hg38 genome assemblies (while only the Hg19 assembly was supported previously). This allowed to train a VarCoPP2.0 model able to predict variant combinations aligned on the Hg38 assembly, and add it to the platform. This model shows similar and even slightly higher performances as compared to the Hg19 model presented here. This represent an important advancement as it will increase the potential usage of ORVAL and VarCoPP2.0

since most genomic data is now aligned to the Hg38 assembly. This new VarCoPP2.0 model has already been used to investigate the pathogenicity of variant combinations in male infertility patients [227,407], providing additional evidence for the involvement of these combinations in disease.

The integration of Hop in ORVAL required more adaptations. First, we wanted to be able to offer both types of analyses, to allow users to predict specific variant combinations with VarCoPP2.0 or to run panel data, but also to test the Hop method on whole exomes. We therefore created a new pipeline for Hop, which could be launched through a new input page, allowing for the input of both HPO terms and a list of genes, which are used as seeds for the computation of the DS, and the input of the number of top combinations that will be returned. In order for the number of predicted combinations to remain tractable memory wise (as keeping all generated combinations from an exome would require too much memory), we implemented a chunked version of the prediction pipeline: 100,000 combinations are predicted in each "chunk", and only the top N combinations are kept in memory, with this list updated at the end of each round of predictions, and N corresponding to the number of combinations specified by the user. For now, the output page of the Hop predictions is kept simple, showing only a table with the top predicted combinations and their respective PS, DS and FS as well as their rank. The version of ORVAL containing Hop is only available as a service *in development*, since the servers need to be upgraded to accomodate for larger datasets.

5.10 Conclusion

In this chapter, we present two novel computational methods, trained on OLIDA, to facilitate the interpretation of variant combinations in the context of human diseases. First we introduced VarCoPP2.0, which is a direct improvement of the VarCoPP predictor, keeping a similar basis, but trained on a larger and more confident dataset, including novel features and a simplified and thus faster model structure. This novel model shows tremendous improvement in performance over the first version of VarCoPP, but is still inefficient at ranking known pathogenic combinations in WES data. In order to address this constraint, we develop, the High-throughput oligogenic prioritizer (Hop), an original and relevant prioritization method that directly uses oligogenic information at several biological levels in order to rank, in WES data, variant combinations based on how likely they are to explain a patient's phenotype. Hop integrates a pathogenicity score (PS) together with a disease-relevance score (DS), in order to prioritize the combinations. The DS is computed by propagating information about the patient's disease in a heterogeneous network, in order to score the proximity of each gene pair to a set of user-defined seeds.

With WES data becoming more easily accessible, the development of oligogenic prioritization is becoming increasingly important as current oligogenic prediction methods were not designed for analyzing such quantities of data. Indeed, although the VarCoPP2.0 predictor used in this work to obtain the pathogenicity score was shown to have high accuracy in a classification setting [254], as a prioritization method, it only ranked the known pathogenic combination in the top 50 in less than 13% of the exomes. This implies that a clinician wanting to identify potentially pathogenic variant combinations in a patient's WES data will need to manually curate a huge number of combinations and investigate their disease relevance in order to identify the potentially disease-causing combination. With Hop, the disease-relevance of gene pairs is scored and combined with the pathogenicity prediction in order to provide the clinician with a ranking of the combinations. We show that by only looking at the top 10 combinations of this ranking, known disease-causing variant combinations can be found in 59.3% of all synthetic exomes (cross-validation and independent set). This significantly reduces the amount of work of the clinician, since, in order to obtain the same accuracy using VarCoPP2.0 alone, more than 3000 variant combinations would need to be investigated.

Moreover, Hop is shown to outperform existing monogenic approaches on the task of ranking digenic variant combinations in WES data. This does not diminish the usefulness of existing monogenic approaches, but rather highlights not only the need to design tools specifically for this task, but also the importance of scoring pathogenicity of variant combinations using method trained on oligogenic disease cases. Indeed, variants involved in oligogenic combinations have been shown to have different characteristics than variants involved in monogenic diseases, such as higher MAF and lower monogenic pathogenicity scores for both digenic and modifier variants [187,256,366]. Hop also introduces a new method to score the disease-relevance of gene pairs, combining the scores of the RWR algorithm to obtain a gene-pair proximity score to a set of disease-related seeds. Although prioritization methods have traditionally relied on phenotype-based approaches, we further investigate the utilization of gene panels as an alternative source of information for disease-relevance scoring. We demonstrate that incorporating both phenotypic and panel data can enhance performance. Moreover, by accommodating both types of information, Hop can perform prioritization even in cases where the patient's phenotype is unknown, as long as a gene panel for the disease is available.

Although Hop is highly successful, the predictor's performance appears to plateau, even when both HPO terms and disease genes are given as prior information. The first possible explanation for this limitation is that Hop strongly relies on prior knowledge and is thus limited by the incompleteness of the available molecular knowledge. This is especially true when such knowledge is obtained through low throughput acquisition methods, such as the PPI or gene ontology networks used in this work [408]. In addition to the limitations caused by incompleteness, using prior knowledge can introduce important biases. This can be seen, for example, by the fact that digenic genes and disease genes in general are more connected in

any knowledge graph than human genes overall (see Figure in Appendix C in case of BOCK). In the case of oligogenic disease, it is possible that the bias present in the HPO to gene network, where only 23.5% of all human genes are linked to HPO terms, introduces limitations. Indeed, the link between an HPO term and a gene is created based on known gene-to-disease associations, from OMIM or Orphanet, and disease-to-HPO associations, from the HPO database [66]. However, these databases contain almost exclusively data on monogenic diseases, which implies that genes involved in oligogenic diseases have fewer HPO terms associations than other disease genes, and is thus likely to explain why knowledge-based approaches for oligogenic diseases are currently limited by prior knowledge.

The limitations introduced by knowledge bias is further illustrated in our results by the difference in performance between the cross-validation exomes and the exomes of the independent set. Indeed, more combinations in the independent set could be annotated with gene panels, indicating that these combinations were linked with diseases that are overall more studied. This is likely to explain why our knowledge-based predictor also predicts these combinations better, especially knowing that, for 79% of the combinations that have an associated panel, the two genes involved in the combinations are present in the gene panel.

Furthermore, it should be emphasized that Hop, like any prioritization method, provides a ranked list of variant combinations for all exomes, without offering confidence estimates or suggesting how many candidates should be kept for further investigation. Given the uncertain frequency of digenic inheritance in diseases, this is important to consider when using the tool. Hop is most effectively employed as a complementary approach to existing methods or in case-control studies for diseases with suspected digenic inheritance. For example, it can be applied to patients for whom a monogenic diagnosis could not be obtained using traditional approaches, or to detect genetic modifiers to these monogenic variants.

Finally, it is important to note that the performance of Hop has so far only been evaluated using synthetic data, and it will be very important to assess the performance of the tool on real clinical WES data. It will also be essential to assess whether using such a prioritization approach can be helpful when analyzing not only single patients but also patient cohorts and assess whether such an approach can help identify oligogenic signatures for specific diseases. This is done in the next chapter of this thesis.

Chapter 6

A computational analysis protocol for detecting novel oligogenic causes

In this chapter, we investigate how we can use the Hop predictor developed in the previous part in order to discover novel oligogenic causes to disease in real patients data. For this purpose, we work on a cohort of patients affected with male infertility, a relatively common condition which remains poorly understood and is expected to have a significant fraction of cases attributed to oligogenic inheritance.

We design a computational analysis protocol for the application of Hop to discover new oligogenic causes in case/control studies. We first define a VCF filtering protocol to ensure that the compared regions and variant datasets of cases and controls are compatible. We then perform gene enrichment analysis to ensure that we don't find any significant associations which would be caused by artefacts in the data, as well as investigate potential monogenic causes in the cohort. We then prioritize each patient with Hop and analyze whether the predictor is capable of finding the manually diagnosed cases. Finally, we perform an exploratory analysis of enriched gene pairs and genes within gene pairs in the whole cohort, as well as in patients subgroups, in an attempt to discover new oligogenic signatures for the disease.

This analysis allows to validate the effectiveness of Hop to identify oligogenic variants in real patients data, as well as highlights its potential to identify oligogenic signatures in whole cohorts, thus providing useful insights into the genetic architecture underlying diseases. These newly identified signatures will need to be further assessed and validated by clinicians in order to be found as really meaningful.

6.1 Motivation and objectives

For many diseases, only a fraction of cases are genetically diagnosed, while a large proportion of patients are thought to have a genetic cause which remains to be identified. This is the case of male infertility, which is a highly heterogeneous condition, with a significant proportion of cases considered as “idiopathic”, meaning that no etiology is currently known (see Section 1.4). A certain number of cases are suspected to have a genetic origin, and a large number of genes, involved in a variety of different biological processes, have been reported as potentially relevant for the disease. However, the diagnostic yield of exome sequencing remains low, in part due to the fact that some cases are probably caused by oligogenic variants, and that this model of inheritance is hard to detect.

In the previous chapter, we describe the development of two novel methods for the detection of oligogenic causes to disease: VarCoPP2.0, which aims to predict the pathogenicity of variant combinations in gene pairs and is more tailored for the identification of variant combinations in gene panels, and High-throughput oligogenic prioritizer (Hop), which integrates the predictions of VarCoPP2.0 together with background information on the patient’s disease in order to prioritize potentially disease-causing combinations directly from WES data. While these two methods have been evaluated in test sets comprising of known pathogenic combinations reported in the latest versions of OLIDA, their potential to discover new oligogenic causes in real patient data has not been assessed.

In this chapter, we investigate the usefulness of our novel prioritization tool, Hop, for discovering novel oligogenic causes to disease. In doing so, we aim to design a reproducible computational analysis protocol, which could be applied to any patient-control cohort for a specific disease in order to detect novel oligogenic signatures for this disease of interest. Our work is applied to the ESTonian ANDrology (ESTAND) cohort for male infertility (see Section 3.1.2), which has been already extensively studied in a collaboration between members of our group and researchers at the University of Tartu [227,407]. Several patients in the cohort were diagnosed with a monogenic cause and a few cases were found to carry oligogenic variants. This cohort therefore provides a great opportunity for the validation of our method to detect these oligogenic variants, and is also promising to make new discoveries about the genetic basis of male infertility, by investigating which gene pairs and variant combinations are found in common between patients in the cohort based on our novel prioritization approach.

6.2 Protocol overview

The developed protocol consists of 3 main steps, followed by different downstream analysis tasks and is schematically represented in Figure 6.1.

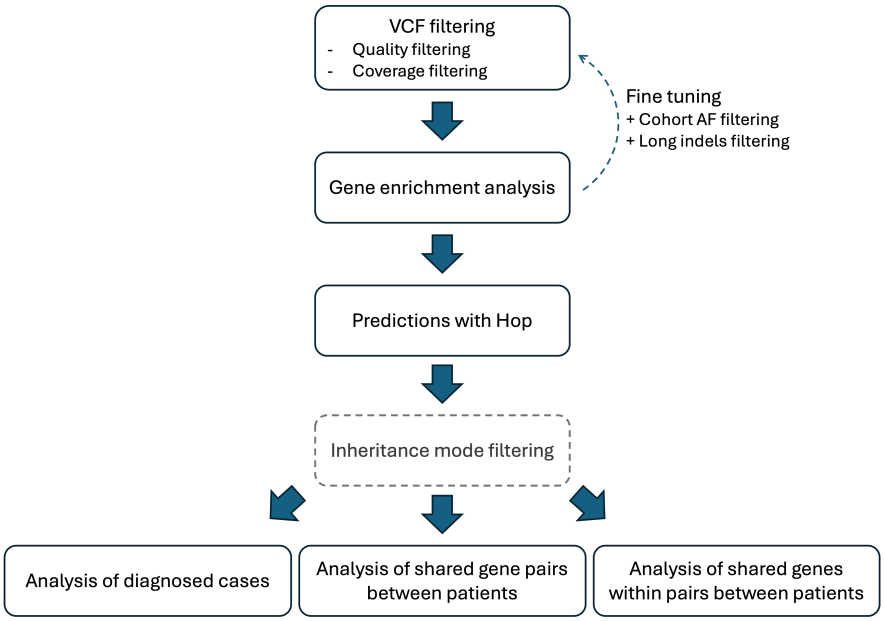


Figure 6.1: Overview of the analysis pipeline for the oligogenic analysis of the male infertility cohort. First the VCF files are filtered to remove variants that are potential sequencing errors, variant calling errors or in regions that are not properly covered by the different sequencing kits. Second we perform gene enrichment analysis to discover potential monogenic associations and assess that all erroneous variants have been removed. This step helps fine tune the VCF filtering step. Finally, the data of each sample is prioritized with Hop and we investigate the use of an inheritance mode filter on the results. We analyse the prioritized combinations by first looking at the diagnosed patients, then investigating enriched gene pairs and finally investigating enriched genes within gene pairs.

First, we determine how to properly filter the VCF files available. This step is crucial to remove potential sequencing or variant calling errors that can lead to spurious associations. In addition, since the samples in our cohort were sequenced using different sequencing technologies, this step is necessary to remove the variants which were specific to the sequencing kits used, and thus ensures that our analysis only includes regions which are well sequenced in all samples.

This filtering step is followed by gene enrichment analysis in different gene sets as well as in the whole-exome. The goal of this gene enrichment is two-fold: first, we want to ensure that there is no enrichment in genes that should not be enriched in patients or controls (which we refer to as *"background genes"*) and second, we want to assess whether we can already

identify monogenic associations in the cohort, whether it is in the candidate gene panel, or in any gene in the exome. This gene enrichment analysis led to the application of two additional filters to the VCF, since it appears from this analysis that some genes were enriched due to other sequencing artefacts (see Section 6.3).

In a third step, all exomes are prioritized with Hop using the gene panel for male infertility from [227] as seeds, and we investigate different downstream analyses tasks to bring together the results. We restrict the number of combinations prioritized to 50.

For the subsequent downstream tasks, we add an optional filter based on the inheritance mode of the genes. With this filter, we do not take into account single heterozygous variants in genes that are known to have an Autosomal Recessive (AR) mode of inheritance. However, we keep homozygous and presumed compound heterozygous variants in those genes. The goal of this optional filter is to not take into account in the results single heterozygous variants in AR genes since the normal copy of the gene can then compensate for the mutated one [409]. While this is the case for monogenic causes, it is not necessarily true under the oligogenic inheritance model where a single heterozygous variant in a recessive gene can have an effect due to its interaction with a second variant [409, 410]. We therefore investigate the two scenarios.

From these results, we briefly analyse general statistics on the Hop predictions and we then follow 3 different analyses routes. In the first one, we assess in details the patients that were already diagnosed with oligogenic variant combinations and monogenic variants by the group of professor Maris Laan. We look at the rank of the manually diagnosed combinations in the Hop output, and investigate the top 10 and top 50 prioritized combinations for each patient, to assess how these variants relate to the ones found manually (Section 6.6). In a second step, we search for enriched gene pairs within infertile men compared to controls and within different subgroups of male infertility, in order to detect specific oligogenic signatures within the patients (Section 6.7). Finally, we also analyze enriched genes within gene pairs, i.e. genes that are found to be frequently present in the prioritized combinations in the patients compared to controls, to investigate potential monogenic anchors or modifier genes for the disease (Section 6.9). This analysis is here again repeated on the whole cohort and in the different patient subgroups.

6.3 Developing an appropriate filtering protocol for VCF files

The ESTAND cohort WES data is split in three separate sets (see Section 3.1.2):

- 429 patients whose exomes have been sequenced at at the Finnish Institute for Molecular Medicine (which we refer to as *FIMM patients*).
- 81 patients whose exomes have been sequenced at the McDonnell Genome Institute of Washington University in St-Louis as part of the GEMINI study (which we refer to as *GEMINI patients*).
- 322 control individuals whose exomes have been sequenced at the Huntsman Cancer Institute High-Throughput Genomics Core Facility as part of the GEMINI study (which we refer to as *GEMINI controls*).

Since these different sets were sequenced in different centers and using different sequencing platforms, our first objective is therefore to develop an appropriate filtering protocol for the VCF files available in order to obtain comparable data between the patients and controls. To do so, we apply a series of filters to the different VCF files and investigate the number of variants present in the VCF and for each sample in the different sets (Table 6.1).

While the process of filtering variants in VCF files prior to analysis is essential, and a key step in any genetics analysis pipeline, there are no clear standards defined. Indeed, the types of filter to be applied highly depend on the sequencing platform that has been used, the metrics available in the VCF file, and the type of analysis that is to be performed.

In our case, we investigate different types of filters with two goals: first, we want to remove any bias that could be linked to the sequencing kit used, and second we want to remove false positives, i.e. variant calls which should have been reference variants but are due to errors in the sequencing or variant calling. Indeed, as mentioned in Section 3.1.2, the individuals were sequenced using three different library kits. The FIMM patients were sequenced using the Human Core Exome kit from Twist Biosicence, while the GEMINI patients were sequenced using an in-house exome targeting reagent and the GEMINI controls were sequenced using the Illumina DNA Prep kit.

The first filter applied is called the “routine” pipeline, and is based on variant quality criteria, only retaining variants with Genotype Quality (GQ) ≥ 20 and of Read Depth (DP) ≥ 10.

To remove the bias due to sequencing kits, we correct for differential read-depth coverage at specific loci similarly to what was done in previous studies [411–413]. This filter, which we refer to as “sufficient coverage” or “SufCov” filter, only retains variants that reside in Consensus Coding Sequence (CCDS) coding regions (± 10 bases of intronic positions), where at least 90% of GEMINI and FIMM-sequenced samples show at least 10-X coverage.

In addition to these initial filters, gene enrichment analysis revealed that several genes appeared to be significantly enriched, due to specific variants that were found to be carried by a large number of samples (see Appendix D). To remove these variants, which are due to either sequencing or variant calling errors, we thus investigate different percentages for cohort allele frequency filters (referred to as "CAF filters"). These filters remove variants that are present in more than $N\%$ of the samples of a specific set, since these are more likely to be erroneous variants.

Gene enrichment analysis also disclosed the presence of potential sequencing errors in long indel variants in specific genes such as PABPC1 (see Appendix D). We therefore further filter all indels which involve the insertion or deletion of more than 3 DNA bases with the "Indel filter".

Finally, we also report on the number of variants per sample after applying the VarCoPP2.0 filters (see Section 5.4.1) as well as the number of variants per sample that would be retained if we apply the Inheritance Mode (IM) filter described above. These filters are applied per sample during the prioritization of combinations by Hop for the VarCoPP2.0 filter and during the analysis of the results for the IM filter.

The variant counts in the complete VCF file and statistics on the variants counts per sample after applying each filter sequentially are shown in Table 6.1. We can see that the largest reduction in variants happens with the Sufficient Coverage filter, that removes differences between the sequencing kits. Out of the 62,417,045 CCDS positions, only 33,249,089 (53.26%) positions are retained. Such a large reduction in the number of genomic positions investigated has been observed in other studies [411], and is an essential step when aiming to run an enrichment analysis between patients and controls which have been sequenced using different kits. If not for this filter, we would simply observe an enrichment in the regions that are not properly covered by both kits.

		GEMINI controls (N=322)		FIMM patients (N=429)	
Filter type	Description	Global count	Sample count	Global count	Sample count
Routine pipeline	Remove variants with DP>=10, GQ>=20	692,322	Min: 33,525 Avg: 72,595.5 Med: 70,948 Max: 115,258	910,225	Min: 68,126 Avg: 112,329.5 Med: 111,040 Max: 146,794
Correct for sequencing kits	Remove sites that do not have 10X coverage in at least 90% of the samples	162,807 (23.52%)	Min: 15,479 Avg: 20,763.47 Med: 20,838.5 Max: 28,364	185,500 (20.38%)	Min: 19,876 Avg: 20,856.36 Med: 20,851 Max: 22,740
Cohort allele frequency 2%	Remove variants with AF>2% in the cohort	99,706 (14.40%)	Min: 439 Avg: 555.85 Med: 544.5 Max: 1,036	123,544 (13.57%)	Min: 440 Avg: 554.80 Med: 542 Max: 2,445
Indel filter	Remove indels longer than 3bps	97,784 (14.12%)	Min: 433 Avg: 545.11 Med: 533.5 Max: 1,018	121,238 (13.32%)	Min: 431 Avg: 544.35 Med: 532 Max: 2,414
VarCoPP2.0 filter	Remove variants with MAF>3.5% synonymous and intronic variants.	N.A.	Min: 380 Avg: 484.3 Med: 475.0 Max: 899	N.A.	Min: 386 Avg: 482.3 Med: 471 Max: 1,742
Inheritance mode filter	Remove single heterozygous variants in recessive genes	N.A.	Min: 129 Avg: 181.9 Med: 177 Max: 386	N.A.	Min: 124 Avg: 177.39 Med: 174 Max: 860

Table 6.1: Variant counts in the VCF files after various filters. The global counts represent all the variants in the VCF file while the sample count show the minimum (Min), average (Avg), median (Med) and maximum (Max) variant counts per sample for each set. The percentage under the global counts show the percentage of variants of the raw files which are retained after applying the filters listed in the rows above and the filter in the current rows. Each filter is shortly described in the description column. The last two filters are applied per sample, when running the prioritization algorithm and the global count is thus not shown. The VarCoPP2.0 and Inheritance mode filters are applied per sample and the counts are therefore not shown in the whole cohort.

After applying all filters, we observe that the number of variants per sample seems to match between the FIMM patients and GEMINI controls sets. We here primarily focus on these two sets while the results for the GEMINI patients set are shown in Appendix D. The number of variants per individual appeared to be significantly reduced in this set after applying the first filters, leading to the decision to remove these patients from the case-control analysis (Section 6.7.1 & 6.9). These patients were however kept for the diagnosed patients analysis, since half of the oligogenic diagnoses were found in individuals from this cohort (see Section 6.6).

On average, before applying the IM filter, each control carried 484.3 variants and each patient carried 482.3 variants that could be prioritized using Hop (Table 6.1). More than half of these variants appeared to be single heterozygous variants in presumed AR genes, which means that each control and patient only carried on average 181.9 variants and 177.4 variants respectively, after applying the IM filter.

6.4 Gene enrichment analysis

In this second part, we assess whether the different filtering steps performed on the VCF files have efficiently removed any batch or sequencing artefacts that were initially present in the data. In order to do this, we investigate the enrichment of variants in two different gene sets: a "background" gene set, which consists of a list of gene that have been reported in the literature as housekeeping genes [414], and for which we expect no significant enrichment, and the candidate gene set for male infertility, which was compiled based on literature search in [227]. We hypothesize that if the filtering is efficient, we would not observe any enrichment in the background gene set, but might observe enrichment in the candidate gene set.

To perform the enrichment analysis, we count, for each gene in the gene sets, the number of samples with at least one "qualifying variant" located in that gene, with different definitions of "qualifying variant" as described below. We thus obtain a contingency table for each gene, which counts the number of controls and patients with and without qualifying variants in the gene. Based on this contingency table, we use the Fisher's exact test, to compute a *p-value* that represents the significance of the difference in enrichment for that gene between the two populations. This process is repeated for different types of "qualifying variants":

- **All variants:** we do not apply further filtering on the variants and thus consider all variants in the VCF as "qualifying".
- **Rare variants:** we only consider variants with MAF $< 1\%$ as "qualifying variants".
- **Rare non-synonymous variants:** only variants with MAF $< 1\%$ which are non-synonymous (i.e. which have a consequence on the protein product) are considered as "qualifying variants".
- **Rare missense variants:** only variants with MAF $< 1\%$ which are missense (i.e. which result in the change of an amino acid) are considered as "qualifying variants".

The enrichment can be visualised as a Quantile-Quantile (QQ) plot, which represents the observed p-values against the expected p-values (Figure 6.2) when investigating the fully filtered VCF files before applying oligogenic prioritization, i.e. the files which have been subjected to the routine pipeline filter, sufficient coverage filter, cohort allele frequency filter and indel filter. The enrichment plots for intermediate steps of the filtering process can be found in Appendix D.

We first investigate the significance of genes within the set of housekeeping and candidate genes, and then look at whether any gene is found to be significant at the whole exome level. The threshold for significance of 0.05 is corrected for multiple testing using Bonferoni correction, to account for the number of genes (20,000 for exome wide significance and 4,300 for gene set significance) and qualifying variant types tested (4 types), and is thus divided by $16,000$ for gene set significance, and by $40,000$ for exome wide significance (Figure 6.2 dotted lines).

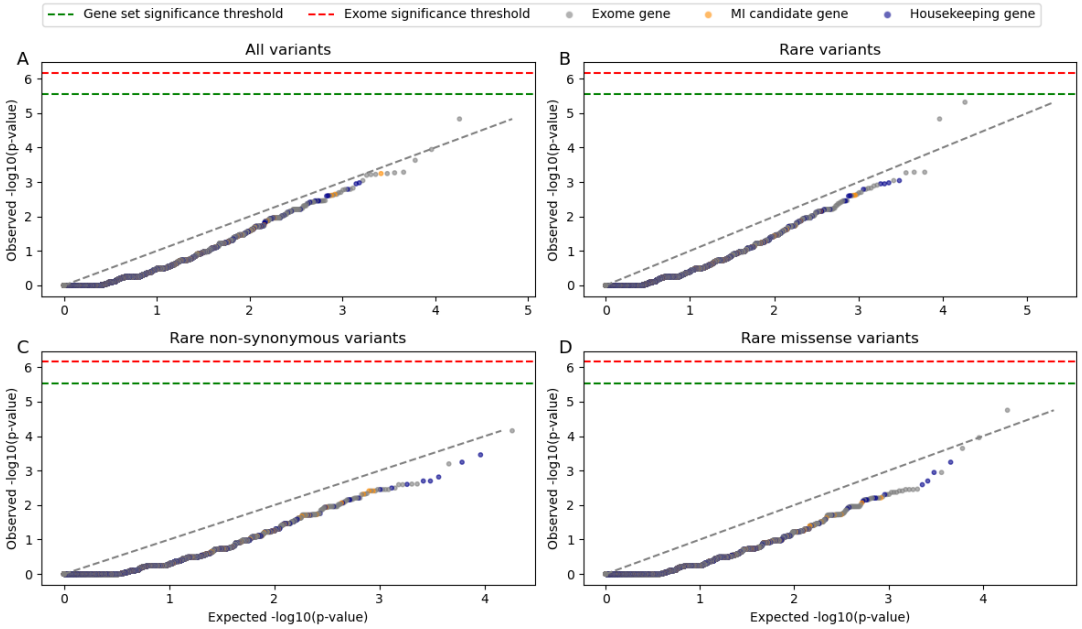


Figure 6.2: QQplot of the gene enrichment in patients vs controls for different types of qualifying variants: all variants (A), rare variants (B), rare non-synonymous variants (C) and rare missense variants (D).

We observe no significant enrichment of any gene with any variant type in the housekeeping gene set and the candidate gene set, and it appears that no gene reaches exome wide enrichment significance either (Figure 6.2). The most enriched genes include the ADCY8 gene ($p-value = 1.5e-5$ when considering all variants and rare variants), FMN2 ($p-value = 4.8e-6$ when considering rare variants), and the DGKZ gene ($p-value = 9.7e-5$ and $p-value = 1.8e-5$ when considering rare non-synonymous and rare missense variant respectively).

Significant enrichment for one or more specific genes was observed during intermediate filtering steps, and the QQplots obtained with the VCF files at these steps are shown in Appendix D. In particular, enrichment of particular genes motivated the addition of specific filters to the procedure. This is the case for the “indel filter”, which was applied after noticing an enrichment in the PABPC1 gene at the whole exome level when considering rare non-

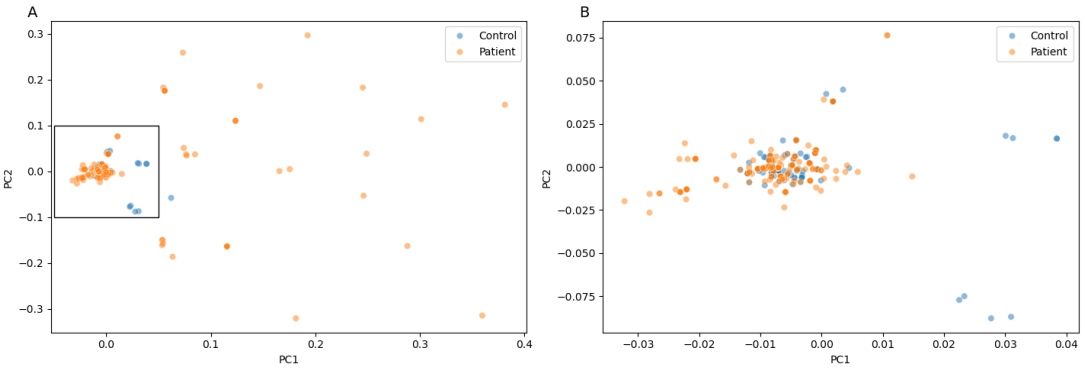


Figure 6.3: Population structure of the patient and control cases based on principal component analysis (PCA). Panel A represents the full PCA while B shows a zoomed version of the square shown in panel A.

coding variants. Investigating the variants present in this gene in the patients, we noticed that many variants included large indels in a region which is potentially error prone. Removing indels larger than 3 base pairs, which should be very rare, removed the enrichment in this gene.

The fact that no significant enrichment is observed in candidate genes is not necessarily surprising, given that the monogenic causes identified in the cohort appeared to be relatively heterogeneous, involving 39 distinct genes (see Appendix E and [227]). Furthermore, given the size of the patient and control sets, at least 23 patients would need to carry a qualifying variant in a gene for this gene to be significantly enriched (assuming none of the controls carried a variant in that gene).

This analysis ensures that any findings that are found through enrichment in gene pairs are not due to enrichment in specific genes caused by batch effects or potential sequencing artefacts.

Complementary to this gene enrichment analysis, we also do a Principal Component Analysis (PCA) of the filtered VCF files of the patients and control population to visualize whether the samples from the two groups overlap. For this analysis, we merge the two filtered VCF files (after applying all but the VarCoPP2.0 filters), and use the PLINK2 software [415], using default settings, to obtain the eigenvector and eigenvalues of the PCA. The results, shown in Figure 6.3, confirms the gene enrichment analysis. Indeed, while it appears that some of the patient samples are generally more spread out, the patient and controls samples overlap in the PCA space.

6.5 Statistics on Hop predictions in the cohort

Following this monogenic enrichment analysis, we run the Hop predictor on all the individuals in the cohort using different gene panels as seeds.

Analysis panels come from [227] which include an infertility panel comprising 638 genes combining the following individual panels:

- **Replicated studies genes (n=156)**: Genes that were reported as associated with male infertility in more than one study.
- **Single study genes (n=330)**: Genes that were reported as associated to male infertility in a single study.
- **POI genes (n=79)**: Genes that are associated with Primary Ovarian Insufficiency (POI), since links between POI and male infertility genetics have been reported [218].
- **HPG axis genes (n=73)**: Genes that are implicated in the Hypothalamic-Pituitary-Gonadal (HPG) axis, which is important for the development and regulation of the male reproductive system [416].

The full infertility panel of 638 include 2 genes located on the Y chromosome (where variants could be associated with Y-linked inheritance, see Section 1.2.2), 70 genes located on the X chromosome (where variants could be associated with X-linked inheritance, see Section 1.2.2) and 566 genes located on autosomal chromosomes.

We run Hop using the combined panel of 638 genes as well as each of the individual panels independently and investigate general statistics on the prioritized combinations. After prioritizing the combinations, we also investigate the use of an Inheritance Mode (IM) filter, which removes the combinations involving a single heterozygous variant in an AR gene. This type of filtering is typically applied in clinics, since AR genes are known to be tolerant to heterozygous variants and clinicians searching for causative variants try to limit the number of variants to investigate manually to only consider the ones that are most likely to be relevant. However, this filtering is not typically applied in our oligogenic predictors, since the oligogenic inheritance model is assumed to be more complex and a single heterozygous variant in an AR gene might act as a co-contributor to the phenotype. To apply this filter, we obtain the inheritance mode of each human gene from the information in [227] for the genes in the candidate gene list and the DOMINO platform [417] for the remaining genes, a tool that predicts the mode of inheritance of human genes. We then create an IM filtered top 50, which only contains the top 50 combinations of each sample where the combinations include any type of variants in AD genes and XL genes (since the samples are all male individuals, all variants in genes located on the X chromosome are hemizygous), as well as homozygous or presumed compound-heterozygous variants in AR genes. Since the phased genotype was not available from the VCF files (i.e. whether variants are on the same chromosome or not) we consider any two variants in the same gene as presumed compound-heterozygous.

Finally, it is important to note that although Hop uses min-max normalization to give equal weight to the *Disease-relevance Score (DS)* and *Pathogenicity Score (PS)* when ranking the combinations in an exome, we used the non-normalized version of the VarCoPP2.0 score to assess the predicted class and confidence zone of each combination that is predicted as pathogenic. Keeping the information about the confidence zones allows us to have an idea of the probability of a variant combination to be a false positive based on its pathogenicity score, and to categorize the ranked combinations based on the VarCoPP2.0 confidence zones.

6.5.1 General statistics on the number of instances and effect of the gene panel

First, we investigate the number of unique instances obtained in the top 50 of each sample, when applying Hop using the different gene panels as seeds. This analysis gives us an idea of the diversity of variants, genes and gene pairs in the combinations prioritized by Hop. In addition, we analyze the number of combinations for which at least one gene in the combination is not in the original seed panel, in order to have an estimate of the potential of the method to generate new discoveries beyond the gene panel. The number of instances shown here is based on the analysis done without applying the IM filter. We further assess the effect on the IM filter by comparing the number of combinations where at least one gene is not in the gene panel when applying this filter on the set of exomes prioritized using the combined gene panel only (Figure 6.4C).

Overall, most samples have 50 unique and distinct gene pairs in the top prioritized combinations (Figure 6.4A). The median number of unique variants in the top 50 combinations prioritized with the combined panel is 34 and the median number of unique genes is 31 (Figure 6.4A). This indicates that in most cases, a gene and the variants that are located in it are involved in several combinations (each combination contains 2 genes and up to 4 variants). We notice that the individual gene panels, and especially the smaller ones, appear to be associated with a larger number of unique variants and genes ranked in the top 50 than the combined panel (Figure 6.4A). This might be due to the fact that with smaller gene panels, the prioritization is less constrained by the *Disease-relevance Score (DS)* and thus ranks more variants in genes that are not part of the original panel in the top of the list, therefore diversifying the set of prioritized variants and genes.

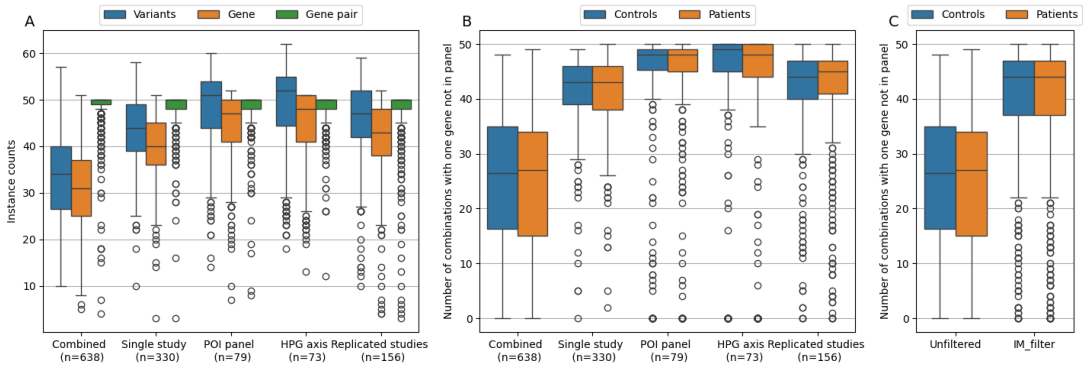


Figure 6.4: General statistics on the instances in the top 50 of each sample, when applying the Hop predictor with the different gene panels. (A) Boxplots of the number of variants, genes and gene pairs in the top 50 prioritized combinations. (B) Boxplots of the number of combinations in the top 50 where at least one gene is not in the gene panel used as seeds. (C) Boxplots of the number of combinations in the top 50 where at least one gene is not in the gene panel for the unfiltered and Inheritance Mode filtered results when the combined panel is used as seeds. Each boxplot is constructed based on all samples in A and separated by the control/patient status of the sample in B and C.

This hypothesis is supported by the observation that the combinations prioritized show a larger proportion of variants in genes that are not present in the original panel when the panel used as seeds is smaller (Figure 6.4B). We observe that for the majority of the samples, around half of the combinations in the top 50 include at least one gene that is not part of the gene panel. This supports the fact that Hop does not limit the search for variant combinations within the gene panel, but extends the predictions to the whole exome.

Moreover, we notice that this proportion is much larger when considering the IM filter than when looking at all variants independently of their zygosity and inheritance mode of the gene they are located in (Figure 6.4C). This is probably caused by the small bias in the contribution of the *DS* to the top ranked combinations observed in Section 5.6.3. A larger proportion of combinations ranked in the top 50 are prioritized due to having a high *DS* than due to having a large *Pathogenicity Score (PS)*. This means that many variant combinations in the top 50 are in genes that are part of the gene panel irrespective of the fact that they involve variant combinations that are predicted with high pathogenicity. When applying the IM filter, we remove the combinations that have single heterozygous variants in AR genes, and therefore go further “down” in the rankings and include more combinations with a lower *DS* involving genes that are not part of the gene panel.

6.5.2 General statistics on pathogenic combinations and effect of the Inheritance-mode filter

We investigate the number of combinations that are predicted as pathogenic and are found in the 99% and 99.9% confidence intervals according to the VarCoPP2.0 predictor. This is done by looking for each sample at the proportion of variant combinations that are found in the three confidence zones (Disease-causing, 99%-confidence zone and 99.9%-confidence zone), analyzing first all combinations prioritized in the top 50 and second the combinations obtained in the top 50 when applying the IM filter.

When analyzing all combinations without filtering on the mode of inheritance, the majority of the samples (458/751, 60.99%) have all of the combinations in their top 50 predicted as disease-causing by VarCoPP2.0 (Figure 6.5A). On average, across the patients, 98.4% of combinations are predicted as pathogenic and 73.4% and 47.3% are found in the 99 and 99.9% confidence intervals. In the control group, the proportions are similar, with 98.4%, 73.9% and 47% combinations predicted as disease-causing and within the 99% confidence and 99.9% confidence zones respectively.

Interestingly, applying the IM filter reduces slightly the proportion of combinations predicted as disease-causing (Figure 6.5B). In the top 50, only 278 out of the 751 samples (37%) had all of the prioritized combinations predicted as disease causing. On average, 97% of combinations are predicted as disease causing in both the patient and control groups, with 80% of combinations found in the 99%-confidence zone and 46% of combinations in the 99.9%-confidence zone in the patient group and 45% of combinations in the 99.9% confidence zone in the control group.

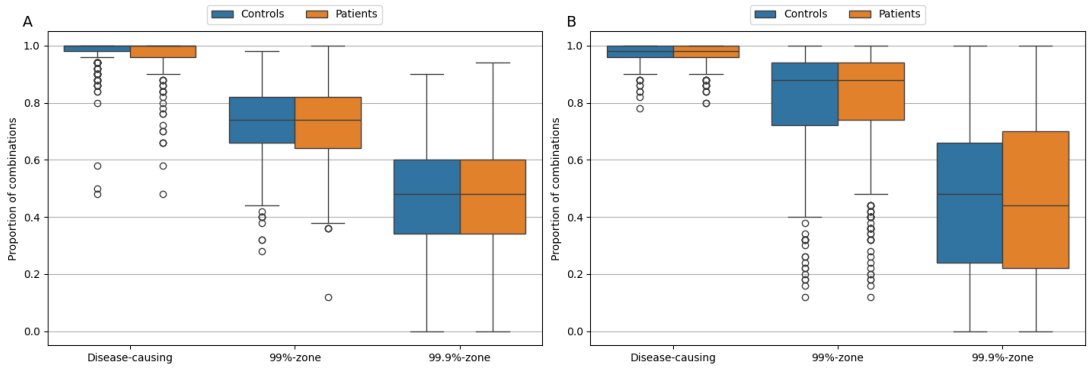


Figure 6.5: Proportion of combinations in the top 50 obtained running Hop without applying the IM filter (A) and applying the IM filter (B) that are predicted as disease-causing, in the 99% confidence interval and in the 99.9% confidence interval for patients and controls.

6.5.3 Analyzing network structure of top combinations

For each patient, we analyze whether the top prioritized combinations appear to be connected in a network, as well as whether we can get insights into the topology of these networks. This is done by generating small networks with the top 10, top 20 and top 50 combinations for each patient, with each node representing a unique variant or compound heterozygous variant, and edges representing prioritized combinations in the top K. We then look at whether all combinations are connected, by computing the number of connected components in each network. A connected component represents a set of nodes that are connected with each other by paths.

Overall, we observe that for the large majority of samples (386/429, i.e. 90% of the patients and 293/322, i.e. 91% of the controls), the top 10, 20 and 50 combinations are connected in a network comprising a single connected component (Figure 6.6). For a small proportion of samples, the network is divided in more components, with one patient having a network made up of 4 components when looking at the top 50 combinations (Figure 6.6C) . These results are obtained on the top 50 combinations after applying the IM filter, but similar results are found on the unfiltered tops, with a slightly larger number of samples having networks with one connected component.

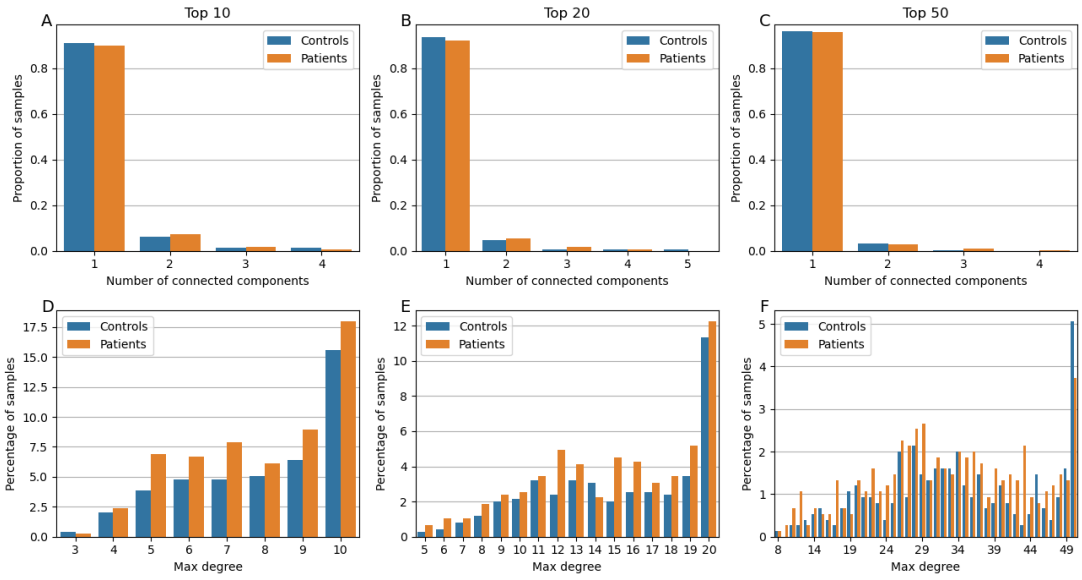


Figure 6.6: Statistics on the network structure of the top combinations for each sample when using the combined gene panel as seeds. Number of connected components in the network generated by the top 10 (A), top 20 (B) and top 50 (C) prioritized combination of each sample. Histogram of the maximum node degrees in in the network generated by the top 10 (A), top 20 (B) and top 50 (C) prioritized combination of each sample. The networks of the top combinations are generated by creating a node for each variant in the top combinations and putting edges between the variants if a combination between the two variants is prioritized in that top.

In a second step, we also investigate the topology of these small networks of prioritized combinations, which can be thought as a representation of the oligogenic signature of each individual. We decide to look at the degree centrality of the networks, and more specifically at the maximum degree of the nodes in the networks to assess whether networks tend to be centered around specific variant nodes. We observe that in each top, there is a significant proportion of samples, both patients and controls, for which the maximum degree is equal to the number of combinations (i.e. the number of edges in the network). This means that for these samples, the network has a "star" structure, with each variant combination stemming from one unique variant. These networks are likely to be associated with a highly pathogenic variant, such as a potential monogenic variant, which can thus act as an "anchor" to variant combinations, as it is always predicted as highly relevant for the disease although with different secondary, and potentially "modifier", variants. The proportion of samples with such networks' structures are 15% for the controls and 17.5% for the patients when considering the top 10 combinations. These percentages decrease as we consider larger number of combinations, with only 5.1% controls and 3.8% patients showing networks centralized around a single variant when looking at the top 50 combinations. We further investigate these samples, when analyzing the patients with a monogenic diagnosis in Section 6.6.2.

Overall, the analysis of the general statistics of the top 50 combinations prioritized by Hop in the different individuals of the cohort reveals that we do not observe significant differences between the patients and the controls in terms of number of instances, proportion of combinations predicted as disease causing by VarCoPP2.0 or the networks generated by the top combinations for each patient. To investigate potential differences in the content of these networks, we further look at the gene pairs shared between patients and controls in Section 6.7, and the genes shared between these groups in Section 6.9.

6.6 Investigation of diagnosed patients

In this section, we investigate how well our Hop approach performs on the patients that have been manually diagnosed by the team of professor Maris Laan [227, 407].

There were 8 patients with an oligogenic diagnosis (4 in the GEMINI patients group and 4 in the FIMM patients group) and 57 patients with a monogenic diagnosis (8 in the GEMINI patients and 49 in the FIMM patients group).

6.6.1 Patients with oligogenic diagnosis

Although, the GEMINI patients were excluded from the full cohort analysis, we report here on the 4 patients with oligogenic combinations from this cohort as well as the 4 patients with oligogenic diagnosis in the FIMM patients group.

Out of the 8 patients with identified oligogenic combinations as causative for the disease, we could only obtain the rank of the combinations for 6 individuals, since two variants which were part of the diagnostic oligogenic combination for samples 233 and 463 were removed during our filtering procedure (see Appendix E for the full details on the diagnosis variants). The sufficient coverage filtering step removed the variant in the SOS2 gene of Sample_233 and removed the variant in the WT1 gene for Sample_463.

Importantly, Sample_233 and Sample_119 were diagnosed using a panel designed for analyzing RASopathies in infertile men, which means that the panel used to diagnose these 2 samples manually was not the same as the panels used here for prioritization. This can thus potentially explain the lower rank obtained for the combination found in Sample_119.

Cohort	Sample_ID	Gene pair	Hop rank (No filter)	Hop rank (IM filter)
FIMM	Sample_119	TP63;SPRED1	85	24
		HUWE1;DYRK1A	27	9
FIMM	Sample_203	DYRK1A;DHX37	39	15
		HUWE1;DHX37	2	2
FIMM	Sample_274	FANCM;PROKR2	5	3
GEMINI	Sample_454	KMT2D;PROK2	119	11
GEMINI	Sample_487	NR5A1;PROP1	19	-
GEMINI	Sample_491	BRIP1;OTX2	60	7

Table 6.2: Rank of the manually identified oligogenic combination using the Hop predictor with the combined panel as seeds on the diagnosed patients. The genomic coordinates of the variants involved in the combinations are in the same order as the list of genes. The rank with no filter corresponds to the rank in all prioritized combinations while IM filter corresponds to the rank in the prioritized combinations when removing all combinations involving a single heterozygous variant in an AR gene. Dashes mean that the rank could not be obtained because one of the variants was filtered out.

For the remaining samples we observe that the combinations are rather well ranked by the predictor, and that all combinations are in the top 50 when only considering the results with the IM filter (Table 6.2).

In addition to these rankings which were obtained with the gene panel that resulted from the combination of 4 panels, we also analysed the ranks obtained when using each panel independently. The results are presented in Figure 6.7.

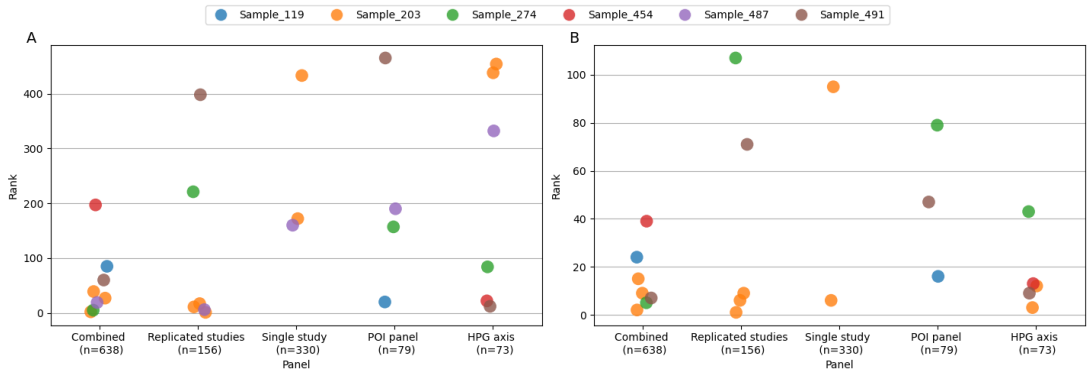


Figure 6.7: Ranks of the combinations found in the 6 samples with oligogenic diagnosis for which the variants had not been filter by our filtering protocol. Ranks are computed without taking into account inheritance mode of genes (A), and taking into account inheritance mode of genes (B). Each color represents a patient, Sample 203 had a diagnosis of 3 variants and is therefore represented by 3 dots.

It appears that the combination of panels is the most helpful panel for prioritization, while the panel containing genes that were only reported in a single study provides the least information. This was to be expected considering that the genes in which oligogenic diagnoses were found are part of either the Replicated studies panel (HUWE1, DHX37, FANCM, WT1, NR5A1), the POI panel (TP63), the HPG panel (PROK2, PROKR2, OTX2, PROP1) or none of the panels (SOS2, SPRED1, DYRK1A, KIF7, CHEK2, KMT2D, CHEK2, BRIP1).

We then analyze the top predicted combinations for each sample individually, to assess which combinations are predicted higher than the manually diagnosed one, and assess whether some variants might have been overlooked during manual diagnosis, as well as understand potential biases from Hop. For each patient, we assess the identified variants using Franklin [1] and OLIDA. Franklin is a platform that automatically classifies variants based on the ACMG criteria, and is thus useful to have an idea of the potential pathogenicity of a variant, even though these criteria are designed for the assessment of monogenic variants [78].

1. https://franklin.genoox.com/clinical-db/home

Sample_119

The top combinations prioritized for Sample_119 appear to be centered around the variant in gene TP63 which was identified as diagnostic variant (Figure 6.8). The majority of the variants found in combinations with this TP63 variant are found in genes that are not part of the panel provided as seeds with the exception of a compound-heterozygous variant in the LEO1 gene. The LEO1 gene was part of the gene panel including genes reported in a single study [227]. It was once reported as associated to severe oligosthenozoospermia, a subtype of oligozoospermia where in addition to having lower sperm counts, patients also exhibit decreased sperm motility [239]. The gene is reported to have AD inheritance mode, and Sample_119 carries compound heterozygous mutations in that gene. The first variant in the LEO1 gene (c.541G>A, p.D181N) is classified as VUS by the ACMG guidelines, and might therefore be relevant for further investigation. On the other hand, the second LEO1 variant (c.823A>G, p.Arg275Gly) is classified as Benign, fulfilling several evidence criteria such as being found in Homozygous state in large population databases, and having been reported as Benign in ClinVar.

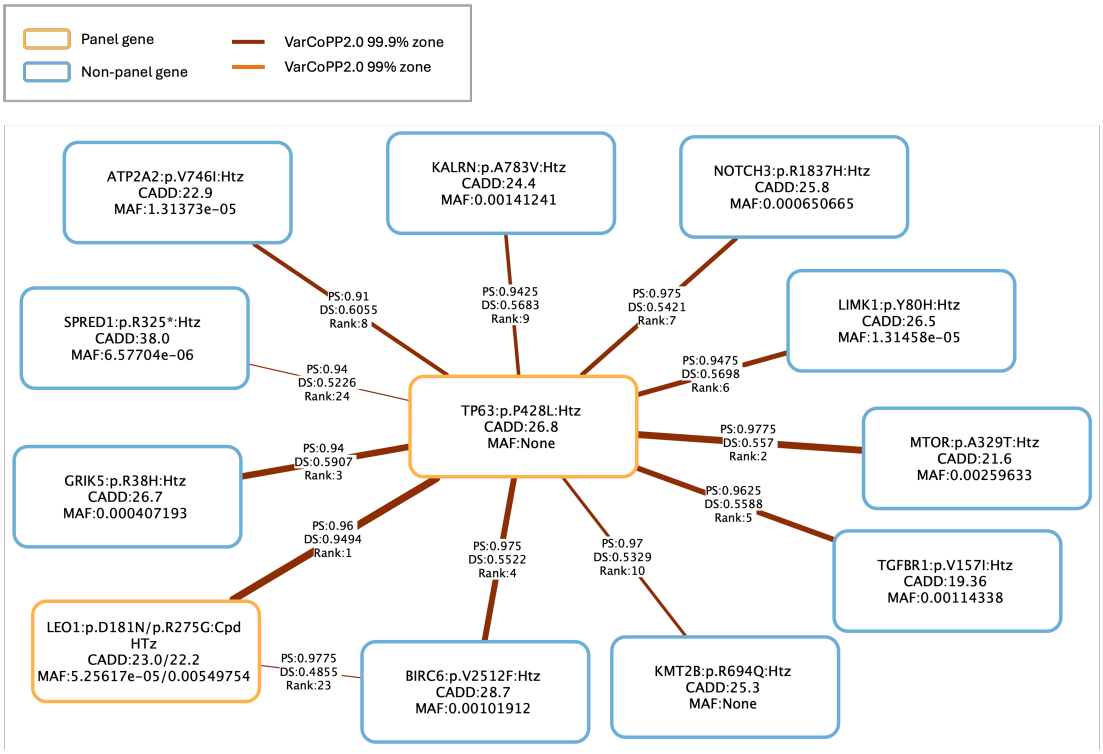


Figure 6.8: Network visualisation of the top combinations prioritised by Hop for Sample_119, after applying the inheritance mode filter. The top 10 combinations are shown in addition to the manually diagnosed combination (with rank 24). For each variant, the protein change is shown if available, otherwise the cDNA change is shown. Variants are also annotated with their CADD phred score and MAF in the GnomADv3.1. Nodes colored in yellow represent variants which are found in a gene from the gene panel and nodes colored in blue represent variants found in other genes. The weight of the edges is inversely proportional to the rank and the color correspond to the different VarCoPP2.0 confidence zones.

The remaining combinations all involve the TP63 variant and variants in genes that are not part of the gene panel and all combinations are predicted to be in the 99.9% confidence interval of VarCoPP2.0 (dark red edges in Figure 6.8, see Section 5.4.5 for a description of the confidence intervals). The difference in rankings between these variants is simply due to small differences in DS. The variant in SPRED1, which was identified as diagnostic variant, is a nonsense variant which is associated with a very high CADD score and would therefore be easily detected by manual analysis of this network of top combinations. We did not identify potentially relevant findings in the other variants (such as pathogenic or likely pathogenic variants, or variants previously reported to be involved in a similar phenotype) for this patient.

Sample_203

The network of the Top 15 prioritized combinations of Sample_203 appear to have a more complex topology, with many variant nodes connected to each other (Figure 6.9). The variants prioritized in the top 5 are variants present in panel genes (TP63, HUWE1, DHX37 and PPP1R12A), including two of the three diagnostic variants for this sample. We investigate the two variants in panel genes which were not reported as diagnostic variants. The TP63 variant (c.1531C>A, p.Pro511Thr) is classified as Benign according to the ACMG guidelines, and has been reported several times in ClinVar as either benign or likely benign. Interestingly, it has been identified in a patient from a cohort study on Müllerian anomalies, although it was not further investigated [418]. The PPP1R12A variant (c.1811G>C, p.G604A) is also classified as Benign according to the ACMG guidelines, as it has been observed in homozygous state in population databases more than expected for the disease and is generally predicted as neutral by *in silico* predictors.

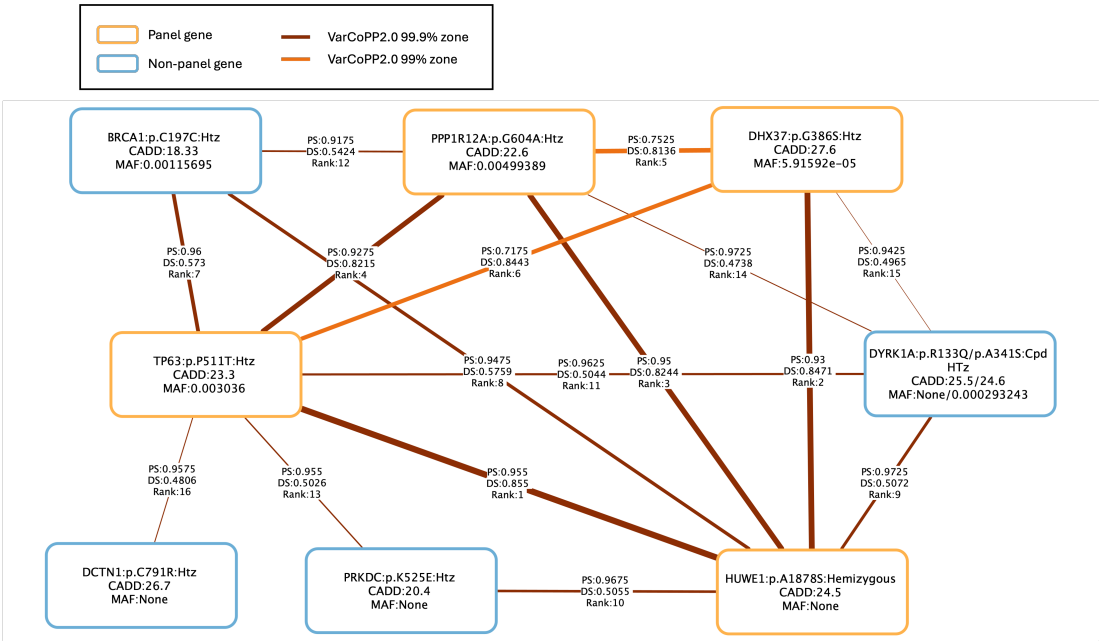


Figure 6.9: Network visualisation of the top combinations prioritised by Hop for Sample_203, after applying the inheritance mode filter. The top 15 combinations are shown, with the manually diagnosed combinations having ranks 2, 9 and 15. For each variant, the protein change is shown if available, otherwise the cDNA change is shown. Variants are also annotated with their CADD phred score and MAF in the GnomADv3.1. Nodes colored in yellow represent variants which are found in a gene from the gene panel and nodes colored in blue represent variants found in other genes. The weight of the edges is inversely proportional to the rank and the color correspond to the different VarCoPP2.0 confidence zones.

The combinations involving the compound-heterozygous variant in the DYRK1A gene, which was not part of the gene panel, are thus predicted with slightly lower ranks, although the combination involving the HUWE1 and the DYRK1A variants is present in the top 10.

Sample_233

For Sample_233, we could not obtain the rank of the manually diagnosed combinations since the major diagnostic variant in the SOS2 gene was removed by the coverage filter. The two other variants in genes CHEK1 and KIF7 are also not found in the top 50, which is probably due to the fact these variants act as modifiers of the main SOS2 variant and are therefore not likely to be prioritized in its absence [407]. We nevertheless investigate the top 10 combinations prioritized by Hop to analyze what other variants might be relevant. The network appears to be centered around variants in the TP63 and ZFPM2 genes which are present in the gene panel, and the majority of the combinations have a *PS* in the 99.9% confidence interval (Figure 6.10).

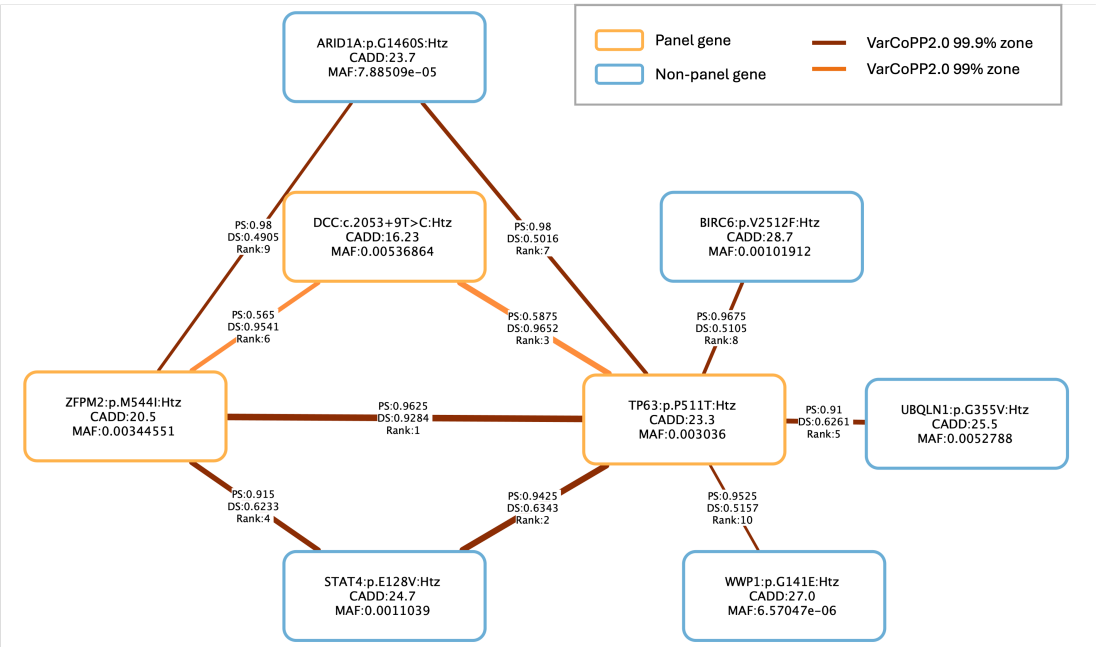


Figure 6.10: Network visualisation of the top combinations prioritised by Hop for Sample_233, after applying the inheritance mode filter. The top 10 combinations are shown. For each variant, the protein change is shown if available, otherwise the cDNA change is shown. Variants are also annotated with their CADD phred score and MAF in the GnomADv3.1. Nodes colored in yellow represent variants which are found in a gene from the gene panel and nodes colored in blue represent variants found in other genes. The weight of the edges is inversely proportional to the rank and the color correspond to the different VarCoPP2.0 confidence zones.

The TP63 variant was also found in Sample_203, and is classified as benign (see above for more detailed description). The ZFPM2 variant (c.1632G>A, p.Met544Ile) is also classified as benign by the ACMG guidelines, although it is linked to conflicting reports in ClinVar. Interestingly, it was found in homozygous state in a patient with 46,XY Disorder of Sexual Development (DSD), and *in vitro* assessment of the variant showed that it significantly reduced the interaction with GATA4 [419]. This interaction is known to regulate the expression of key genes involved in sex determination [420–422]. However, a more recent study reclassifies this variant as benign, based on updated *in silico* tools and *in vitro* assays of the molecular activity of the mutant in the context of gonadal signalling [423]. This highlights the importance of reproducing functional findings in order to correct potential errors. Notably, the variant was also reported in different studies as potentially disease causing in cardiac patients [424, 425]. This patient presented with a very severe phenotype including typical facial features of RASopathies, gonadal dysgenesis, impaired mobility and cognitive and intellectual performance. He had also been diagnosed with Lennox–Gastaut syndrome, characterized by early-onset epilepsy and pancreatic cancer [407]. The DCC variant (c.2053+9T>C) is classified as benign by the ACMG guidelines, and was reported several times as such in ClinVar. The BIRC6 (c.7534G>T, p.Val2512Phe) and WWP1 (c.422G>A, p.Gly141Glu) variants are VUS and could therefore be further investigated for potential involvement in the severe phenotype.

Sample_274

The network of prioritized combinations for Sample_274 only includes variants in genes that were part of the gene panel (Figure 6.11). All combinations therefore appear to have high DS although they have relatively lower PS compared to the other patient networks. Only the top ranked combination appears to be in the 99.9% confidence interval of the VarCoPP2.0 predictor, while the remaining combinations are in the 99% zone or simply predicted as disease causing. The two combinations that are ranked higher than the manually diagnosed combinations involve the FANCM variant, which is one of the two diagnostic variants, in combination with variants in the CHD7 and POLA1 genes.

The CHD7 variant (c.307T>A, p.Ser103Thr) has been reported several times in ClinVar as benign. Although it has been found in patients with IHH [426, 427], it was shown to be able to rescue the CHD7 knock-out phenotype in a zebrafish model [427], and can thus be considered as benign. The fact that this variant is ranked highly by Hop is normal since the gene is indeed relevant for the disease (with reported disease-causing variants in the gene) and the CADD score is relatively high, which was shown to be an important feature for the VarCoPP2.0 model (see Section 5.4.6).

The POLA1 variant (c.4356C>T, p.Tyr1452=) is a synonymous variant, with a relatively low CADD score, and is classified as benign by the ACMG guidelines. The fact that this combination is predicted with relatively high *PS* is surprising. To further investigate, we looked at the feature contribution analysis for this specific combination, which reveals that the main features driving the prediction are the pair features, since the genes are close according to the biological distance and have high biological process similarity.

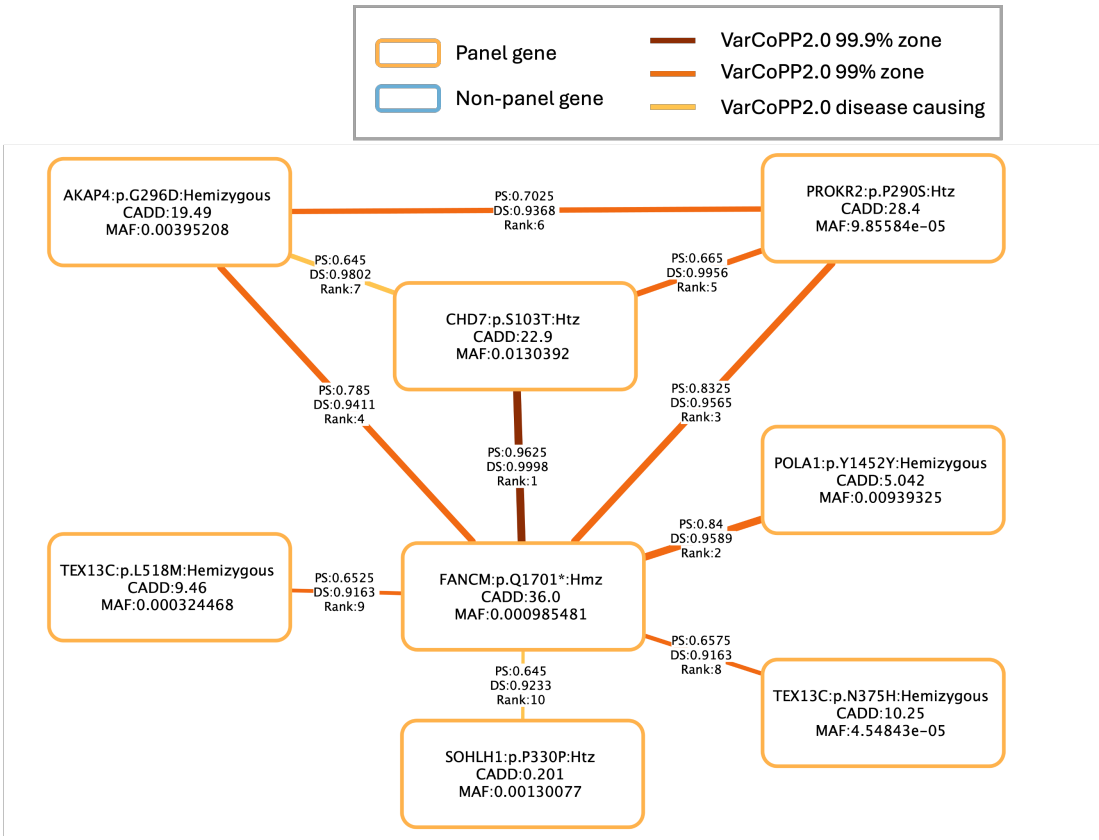


Figure 6.11: Network visualisation of the top combinations prioritised by Hop for Sample_274 after applying the inheritance mode filter. The top 10 combinations are shown, with the manually diagnosed combination at rank 3. For each variant, the protein change is shown if available, otherwise the cDNA change is shown. Variants are also annotated with their CADD phred score and MAF in the GnomADv3.1. Nodes colored in yellow represent variants which are found in a gene from the gene panel and nodes colored in blue represent variants found in other genes. The weight of the edges is inversely proportional to the rank and the color correspond to the different VarCoPP2.0 confidence zones.

Sample_454

The network of prioritized combinations for Sample_454 is centered around variants in genes PROK2 and KLB, which are both part of the combined panel (Figure 6.12). The PROK2 variant is a diagnostic variant, in combination with the KMT2D variant. The KLB variant (c.2329_2331del, p.Phe777del) is classified as likely benign by the ACMG criteria, and has once been reported as VUS in Clinvar. It has been identified in homozygous state in large population datasets.

The remaining variants are found in genes that are not part of the gene panel, with 4 variants linked to both PROK2 and KLB variants (including the KMT2D diagnostic variant) and one variant linked only to the PROK2 variant and one variant linked only to the KLB variant. The KMT2D variant has the highest CADD score among these and would therefore quickly be identified by manual analysis. All the variants in the non-panel genes, except for the NEDD4L variant, are classified as VUS and are therefore relevant for further investigation in an oligogenic inheritance context.

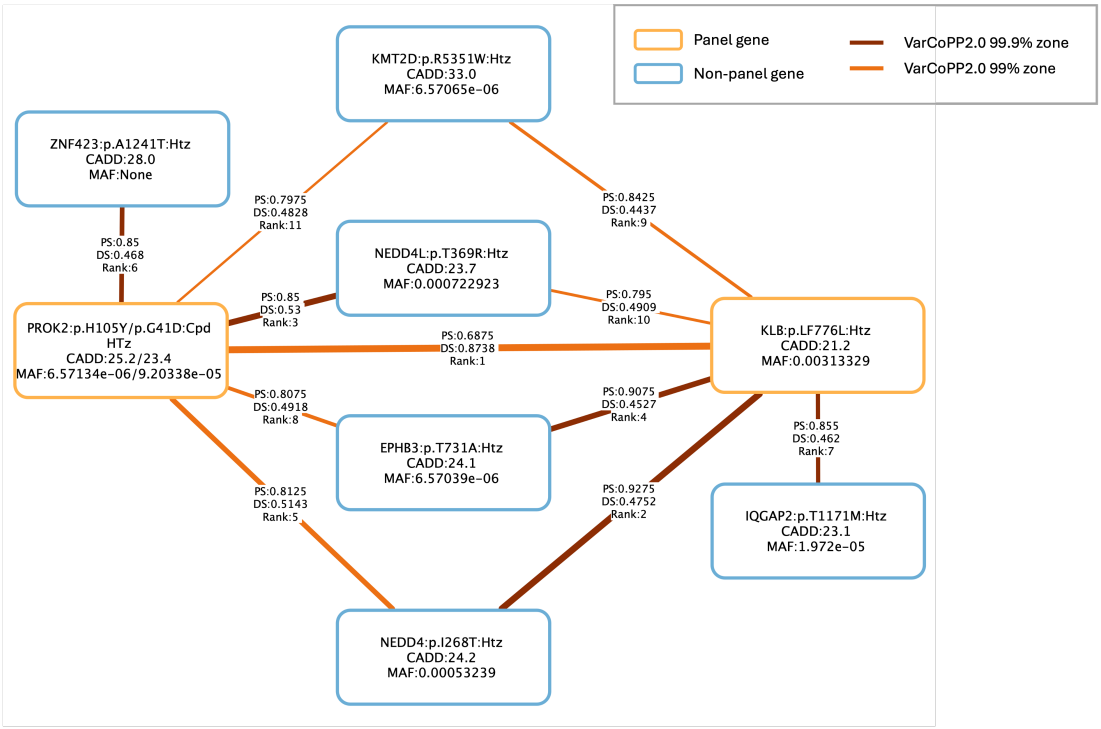


Figure 6.12: Network visualisation of the top combinations prioritised by Hop for Sample_454 after applying the inheritance mode filter. The top 11 combinations are shown, with the manually diagnosed combination at rank 11. For each variant, the protein change is shown if available, otherwise the cDNA change is shown. Variants are also annotated with their CADD phred score and MAF in the GnomADv3.1. Nodes colored in yellow represent variants which are found in a gene from the gene panel and nodes colored in blue represent variants found in other genes. The weight of the edges is inversely proportional to the rank and the color correspond to the different VarCoPP2.0 confidence zones.

Sample_463

The network of top 10 combinations for Sample_463 only contains one variant in a gene from the gene panel (Figure 6.13, TUBB3 variant). This variant is intronic, has a relatively low CADD score, and is classified as benign by the ACMG guidelines. The rank of the manually diagnosed combination could not be obtained since the CHEK2 variant was filtered out during the coverage filtering step. The prioritized combinations therefore appear to either have high PS (5 combinations are in the 99.9% confidence zones) but low DS or have higher DS but low PS (the combinations are not in any confidence interval).

This type of output from Hop brings back the idea, discussed in Section 5.10 that a prioritization algorithm will always provide a ranking, even though the identified combinations are not necessarily relevant (here they are either not likely to be disease causing or not likely to be relevant to the disease), and we should thus be careful about the outputed scores.

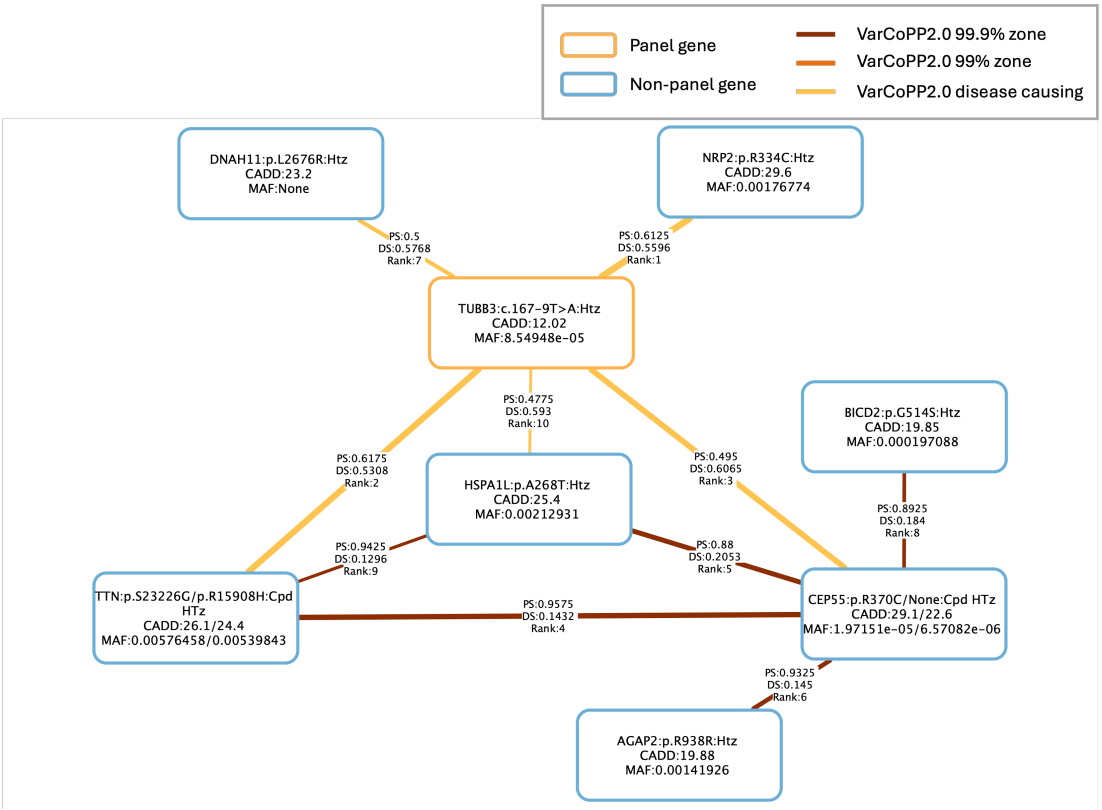


Figure 6.13: Network visualisation of the top combinations prioritised by Hop for Sample_463 after applying the inheritance mode filter. The top 10 combinations are shown. For each variant, the protein change is shown if available, otherwise the cDNA change is shown. Variants are also annotated with their CADD phred score and MAF in the GnomADv3.1. Nodes colored in yellow represent variants which are found in a gene from the gene panel and nodes colored in blue represent variants found in other genes. The weight of the edges is inversely proportional to the rank and the color correspond to the different VarCoPP2.0 confidence zones.

Sample_487

For Sample_487, the manually diagnosed combinations included a heterozygous variant in the PROP1 gene which is known to be AR. We therefore analyze the top prioritized combinations for that patient without applying the IM filter. The majority of the variants identified in the top 19 are located in genes that are part of the gene panel (Figure 6.14). However, it is important to note that, except for the NR5A1 gene (which contains a diagnostic variant) and the GLI2 gene, the remaining genes in the network have an AR mode of inheritance. The variant in the GLI2 gene (c.392G>A, p.Arg131His) is classified as VUS by the ACMG criteria and was reported as such in ClinVar, and might thus be worth further investigation due to its high CADD score.

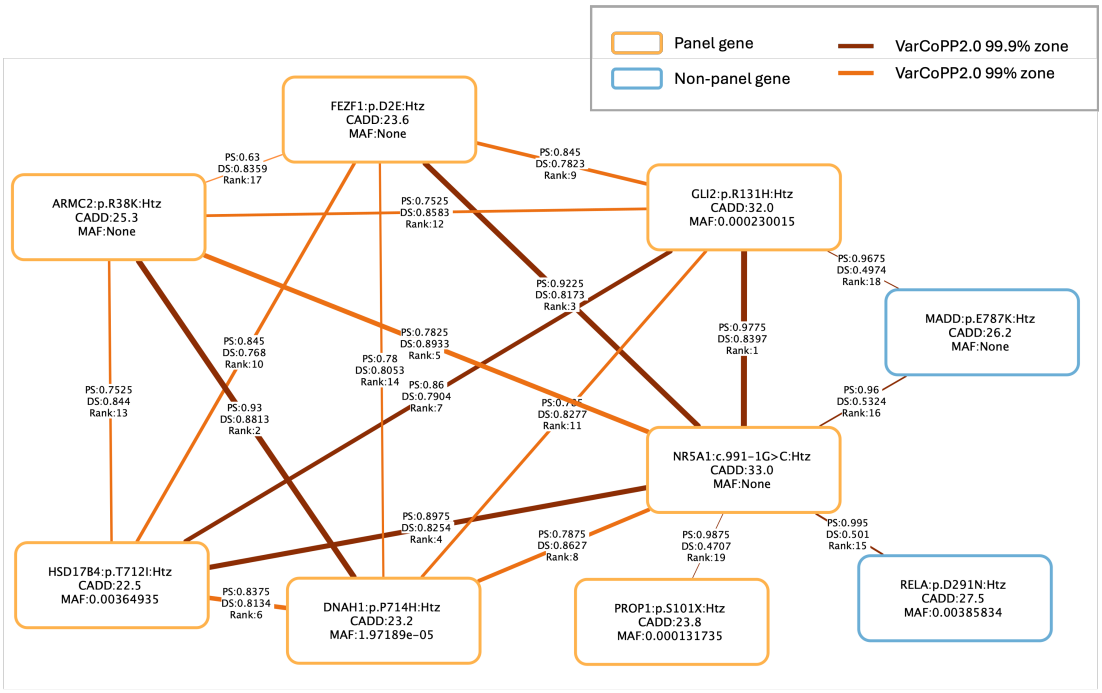


Figure 6.14: Network visualisation of the top combinations prioritised by Hop for Sample_487 without applying the inheritance mode filter. The top 19 combinations are shown with the manually diagnosed combinations having rank 19. For each variant, the protein change is shown if available, otherwise the cDNA change is shown. Variants are also annotated with their CADD phred score and MAF in the GnomADv3.1. Nodes colored in yellow represent variants which are found in a gene from the gene panel and nodes colored in blue represent variants found in other genes. The weight of the edges is inversely proportional to the rank and the color correspond to the different VarCoPP2.0 confidence zones.

Sample_491

The network of prioritized combinations in Sample_491 is centered around the OTX2 variant, which is located in a gene present in the gene panel and which is one of the diagnostic variant for this patient (Figure 6.15). Two other prioritized variants are within genes from the gene panel and therefore have higher *DS* scores, although their *PS* scores are lower than the one of the manually diagnosed combination (OTX2;BRIP1).

The TKTL1 variant (c.188A>G, p.Tyr63Cys) is classified as VUS by the ACMG guidelines and was never reported in ClinVar or in other publications. This gene was only recently reported in a study of patients with azoospermia, and has been shown to have particularly high expression in the testis [428]. On the other hand, the SPECC1L (c.2882C>T, p.Ala961Val) is classified as benign and was previously reported as such in ClinVar.

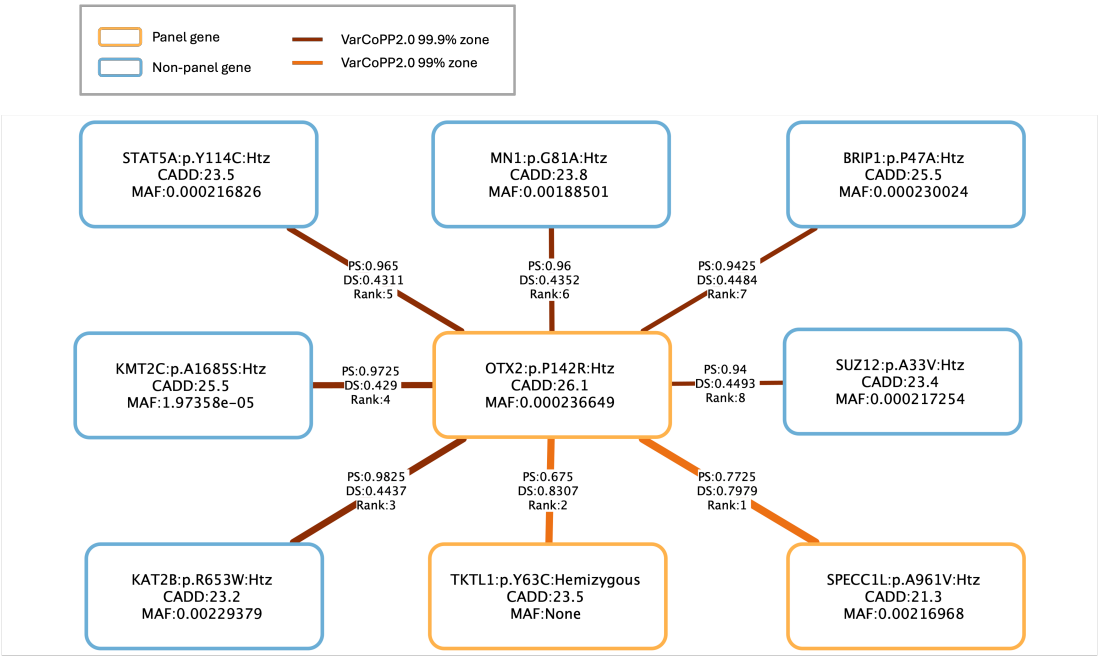


Figure 6.15: Network visualisation of the top combinations prioritised by Hop for Sample_491 after applying the inheritance mode filter. The top 8 combinations are shown, with the manually diagnosed combination at rank 7. For each variant, the protein change is shown if available, otherwise the cDNA change is shown. Variants are also annotated with their CADD phred score and MAF in the GnomADv3.1. Nodes colored in yellow represent variants which are found in a gene from the gene panel and nodes colored in blue represent variants found in other genes. The weight of the edges is inversely proportional to the rank and the color correspond to the different VarCoPP2.0 confidence zones.

We further analyzed the significance of the variants that are prioritized higher than the BRIP1 variant, since some had higher *PS*. The KAT2B variant (c.1957C>T, p.Arg653Trp) and the MN1 variant (c.242G>C, p.Gly81Ala) were both classified as benign according to the ACMG guidelines, and reported in ClinVar as such. The KMT2C variant (c.5053G>T, p.Ala1685Ser)

is classified as VUS, and was identified as associated to increased Body Mass Index (BMI) in an exome wide association study performed in japanese individuals [429]. Since this patient is also affected with obesity, this variant might thus explain this part of his phenotype. Finally, the STAT5A variant (c.341A>G, p.Tyr114Cys) is also classified as VUS. The STAT5A gene has been found to affect female fertility in mice models, but no link has yet been made with male infertility.

The detailed analysis of each patient network presented here indicates that in most cases, the combinations prioritized by Hop for each individual make sense, and include the manually diagnosed combinations. This type of visualisation can therefore be used by clinicians to potentially diagnose patients. The disease-causing variants can indeed easily be identified in most networks, as they either involve a highly connected variant, or involve variants with very high CADD scores, standing out from the other nodes in the network. As we have seen in the analysis of Sample_463, this visualisation also indicates when the ranking produced is not relevant, as scores appear to be much lower. Even though manual analysis of such networks for all patients would be time consuming, we believe it might still represent an improvement over achieving manual diagnosis by looking solely at lists of variants, as it provides a visual representation of the predictions.

6.6.2 Patients with monogenic diagnosis

Overall, in 57 patients, a monogenic diagnosis was obtained through manual analysis [227, 407]. This means that one or two variants in a single gene were found to explain the patient's disease. The list of patient and diagnosis can be found in Appendix E. The majority of the diagnoses were found in the FIMM patients cohort, with only 8 patients from the GEMINI cohort having a monogenic diagnosis.

We investigated the ranks of the monogenic diagnostic variants in the Hop predictions. A small proportion (10/54) of diagnostic variants were removed by our VCF filtering procedure. These variants, as well as the filters that removed them, are detailed in Appendix E. Therefore, for 11 patients (6 from the FIMM cohort and 5 from the GEMINI cohort), the rank of the monogenic diagnostic variants could not be retrieved by the predictor. For the remaining patients, we computed the rank of the variants as the rank of the first combination in the prioritized combinations that contained this variant. This was done with and without applying the IM filter and for the different gene panels.

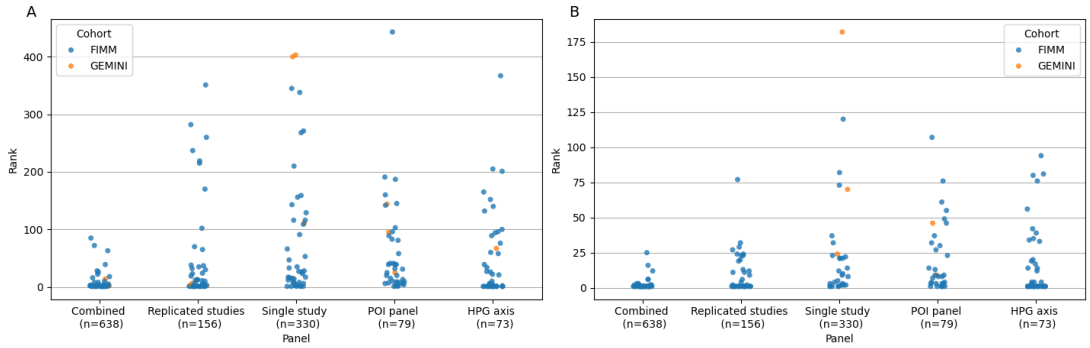


Figure 6.16: Rank of the first combination including the monogenic diagnostic variant, without applying any filtering on the results (A) and by filtering the variants based on the genes inheritance mode (B). The results are shown using the different panels as seeds for the prioritization.

Similarly to the oligogenic variants, the combinations involving monogenic diagnostic variants appear to be ranked highly in the predictions, with the majority of the combinations being ranked in the top 25 when using the combined panel as seeds (Figure 6.16). Here again it appears that ranking using the combined gene panel as seeds lead to the best retrieval of known combinations.

To better visualise the proportion of combinations that were ranked in the top 100 by Hop, we compute the CDF curves, similarly to what was done in Chapter 5, for the rankings generated with the different panels and using the two different filtering of the results (Figure 6.17).

Using the combined panel, for all but one patient, the first combination containing the monogenic diagnostic was found in the top 100 in the unfiltered results and in the top 50 when filtering the results using the IM filter. The diagnostic variant that could not be retrieved is a hemizygous variant in the GLUD2 gene (c.412C>T p.Gln138*), which was found as diagnostic for Sample_114.

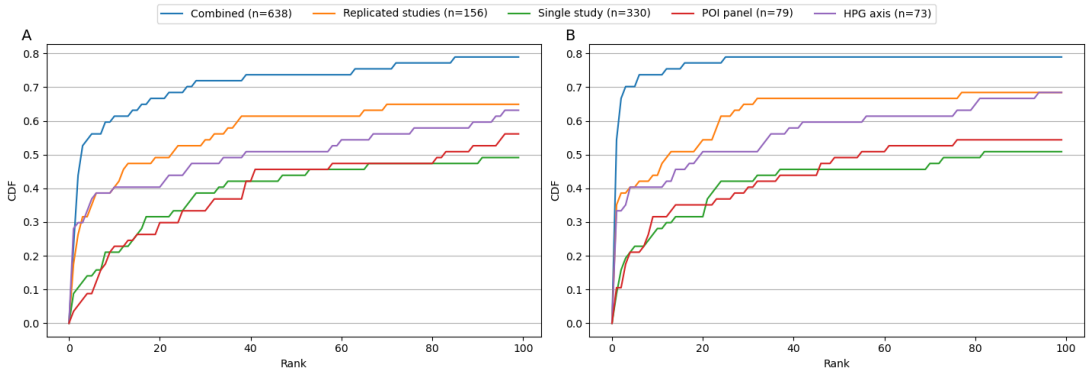


Figure 6.17: CDF plot of the performance of Hop in identifying the monogenic variants found in the ESTAND cohort. Each curve represent the use of a different gene panel as seeds for the prioritization. The CDF curve represents the proportion of patients for which the known monogenic variant was found in one of the combinations higher than the rank value shown on the x-axis (varied between 1 and 100). Since for 11/57 patients, the diagnosed monogenic variant has been filtered out, the maximum proportion of diagnosed patients that could be found with Hop is 0.81.

In addition to investigating the ranks of the monogenic variants, we checked whether the networks of the top prioritized combinations in patients with monogenic diagnoses have a “star” topology and are centered around the monogenic diagnostic variants. In the networks generated by the top combinations when not taking into account the inheritance mode of the genes, only one sample had a star network topology. However, when applying the IM filter, 17, 15 and 7 patients out of the 57 patients with monogenic diagnostic variants had a “star” network topology structure for the networks generated by the top 10, top 20 and top 50 combinations respectively. This represents 29.8%, 26.3% and 12% of the patients respectively, which is higher than the percentage of samples having such network structures when considering all patients (see Section 6.5.3). Furthermore, the diagnostic variant was the variant with the most connections in the network for 21 patients when looking at the top 10 and 18 patients when looking at the top 20 and top 50 combinations. This indicates that some of these monogenic diagnostic variants can act as “anchors” to the predictions, leading to the prioritization of multiple variant combinations involving the diagnostic variant together with potential modifiers.

Analysis of the manually diagnosed patients using our Hop predictor therefore allows to validate that the combinations detected by the predictor match with the manually detected combinations and confirms the relevance of our tool for the identification of disease-causing variants in the ESTAND cohort.

6.7 Analyzing shared gene pairs among patients

Following from these preliminary results, we therefore move towards the investigation of general patterns of digenic variant combinations in the cohort based on the Hop predictions in all patients. All results presented onwards are based on the prioritized combinations with the combined gene panel collecting all 638 candidate genes for male infertility as this panel seemed to yield better results. We however continue to compare the effect of applying or not applying the IM filter.

6.7.1 Patients share more gene pairs than controls

We first look at how many gene pairs are shared between patients and between controls in the top 10, top 20 and top 50 prioritized combinations. We define a gene pair as shared by two individuals if it contains at least one variant combination that is prioritized in the top K combinations of both individuals. This is done in the results filtered with the IM filter.

The number of distinct gene pairs that are shared by more than N samples, for N varied between 2 and the maximum number of samples sharing a particular gene pair is shown in Figure 6.18. Overall, when considering the top 10 combinations, 194 pairs are shared by at least 2 patients (out of 3882 distinct gene pairs present in the top 10 of all patient samples), while 94 gene pairs are shared by at least 2 control samples (out of the 3005 distinct gene pairs found in the top 10 of all controls individuals). Furthermore, based on the top 10, 246/429 (57%) patient individuals share at least one gene pair with another patient, while only 124/322 (38.5%) control individuals share at least one gene pair with another control individual. In the top 20 and top 50, we observe more shared gene pairs among at least 2 samples. Nevertheless, even when considering the top 50, there are 36 patients (8.4%) which do not share any gene pair with another patient sample, and 45 control samples (14%) which have only have variants in unique gene pairs. The maximum number of patients sharing a particular gene pair is 6 (and this is the case for 4 distinct gene pairs in the top 50), while at most 5 control samples carry a variant combination in the same gene pair.

It therefore appears that patient share more gene pairs in their top 10, 20 and 50 than control samples, yet, since the patient group is larger than the control group, this might be due to chance only. To assess whether this difference is statistically significant due to differences in the groups, and thus to investigate whether this is due to the sharing of potentially relevant gene pairs among patients, we perform a permutation analysis. The “patient” and “control” labels of the samples are randomly shuffled 10,000 times and each time, we recompute the proportion of shared gene pairs in the top 10, top 20 and top 50 combinations of each set. We then compare the z-score obtained from the true proportion of shared combinations between patients and controls with the 10,000 z-scores obtained from the random sets and computed a $p-value$ as the number of times these permuted z-scores are higher than the true z-score divided by the number of permutations.

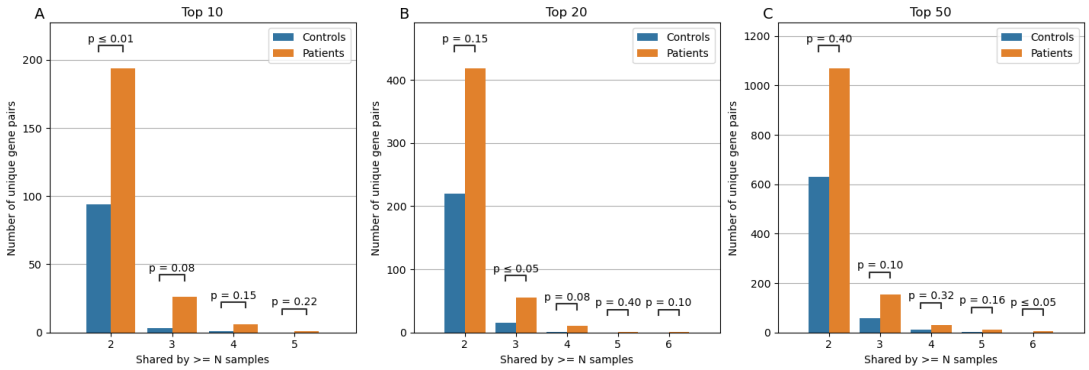


Figure 6.18: Number of distinct gene pairs that are shared by more than N patients or controls, for varying values of N, when considering combinations in the top 10 (A), top 20 (B) and top 50 (C) variant combinations. A gene pair is shared between two patients if each patient has at least one variant combination in that gene pair in their top K prioritized combinations for the three aforementioned values of K. P-values represent the statistical significance of the difference between the counts, by assessing the difference in ratio of shared combinations among patients and controls groups using a permutation test.

This indicates that although the number of gene pairs shared by the patient group seem to be much larger than the number of gene pairs shared among the control group, these differences are mostly due to the difference in sample size between the groups. We only observe a few statistically significant differences (Figure 6.18). For instance, there seem to be more gene pairs that are shared by at least two patients in the top 10, as well as significantly more gene pairs that are shared by at least 3 patients when considering the top 20 of each sample (Figure 6.18).

This analysis suggests that patients share more gene pairs than controls, which indicates that potential oligogenic signatures are present in the patient group, and that they are prioritized by our method. The control group does not present with such enrichment signal, since the variant combinations that are prioritized in these samples are mostly due to chance and biases of the method, and there are thus less common findings between the different samples.

To further analyze the difference between these shared gene pairs, we visualise the network of gene pairs that are shared by more than 3 samples in the patient and controls sets, when taking into account the top 20 combinations of each sample, since this top showed a statistically significant difference between patients and controls (Figure 6.18B). Each node in this network represents a gene and each connection between two genes represent a pair that is shared by more than 3 samples. By making the size of each node in the network proportional to the degree of the node (i.e. the number of shared gene pairs it is a part of) we can visualise the genes that appear to be the most important genes among these shared combinations. This is done for both the patients and controls set separately, to assess whether we distinguish different genes in the two sets (Figure 6.19).

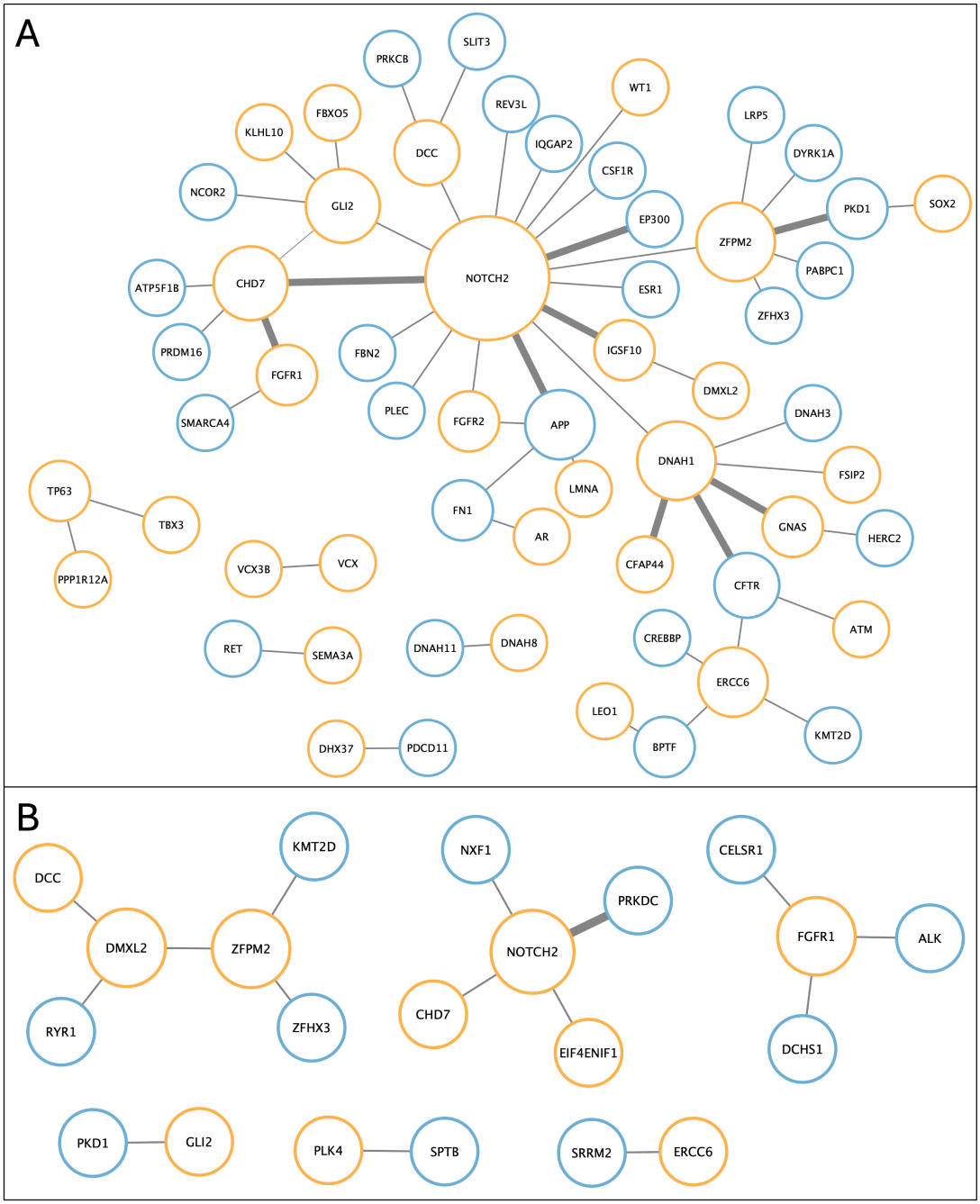


Figure 6.19: Networks of gene pairs shared by at least three samples when considering the top 20 of each sample, for the patient group (A) and the control group (B). The orange nodes are genes that are part of the gene panel used as seeds for the prioritization while the blue nodes are novel genes. The size of the node is proportional to the degree of the node and the width of the edges is proportional to the number of patients or controls that share the gene pair.

We observe that in addition to being larger, the network of gene pairs shared between patients is more connected than the network of gene pairs shared between controls (Figure 6.19). Some genes appear to be common between both networks, such as NOTCH2, which is the gene with the highest degree in both the patients and control sets (Figure 6.19). There also appear to be common gene pairs, such as the NOTCH2;CHD7 pair which is shared by 4 patients and 3 controls. This indicates that not all gene pairs that are shared between patients are relevant for further analysis as they might be due to biases of the predictor, which ranks variant combinations in these gene pairs highly in both patient and control samples. These are probably due to biases in the DS, as discussed in Chapter 5, and highlights the importance of performing case-control studies to correct for these biases.

Noticeably, in the patient network, the APP and CFTR genes, which were not in the gene panel, appear to be common to several shared gene pairs. CFTR is a well known gene associated with male infertility (see Section 1.4), but was excluded from the panel of [227] since variants in this gene were assessed prior to the study and no pathogenic or likely pathogenic variants based on the ACMG criteria had been identified. However, since these criteria are for the assessment of monogenic variants, other types of variants in this gene might be relevant for an oligogenic model and might have been picked up by our approach. The APP gene encodes for the Amyloid precursor protein, which is mostly known for playing a central role in the development of Alzheimer's disease [430]. However, this protein is expressed in multiple tissues outside of the nervous system and several lines of evidence have started to show its potential implication in fertility, such as the fact that it is expressed in the rat testis and seem to be important for spermatogenesis [431], the observation of infertility in mice with double knock-out of APP and APLP2 [432], and analysis of the APP interactome in testes [433].

6.7.2 Gene pair enrichment analysis

Following the observation that patients appear to share certain gene pairs in their top prioritized combinations, we note that some of these gene pairs might be due to biases from the predictor, as they are also found in the control set. To address this, in this section, we analyze which gene pairs appear to be significantly more frequent in patients compared to controls.

To do so, we perform the following permutation procedure [369]:

1. We collect the most frequent gene pairs in the Top10, Top20 and Top50 prioritized combinations of the patient cohort.
2. We calculate the frequency of these gene pairs in the equivalent ranking in the control population.
3. For each gene pair, we test whether the frequency of the gene pair in the patient population is significantly higher than the frequency of the gene pair in the control population, by using a normal one sided test for proportions, and calculate the z-score of the test.

4. We then perform a sample permutation, by randomly shuffling the "patient" and "control" labels of the samples 1000 times. For each iteration n (n), we re-computed the frequency of the gene pairs in the patients and controls sets and computed the associated $z-score_n$.
5. Using these permutations, we compute a $p-value$ for each gene pair according to the following equation

$$p-value = \frac{\sum_{n=1}^{N} I(z-score_n \geq z-score)}{N}$$

where N represents the number of permutations. This $p-value$ thus represents the proportion of permutations for which the $z-score$ obtained with the shuffled labels is larger than the actual $z-score$.
6. We consider gene pairs with a $p-value < 0.05$ as significantly more frequent in patients, and relevant for further investigation.

This is done first for all patients compared to controls and then for the different patient subgroups.

6.7.3 Enriched gene pairs in infertile men compared to controls

First, we assess whether we could discover enriched gene pairs across the whole cohort. The procedure described above is thus applied to the whole FIMM patients cohort of 429 infertile men compared to the 322 control individuals sequenced as part of the GEMINI study. The significant gene pairs for the IM filtered and unfiltered datasets and for the different tops are shown below (Table 6.3).

Only two gene pairs appeared to be significantly enriched among patients compared to control individuals when filtering variants according to the mode of inheritance, and 6 distinct gene pairs are significantly enriched when not taking into account the inheritance mode of the genes (Table 6.3). The fact that we do not observe many significantly enriched gene pairs is probably due to the low statistical power induced by the sample size. With 429 patients and 322 control individuals, a gene pair has to be shared by a minimum of 6 patients for it to be significantly enriched (assuming no control is found to carry a variant combination in this gene pair). As we have seen from the number of shared gene pairs among patients (Section 6.7.1), there are no such gene pairs in the top 10, and only one such gene pair in the top 20 and 4 gene pairs in the top 50. Given the high heterogeneity of male infertility genetics and the phenotypic heterogeneity of the ESTAND cohort, obtaining such as small number of statistically significant gene pairs is therefore not surprising.

Unfiltered				Inheritance mode filtered			
Gene pair	$p-value$	Num. patients carriers	Num. Control carriers	Gene pair	$p-value$	Num. patients carriers	Num. Control carriers
Top 10							
Top 20							
TTC21A;DNAH1	0.0308	6	0	CHD7;GLI2	0.035	6	0
SLX4;RECQL4	0.036	6	0				
Top 50							
DNAH11;DNAH8	0.0012	13	0	NCOR1;NOTCH2	0.0348	6	0
DNAH1;DNHD1	0.0168	7	0	CHD7;GLI2	0.035	6	0
TTC21A;DNAH1	0.0182	7	0				
GNAS;DNAH1	0.031	6	0				

Table 6.3: Significantly frequent gene pairs in the top 20 and top 50 ranking of the patients of the ESTAND cohort. Genes colored in blue are not part of the initial panel.

We therefore further analyze these gene pairs, by looking, for each patient, at the variant combinations that are found within these gene pairs and the characteristics of those combinations, as well as the detailed scores that led to their prioritization (i.e. the DS and PS for each combination). These results are presented in Figure 6.20 for the CHD7;GLI2 pair and Figure 6.21 for the NCOR1;NOTCH2 pair. The detailed results for the gene pairs found to be statistically significant without taking into account the inheritance mode of the genes are shown in Appendix F, and are briefly discussed at the end of this section.

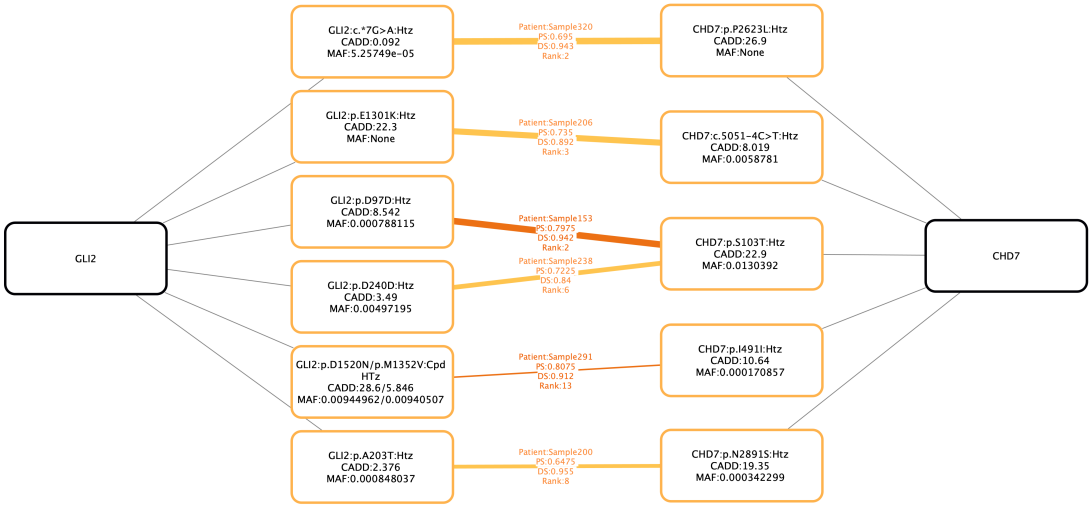


Figure 6.20: Variant combinations identified in the gene pair GLI2;CHD7 in different patients of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS score and DS score. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

Details on the variants found in patients in the CHD7;GLI2 pair show that part of these combinations are probably false positives due to the high contribution of the DS to the prioritization. Indeed, in 5/6 combinations, one of the variant involved has a very low scaled CADD score (<10, which we would consider the threshold for a variant involved in an oligogenic combination). This is due in some cases to these variants being synonymous variants, or intronic variants. Furthermore, the PS of these combinations is relatively low, with only 2 out of 6 combinations being in the 99% confidence zone. The GLI2 and CHD7 genes are both in the gene panel used for the prioritization and have well established links to male infertility. Additionally, each gene is linked respectively to 16 and 70 oligogenic combinations in OLIDA, out of which 2 variant combinations are found to involve both these genes (OLI1746 and OLI575) [253]. These combinations are reported as causative of DSD in two distinct patients [193, 434], and involve CHD7 and GLI2 variants together with other genes. However, both combinations are associated with confidence scores of 0, indicating that they are not associated to sufficient evidence for confirmed pathogenicity involvement according to our criteria. This also means that these combinations were not used in the training set of VarCoPP2.0, and that this finding is not due to biases in our tools.

We investigate further each variant identified using Franklin to assess whether the intronic variants are predicted to affect splice sites, and also assess the ACMG guidelines status of these variants. We also search for the presence of these specific variants in OLIDA to check whether they have previously been reported as oligogenic causes.

In GLI2, most variants are classified as likely benign, and have been reported as such in ClinVar. This is the case for the p.D97D, p.D240D, p.A203T, p.D1520N and p.M1352V variants. However, the reports in ClinVar were mostly assessing the involvement of these variants in Holoprosencephaly, a birth defect that appears to be unrelated to male infertility. The intronic variant c.*7G>A is classified as VUS by the ACMG guidelines, although it is predicted benign by SpliceAI [435]. It is rare in population databases and was never reported in ClinVar. The variant p.E1301K is also classified as VUS, it presents with low frequency in population databases and was also never reported in ClinVar.

Interestingly, one of the GLI2 variants (c.4558G>A, p.Asp1520Asn) is found in the OLIDA combination OLI1159, as it was recently reported in a patient with Kallmann syndrome, in combination with a FGFR1 variant. This combination is associated with a confidence score of 0 and is thus not included in the VarCoPP2.0 training set. In our cohort, this variant is prioritized as part of a compound heterozygous variant and in combination with a synonymous CHD7 variant (c.1473C>T, p.Ile491=), which has been reported as benign and likely benign in ClinVar in relation to Charge syndrome and CHD7 related condition.

In CHD7, 3 out of the 5 variants are also found in OLIDA, indicating that they were previously reported as an oligogenic cause to disease. In particular, the p.Asn2891Ser variant is found in two combinations (OLI1294 and OLI1301), with the OLI1294 obtaining a FINALmeta score of 1. It is important to note that this combination was added in OLIDAv3 and that this variant was therefore not present in the training set of VarCoPP2.0. The remaining variants (c.5051-4C>T found in OLI1600 and c.307T>A,p.Ser103Thr found in OLI1172) are found in combinations that have a final confidence score of 0. These combinations were however reported in individuals affected with either Kallmann syndrome or Hypogonadotropic Hypogonadism, which are known to cause infertility.

The c.7868C>T, p.P2623L variant, which is not found in OLIDA, is classified as VUS by Franklin and has been reported as such in ClinVar for the Hypogonadotropic Hypogonadism phenotype.

Overall, it therefore appears that although the variant combinations found within the GLI2;CHD7 gene pair initially did not seem relevant due to the low CADD scores of the variants and the relatively low VarCoPP2.0 scores, several of these variants were previously reported as part of oligogenic variant combinations found in patients with similar phenotypes, and a few of the other variants are classified as VUS using ACMG criteria, meaning that their potential role in the phenotype of the patients need to be further assessed. The CHD7 and GLI2 genes together are good candidates as they have also been reported in the same gene combination in OLIDA. The biological mechanisms that could potentially explain their joint involvement in the phenotype are further explored in Section 6.8.

The NOTCH2;NCOR1 pair involves the NOTCH2 gene which is part of the gene panel and NCOR1, which is not present in the candidate gene list. We here observe that the variant combinations found have higher PS, and that the majority of combinations (4 out of 6) are in the 99.9% confidence interval of the VarCoPP2.0 predictor (Figure 6.21). However, the combinations present with lower DS as they involve variants in a gene that is not in the original panel.

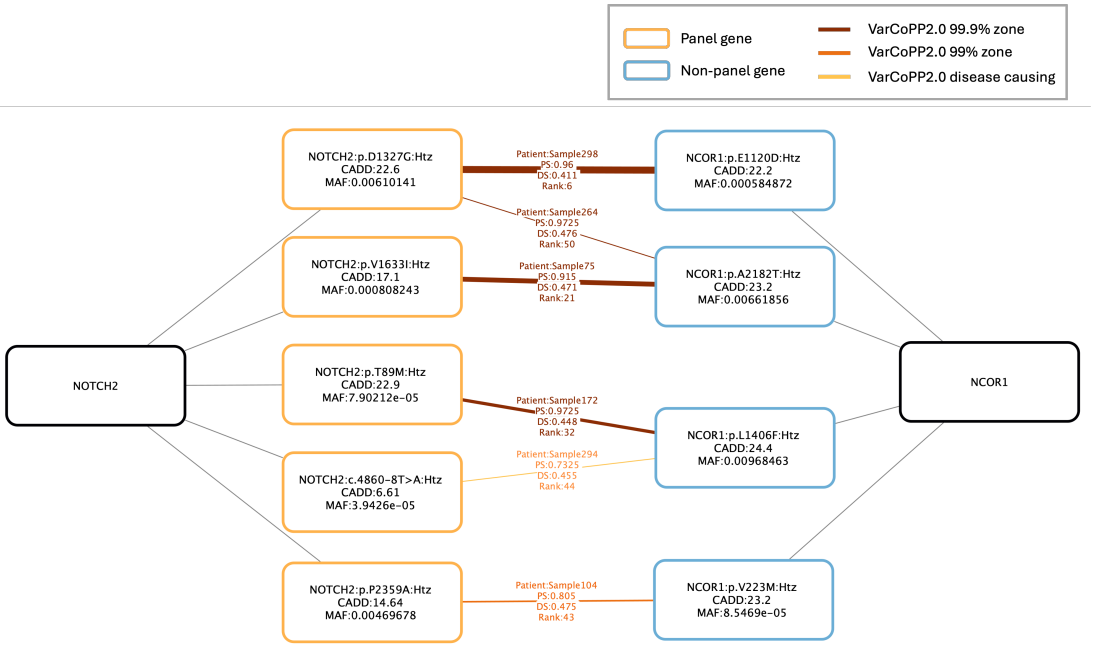


Figure 6.21: Variant combinations identified in the gene pair NCOR1;NOTCH2 in the top 50 of 6 patients of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

We here again proceed to further analyze the variants involved in the combinations found in this gene pair using Franklin for monogenic interpretation and OLIDA for searching for previous reports of these variants in oligogenic cases.

Although NCOR1 is not included in the gene panel in [227], it has been reported in OLIDA as part of a variant combination that was found in a DSD patient. This combination (OLI573) involves two NCOR1 variants including the NCOR1:c.6544G>A,p.Ala2182Thr, which is also found here in the prioritized combinations of Sample75 and Sample264 (Figure 6.21). OLI573 involves 6 genes and was assigned a final confidence score of 0. According to Franklin, this NCOR1 variant is nonetheless classified as benign as it has been reported as such in ClinVar.

In NOTCH2, one variant is found in OLIDA, the c.3980A>G, p.Asp1327Gly which is found in OLI601, a combination involving 7 genes in a DSD patient. This combination is also associated with a final score of 0, although it has a functional gene score of 1 and statistical score of 1. For the remaining NOTCH2 variants, apart from the c.266C>T, p.Thr89Met variant which is classified as VUS and was never reported in ClinVar, all other variants (c.4680-8T>A, c.4897G>A;p.Val1633Ile and c.7075C>G; p.Pro2359Ala) are classified as Benign or Likely

Benign by Franklin. However, the conditions for which these variants are reported as benign are often not specified. Interestingly, the c.7075C>G, p.Pro2359Ala variant was reported to be involved in Primary Ovarian Insufficiency [436]. Functional analysis revealed that the p.Pro2359Ala variant led to a significant decrease in transcription activity of NOTCH2 [436].

In NCOR1, out of the three remaining variants, 2 are classified as VUS and were never reported in ClinVar (c.667G>A;p.Val223Met and c.3360G>C,p.Glu1120Asp), while the last variant (c.4218A>C; p.Leu1406Phe) was reported as benign in ClinVar for an unspecified condition. Of note, the c.3360G>C,p.Glu1120Asp variant was found in one patient with digital papillary adenocarcinoma, although the authors do not report any evidence of its involvement in the pathogenesis [437]. We further investigate the potential disease mechanisms of the involvement of these two genes in infertility in Section 6.8.

The pairs that are found to be statistically significantly enriched when not taking into account the inheritance mode of the genes involved variants in genes of the DNAH family, which encode for dynein axonemal heavy chain proteins, and are important for cell motility [438]. For example, we find the pair DNAH11;DNAH8 as being the most enriched gene pair in the top 50. However, although DNAH8 is known to be important for sperm flagella and is highly expressed in sperm cells [439], DNAH11 is almost exclusively expressed in the brain. The genes in this gene pair are therefore not likely to act together to cause infertility. The pair DNAH1;DNHD1 might however be relevant, as both genes are expressed in spermatids and have been shown to be associated with morphological abnormalities of the sperm flagellum. Furthermore, variant combinations in this gene pair appear to have high pathogenicity scores (Appendix F). Although both genes are described to have an AR mode of inheritance, and that the variant combinations found in this gene pair involve, for the most part, single heterozygous variants, they could be relevant under a digenic inheritance model and deserve further investigation.

6.7.4 Enriched gene pairs in patients subgroups

As mentioned in Section 3.1.2, the patients of the ESTAND cohort can be divided in 3 main subgroups according to their phenotype. In this part, we therefore assess whether we could find significantly more frequent gene pairs in these three subgroups as compared to the controls. Because the size of the patient group is smaller compared to the controls, it it easier to have statistical significance, and the patients are more likely to have similar oligogenic causes because they present with a more similar phenotype.

Patients with azoospermia

There are 183 patients in the cohort diagnosed with Non-Obstructive Azoospermia (NOA), 80 of which were sequenced as part of the GEMINI study and 103 sequenced at the FIMM. We here assess the enriched gene pairs in the 103 FIMM samples compared to the 322 controls, using the same procedure as described above.

Unfiltered				Inheritance mode filtered			
Gene pair	$p-value$	Num. patients carriers	Num. Control carriers	Gene pair	$p-value$	Num. patients carriers	Num. Control carriers
Top 10							
BRCA2;SLX4	0.0486	3	1	SMARCA4;FGFR1	0.0494	3	1
Top 20							
BRCA2;FANCA	0.0466	3	1	SMARCA4;FGFR1	0.043	3	1
BRCA2;SLX4	0.0466	3	1				
ATP2B4;DCC	0.0488	3	1				
Top 50							
CREBBP;AKAP9	0.0108	3	0	EP300;NOTCH2	0.0484	3	1
DNAI1;DNAH8	0.044	3	1	TGFBR3;ERCC6	0.0496	2	0

Table 6.4: Significantly frequent gene pairs in the top 10, top 20 and top 50 ranking of the patients with azoospermia.

We do not observe many significant findings in the group of patients with NOA, with only 3 gene pairs with a $p-value$ less than 0.05 when applying the IM filter and 5 significant gene pairs in the unfiltered results (Table 6.4). Furthermore, the majority of these gene pairs are still found to be carried by at least one control individual, and we therefore do not investigate these pairs further. The details on the variant combinations found in these gene pairs can be found in Appendix F.

Patients with oligozoospermia

There are 181 patients with Oligozoospermia (OZ), 3 were sequenced using the GEMINI platform and were thus excluded from the analysis, while 178 were sequenced as part of the FIMM group and are thus here compared to the 322 infertile men.

Unfiltered				Inheritance mode filtered			
Gene pair	$p-value$	Num. patients carriers	Num. Control carriers	Gene pair	$p-value$	Num. patients carriers	Num. Control carriers
Top 10							
SBF1;BMPR1B	0.0436	3	0	SLIT3;DCC	0.0488	3	0
CHD7;RECQL4	0.0448	3	0				
SETX;BLM	0.0484	3	0				
Top 20							
ATM;RECQL4	0.0048	5	0	CHD7;GLI2	0.0174	4	0
CHD7;GLI2	0.041	3	0	SLIT3;DCC	0.0416	3	0
SBF1;BMPR1B	0.0456	3	0	CHD7;FGFR1	0.0452	3	0
RECQL4;MCM9	0.0498	3	0	TP63;PPP1R12A	0.0468	3	0
Top 50							
DNAH11;DNAH8	0.0058	6	0	CHD7;GLI2	0.0196	4	0
ATM;RECQL4	0.01	7	2	NOTCH2;JAG1	0.0392	3	0
CHD7;GLI2	0.0134	4	0	TP63;PPP1R12A	0.0408	3	0
NOTCH2;ATM	0.0142	7	2	NOTCH2;GLI2	0.0424	3	0
CFAP61;DNAH1	0.0146	4	0	SLIT3;DCC	0.0424	3	0
TOP2A;RECQL4	0.0148	4	0	NOTCH2;MCM2	0.0426	3	0
RECQL4;MCM9	0.0208	4	0	TP63;NOTCH3	0.043	3	0
ROS1;USP42	0.0358	3	0	APP;FN1	0.0446	3	0
CSMD1;SH3TC2	0.0404	3	0	CHD7;FGFR1	0.0446	3	0
CHD7;MCM8	0.0422	3	0	NOTCH2;DCHS1	0.0454	3	0
NOTCH2;EIF2B2	0.0432	3	0	CHD7;TENM4	0.048	3	0
NOTCH2;TTC21A	0.0442	3	0				
SBF1;BMPR1B	0.0446	3	0				
USP42;PRDM9	0.0446	3	0				
ATM;ME1	0.0448	3	0				
EPHB1;SETX	0.0452	3	0				
KLC2;DNAH1	0.0456	3	0				
EPHB1;DCC	0.0466	3	0				
DCC;POTEJ	0.0476	3	0				
SPTBN2;DCC	0.0478	3	0				
BRCA2;EXO1	0.0284	0	8				
MLH3;RECQL4	0.0414	0	7				

Table 6.5: Significantly frequent gene pairs in the top 10, top 20 and top 50 ranking of the patients with oligozoospermia.

We find more significant gene pairs in the group of patients with OZ. Of note, 4 of the samples carrying variant combinations in the CHD7;GLI2 pair appear to be in this group, and several pairs that are found to be significantly enriched involve novel genes that are not present in the original panel (Table 6.5 blue genes), including a gene pair that only involves novel genes (the APP;FN1 pair). These findings highlight the fact that Hop and our analysis protocol allows for potential new discoveries from WES and is not too strongly biased towards the seed genes. The full results on the variant combinations are shown in Appendix F.

Patients with cryptorchidism

There are in total 152 patients with cryptorchidism. 4 were sequenced with the GEMINI samples and 148 belong to the FIMM group. We here focus our analysis on the 148 FIMM patients and look for gene pair enrichment in these patients as compared to the 322 controls. The cryptorchidism group appears to have fewer significant gene pairs than the OZ group, when considering the genes' inheritance mode. Furthermore, except for the NCOR1;NOTCH2 combination which was already discussed above, all these gene pairs involve the GNAS gene.

Unfiltered				Inheritance mode filtered			
Gene pair	$p-value$	Num. patients carriers	Num. Control carriers	Gene pair	$p-value$	Num. patients carriers	Num. Control carriers
Top 10							
ROS1;DNAH1	0.033	3	0	GNAS;DNAH1	0.0334	3	0
Top 20							
SLX4;RECQL4	0.0286	3	0	GNAS;DNAH1	0.0264	3	0
TTC21A;DNAH1	0.0292	3	0	HERC2;GNAS	0.0278	3	0
POLR3A;SETX	0.03	3	0				
WDR11;SEMA5A	0.0304	3	0				
ROS1;DNAH1	0.034	3	0				
BRCA2;EXO1	0.0446	0	8				
Top 50							
GNAS;DNAH1	0.0028	5	0	GNAS;DNAH1	0.0314	3	0
DNAH11;DNAH8	0.0032	5	0	NCOR1;NOTCH2	0.0326	3	0
HCN4;ATP2B4	0.0276	3	0	GNAS;MPG	0.0328	3	0
NOTCH2;VWA3A	0.0288	3	0	HERC2;GNAS	0.0336	3	0
FAT4;CELSR3	0.03	3	0				
DNAH1;DNHD1	0.0304	3	0				
POLR3A;SETX	0.031	3	0				
RNF213;ZFHX2	0.031	3	0				
CDC14A;DNAH1	0.0312	3	0				
ROS1;DNAH1	0.0312	3	0				
CENPE;DNAH1	0.0322	3	0				
TTC21A;DNAH1	0.0326	3	0				
WDR11;SEMA5A	0.0334	3	0				
CFAP65;TMEM247	0.0338	3	0				
SBF1;SEMA5A	0.0342	3	0				
SOHLH1;MCM8	0.0354	3	0				
DNAH10;DNAH6	0.037	4	1				
BRCA2;EXO1	0.048	0	8				

Table 6.6: Significantly frequent gene pairs in the top 10, top 20 and top 50 ranking of the patients with cryptorchidism.

Overall, we observe in the analysis of the different subgroups that we find less enriched gene pairs in the NOA patients as compared to the other subgroups of patients. This is likely to be due to the size of the sample set, since there are only 103 samples with NOA, while there are more patients in the other subgroups. However, this might also be due to the fact that the NOA phenotype is more extreme and thus more likely to be caused by monogenic variants than oligogenic ones.

6.8 Analysis of the mechanisms underlying enriched gene pairs

In order to analyze the potential biological mechanisms underlying the digenicity of certain pairs, we further investigate them using ARBOCK (see section 3.5.6 and [374]), to generate graph explanations for each of the identified gene pair.

We investigate the two pairs that were found to be enriched in the whole group of patients: CHD7;GLI2 and NCOR1;NOTCH2. For each pair, we use the ARBOCK model with and without phenotype to predict the pathogenicity of the pair and generate explanations for the predictions, which are then visualised as small networks involving the genes, GO terms, HPO terms and other entities of the graph which are likely to be relevant to the disease mechanisms as predicted by the tool (see Section 3.5.6 and [374]).

We observe that the graph explanations of the CHD7;GLI2 mechanisms involve other genes known to be associated to male infertility such as SOX2 and FGFR1 (Figure 6.22).

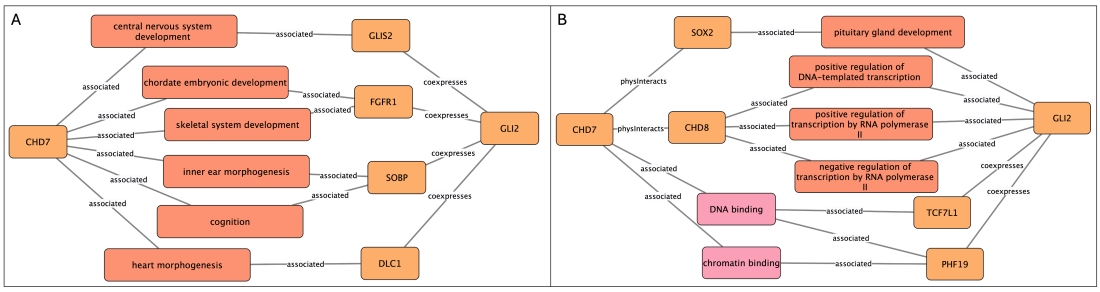


Figure 6.22: Explanation graphs generated by ARBOCK for the CHD7;GLI2 gene pair. Explanation from subgraph (A) is associated with a prediction score of 0.97 and explanation from subgraph B is associated with a prediction score of 0.84. Nodes are colored according to the color code of BOCK (see Figure 3.10), genes are in orange, Biological Processes in red and Molecular functions in pink.

It appears that the CHD7 and GLI2 genes are involved in several developmental processes, either directly or through physical interaction and coexpression with other genes. In particular, one of the graph explanations shows that through interaction with SOX2, CHD7 and GLI2 are both involved in pituitary gland development (Figure 6.22B). This mechanism was investigated in functional experiments, which show that through physical interaction, CHD7 and SOX2 regulate the expression of different genes in neural stem cells [440]. These genes are part of the Notch and Shh pathways including GLI2 [440]. Furthermore, this analysis shows that SOX2 and CHD7 bind the genes JAG1, GLI2 GLI3 and MYCN and activate the expression of JAG1, GLI2 and GLI3 [440]. Since the pituitary gland is an important regulator of hormones crucial for spermatogenesis (in particular LH and FSH), alterations in these two genes are likely to have an effect on male fertility [441].

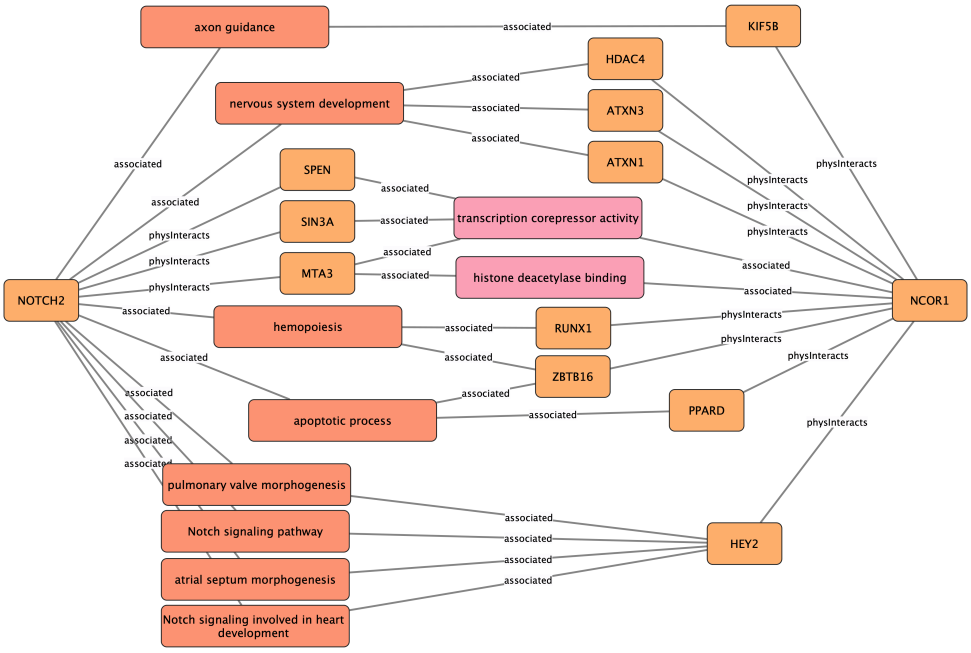


Figure 6.23: Explanation graph generated by ARBOCK for the NOTCH2;NCOR1 gene pair. Nodes are colored according to the color code of BOCK (see Figure 3.10), Genes are in orange, Biological Processes in red and Molecular functions in pink.

The explanation graphs generated by ARBOCK for the NOTCH2;NCOR1 pair appear to be more complex, and we here show one such graph which highlights different processes involving both genes (Figure 6.23).

The genes seem to be involved in axon guidance and nervous system development, though these associations are indirect in the case of NCOR1. It appears from the ARBOCK explanations that NCOR1 is linked to Notch signalling, involving the NOTCH2 gene, through physical interaction with HEY2. The Notch signalling pathway is important for controlling cell proliferation, differentiation and death in many organs, which explains why so many biological processes appear to be found by ARBOCK. In particular, Notch signalling also plays a role in the development of both female and male reproductive systems, and NOTCH2 has been found to be highly expressed in different parts of the male reproductive tract [442]. Notch signalling has been suggested to also be important in sperm maturation [443]. Furthermore, NCOR1 has been shown to play an important role in developmental processes via regulation of the Notch signalling pathway [444]. Finally, a mouse model with mutant NCOR1 was shown to have decreased levels of thyroid hormone and thryoid stimulating hormones, which are important for male reproduction [445, 446]. Although the mechanism needs to be clarified, these different lines of evidence point towards the potential involvement of these genes in infertility.

6.9 Analyzing shared genes within gene pairs in infertile men

In this section, we turn to investigating whether specific genes appear to be enriched in the prioritized gene pairs in patients as compared to controls. We apply the same methodology as for the gene pair, except that this time, the entities counted are the genes found within the top prioritized combinations.

With this analysis we aim to find genes that appear to be often shared between patients but not within controls. These genes might either partially cause infertility, with their effects influenced by alternative genes, or act as modifiers, interacting with various other genes. This analysis also gives us more statistical power as it is more likely to find patients with common genes in the cohort than patients with the exact same gene pairs. Furthermore, this allows to focus on some of the novel exome genes (i.e. genes which are found by Hop which are not part of the gene panel) and assess their relevance.

6.9.1 Enrichment analysis of genes within gene pairs in infertile men compared to controls

First, we investigate the genes that are found to be significantly enriched within gene pairs among all infertility patients as compared to controls. We observe that, when taking into account the IM of the genes, we find much more novel genes (i.e. genes that were not in the original panel) to be enriched than when looking at the unfiltered results (Table 6.7). This can be caused by the fact that since we use the same gene panel for prioritization in both patients and controls and that the top combinations appear to be slightly biased towards the seed genes, due to the DS, theses genes are more likely to be found in the top prioritized combinations of both controls and patients, and therefore to not be significantly enriched in patients. It is important to note that since we performed a single gene enrichment analysis at the beginning of this chapter, the results presented here represent an enrichment of the genes within the prioritized pairs, and are not caused by an enrichment in the number of variants within these genes.

Unfiltered				Inheritance mode filtered			
Top 10							
Gene	$p-value$	Num. patients carriers	Num. Control carriers	Gene	$p-value$	Num. patients carriers	Num. Control carriers
ASZ1	0.005	10	0	PRKCB	0.006	9	0
CDC14A	0.022	12	2	NOTCH2	0.007	57	24
SBF1	0.024	24	8	VCX3B	0.017	11	1
CCDC146	0.039	13	3	PTPRT	0.018	7	0
ROS1	0.043	42	19	TUB	0.02	10	1
DHX37	0.044	15	4	ACTRT1	0.021	7	0

RNF213	0.044	29	12	GLI2	0.022	28	10
				ATM	0.027	6	0
				PIAS2	0.028	9	1
				LRP5	0.032	6	0
				ESR1	0.037	13	3

Top 20

Gene	$p-value$	Num. patients carriers	Num. Control carriers	Gene	$p-value$	Num. patients carriers	Num. Control carriers
ASZ1	0.007	12	1	PRKCB	0.002	12	0
SPO11	0.008	8	0	PTPRT	0.002	13	0
NOTCH2	0.012	52	21	NOTCH2	0.004	62	25
VCX3B	0.013	8	0	VCX3B	0.008	12	1
FEZF1	0.014	8	0	DNMT3A	0.012	8	0
BCORL1	0.015	8	0	GLI2	0.013	30	10
PPM1J	0.021	18	5	SOX30	0.013	8	0
CDC14A	0.027	14	3	BRD2	0.015	8	0
CCDC146	0.028	14	3	FAT4	0.016	11	1
ROS1	0.033	47	21	ESR1	0.018	15	3
RNF17	0.036	9	1	DHX37	0.023	21	6
CSMD1	0.041	39	17	ABL2	0.026	10	1
AXDND1	0.048	10	2	AP2A2	0.027	9	1
				BPTF	0.028	25	9
				TUB	0.028	12	2
				BCORL1	0.034	9	1
				TYRO3	0.034	11	2
				MAGI1	0.043	13	3
				TNRC18	0.043	8	1
				AAK1	0.047	8	1
				CASZ1	0.048	11	2
				FBXO5	0.049	10	2

Top 50

Gene	$p-value$	Num. patients carriers	Num. Control carriers	Gene	$p-value$	Num. patients carriers	Num. Control carriers
NPHP3	0.004	9	0	PTPRT	0.001	27	4
ASZ1	0.005	13	1	ELL	0.002	12	0
FEZF1	0.005	10	0	VCX3B	0.003	16	2
NOTCH2	0.005	60	23	NOTCH2	0.005	66	27
PABPC1	0.005	20	4	SOX30	0.005	13	1
PRKCB	0.005	10	0	DHX37	0.006	26	7
SORL1	0.005	19	4	PABPC1	0.006	29	9
GLI2	0.008	27	8	FAF1	0.008	14	2
VCX3B	0.008	12	1	NCOR1	0.009	27	8

AAK1	0.01	8	0	PRKCB	0.009	12	1
CAND2	0.01	8	0	TNRC18	0.009	25	7
TUB	0.01	11	1	COL7A1	0.011	40	14
FLNB	0.013	16	3	MLLT10	0.011	12	1
MAGI1	0.013	13	2	TEAD4	0.013	11	1
POTEJ	0.02	27	8	SPEG	0.015	21	6
CCDC146	0.028	14	3	CASZ1	0.016	16	3
MYO6	0.031	13	3	AAK1	0.017	10	1
BCORL1	0.032	9	1	DGKZ	0.019	15	3
FBXO5	0.034	11	2	FAT4	0.02	26	9
CSMD1	0.035	47	22	SH3GL1	0.02	10	1
RNF17	0.036	9	1	IQGAP2	0.021	23	8
SPO11	0.037	11	2	YEATS2	0.021	10	1
SBF1	0.039	42	19	DNAJC13	0.022	21	6
ITGB4	0.045	10	2	BCORL1	0.023	10	1
AXDND1	0.048	10	2	PES1	0.024	11	1
NCOR1	0.049	23	8	RIMS1	0.024	12	2
				ADGRL3	0.027	14	3
				CHD6	0.029	14	3
				FHIP2B	0.029	10	1
				MGA	0.031	18	5
				XAB2	0.032	21	7
				ESR1	0.039	15	4
				GORASP1	0.041	21	7
				MAGI1	0.041	19	6
				LAMA5	0.043	19	6
				CIT	0.049	18	6
				FBXO5	0.049	10	2

Table 6.7: Significantly frequent genes within gene pairs in the top 10, top 20 and top 50 ranking of the infertile patients compared to normozoospermic males.

Since many genes appear to be significant and further analysing the individual variants in these genes would be too time consuming, we focus on investigating the enriched Gene Ontology (GO) terms associated to these genes are present and on assessing whether some of the novel genes found by this analysis can be relevant for male infertility and could be therefore considered as novel candidates for the disease.

For the biological process enrichment analysis, we use PANTHER [447], a platform that performs enrichment analysis of certain types of annotations for a gene list of interest. We compare all unique genes found in the top 10, top 20 and top 50 combined to the list of all human genes to test for over-representation of biological processes, molecular function and cellular component from the GO [378].

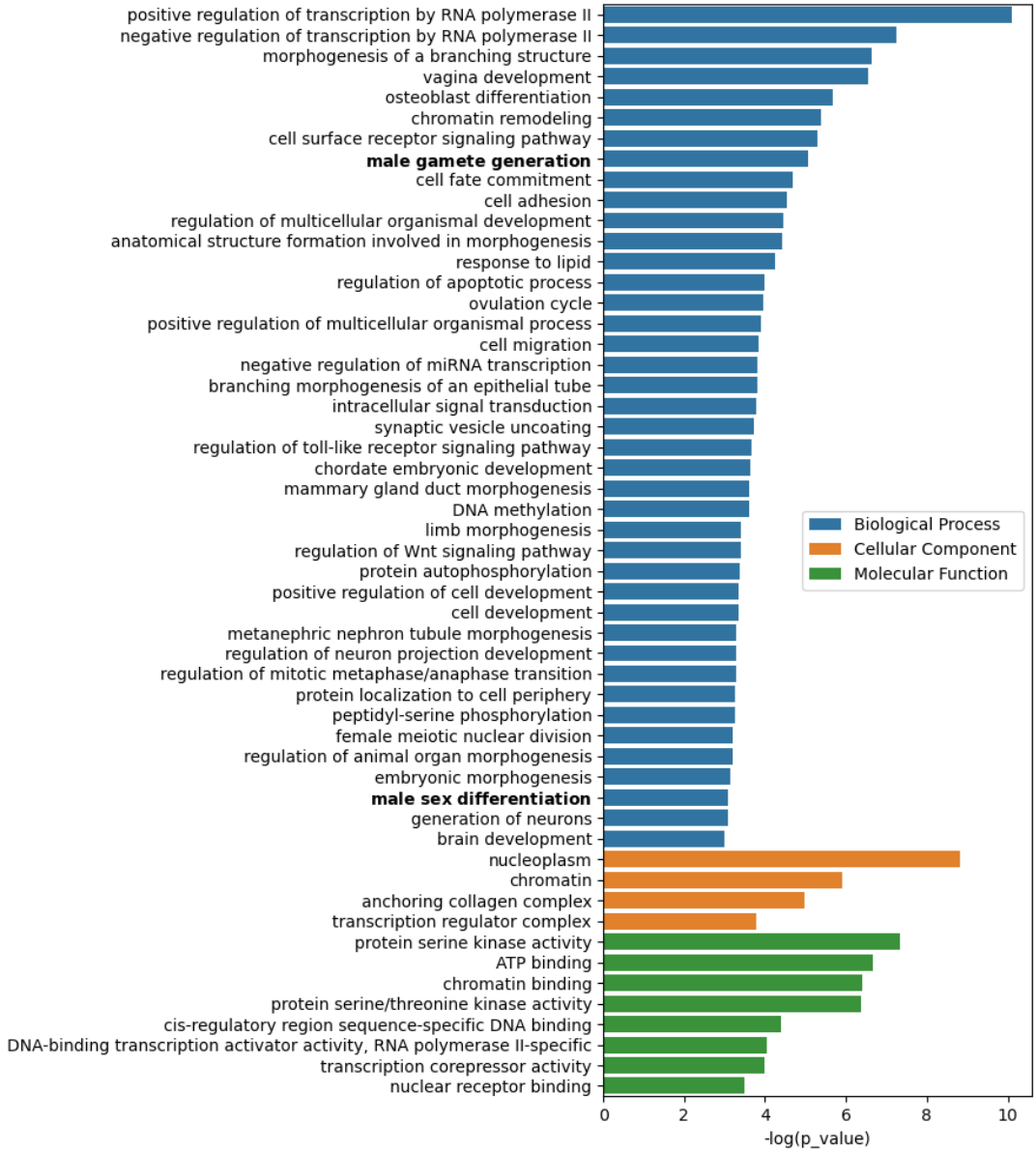


Figure 6.24: GO enrichment analysis for all the genes that are found significantly enriched in all patients compared to controls when considering the inheritance mode of the genes.

We find a large number of enriched GO terms and therefore only visualise the parent term when there are several terms from the same branch in the ontology (Figure 6.24). Among the relevant biological processes, we notice that the male sex differentiation (GO:0046661) and male gamete generation (GO:0048232) terms are significantly over-represented in our gene set. We also observe several processes linked to the female reproductive system such as vagina development, ovulation cycle and mammary gland duct morphogenesis. This can

indicate that there are more links between the male reproductive genetics and female reproductive genetics, and that the genes identified can thus be relevant for both reproductive systems. However, this could also be a bias due to the use of POI genes as seeds for the algorithm. We also notice several terms linked to developmental processes.

Moreover, we analyse a few of the novel genes found to be enriched (in blue in Table 6.7), by searching in different databases including OMIM, Franklin and the human protein atlas [448] to investigate whether these genes might be relevant for male infertility.

PRKCB, the most enriched gene within gene pairs when considering the top 10 and top 20 combinations with IM filtering (Table 6.7), encodes for the protein kinase C beta and is part of the protein kinase C family. It is important for mitochondrial energy homeostasis and autophagy and was shown to be differentially expressed during spermatogenesis [449]. Furthermore it is found to be relatively highly expressed in the testis and more specifically in sertoli cells, which are important for spermatogenesis [450].

The TUB gene encodes a protein that is a member of the Tubby family of bipartite transcription factors. In human, it has been associated to increased body weight and obesity in several studies [451–453], which can potentially explain the presence of this gene in our cohort of patients, since a large proportion of the patients are also affected with obesity [227]. Interestingly, mouse models involving the knock-out of other genes of the Tubby family, such as TULP2, have been found to be infertile [454, 455]. However, contrary to the TULP2 gene, TUB is not found to be highly expressed in testes and its relevance to the male infertility phenotype is therefore not likely.

SOX30, is a member of the SOX family of transcription factors which are involved in the regulation of embryonic development and determination of cell fate. It is found to be particularly highly expressed in the testes according to both Franklin and the Human Protein Atlas. Experiments in mouse models have shown that it has an important role in meiosis and that homozygous Sox30 deletion led to spermatogenic arrest, leading to an infertile phenotype in mice [456]. Epigenetic modifications in this gene have also been associated to NOA patients [457]. Although there are currently no clear genetic associations between SOX30 variants and male infertility in humans, a variant in this gene was reported as part of an oligogenic combination in a DSD patient in OLIDA (OLI575 [434]) and we found one ClinVar report (Accession ID VCV002573119.1) for a VUS found in a patient affected with azoospermia. These findings therefore seem to support the potential involvement of SOX30 in male infertility and highlight the need to perform further investigation on this gene. In particular, the gene pairs involving this gene should be investigated further in order to better understand the potential oligogenic mechanism.

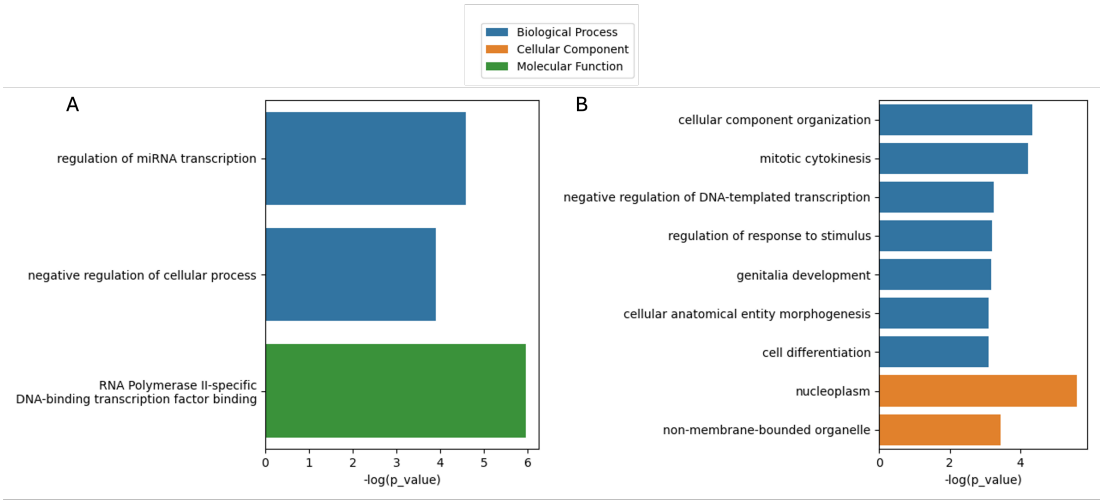


Figure 6.25: Gene ontology enrichement analysis for the set of genes found to be statistically siginificantly enriched in the NOA patient (A) and OZ patient (B) groups.

6.9.2 Enriched genes in patients subgroups

In this final section, we investigate the enriched genes within gene pairs in the different patients subgroups in order to investigate whether we found particular genes in these different subtypes of the disease. The tables showing enriched genes within patients subgroups are available in Appendix G. We here highlight the GO enrichment analysis for these genes as well as discuss a few of the genes which are found to be enriched specifically in each group.

Patients with azoospermia

In the NOA patient group, there were a total of 18 distinct genes that were significantly enriched, when taking into account the inheritance mode. The large majority of these genes were novel, with only NOTCH2 and DHX37 that were part of the original gene panel (Appendix G). The gene ontology enrichment analysis of this gene set revealed the enrichment of three GO terms, which are linked to transcription and regulation of cellular processes (Figure6.25A).

Among the novel genes found to be enriched in NOA patients, SMARCA4 is highly expressed in testes and is not found to be enriched in any other patient subgroups. The protein encoded by this gene is part of a large complex which is required for transcriptional activation of genes normally repressed by chromatin. SMARCA4 has been associated to Coffin-Siris syndrome and cancers [458], and a link with azoospermia is therefore not evident.

Patients with Oligozoospermia

We find a total of 37 distinct enriched genes in the OZ patient group. Here again, many are novel with only 6 genes from the original panel. The GO enrichment analysis shows enrichment of specific relevant biological processes such as genitalia development (GO:0048806) (Figure 6.25B).

Noticeably, both PIAS2 and FAF1 which are found to be enriched in OZ patients are highly expressed in testes. In particular, PIAS2 was reported to be potentially important for spermatogenesis as it is highly expressed in spermatids and spermatocytes [459]. It is also found to be highly connected in the network of signalling proteins in human spermatozoa [460].

Patients with cryptorchidism

The cryptorchidism patient group presents with 39 enriched genes (Appendix G). Only one gene ontology term is found to be over-represented in this set of genes, which is why this group is not shown in Figure 6.25. This term is the cellular component “anchoring collagen complex”, which is linked to two genes in the cryptorchidism enriched gene set.

Of note, BMPR1A is found to be significantly enriched in the cryptorchidism group, while it is not enriched in the other subgroups. This gene is in the 638 candidate genes panel, as it was found to be associated to POI, and is also known to cause Juvenine Polyposis Syndrome, which has been associated to cryptorchidism in some cases [461, 462]. BMPR1A is involved in the anti-müllerian hormone signalling pathway, which is important for proper testicular descent, indicating that mutations in this gene might be relevant for cryptorchidism [463].

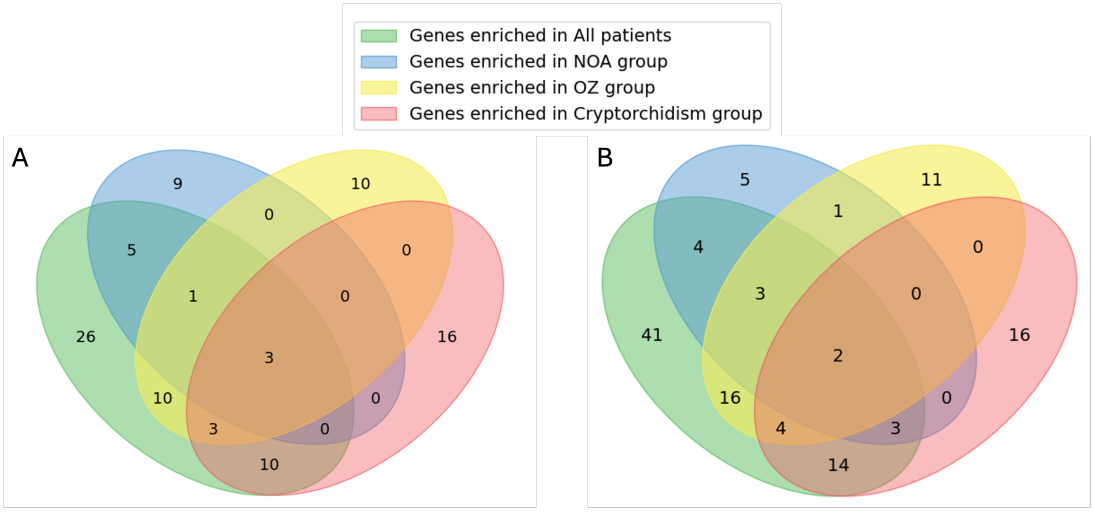


Figure 6.26: Number of shared enriched genes when performing gene enrichment analysis within gene pairs for the different patient groups without any filtering of the results (A), after Inheritance Mode (IM) filtering (B).

We show in Figure 6.26 the number of genes that are found to be significantly enriched in the different patient groups (including all patients). It further indicates that the NOA patient group shows the smallest number of enriched genes, and that it is more distinct than the other patient groups. Indeed when considering all results without IM filtering, 9 genes are only enriched in this NOA group and 9 genes shared with at least one other group, while for the other groups, the number of shared enriched genes is larger than the number of genes unique to the group. In particular, we notice that no enriched genes are found in common between the three subgroups of patients (NOA, OZ and Cryptorchidism in Figure 6.26), suggesting potentially distinct genetic origins.

Although the majority of the findings presented here deserve more careful analysis and attention, they highlight that the genes found by our approach include relevant candidates for male infertility. In particular, we find many novel genes that were not present in the original panel, which could be assessed further.

6.10 Conclusion

In this chapter, we demonstrate the usefulness of our prioritization method, Hop, in analysing a real patient dataset. In doing so, we further describe the different steps that should be taken to apply this predictor to such a dataset, defining for the first time a protocol for the analysis of oligogenic variant combinations in WES data of a case/control cohort. This protocol is here applied to a cohort of patients affected with male infertility, but is designed to be fully reproducible and can therefore be applied to case/control cohorts for other diseases.

The first part of this protocol highlights the need to properly filter VCF files and analyse enrichment of single genes in order to detect and remove potential sequencing or variant calling errors which would produce spurious associations. This analysis should be performed for all genetic association studies, but is especially important in the case of oligogenic analysis because of the number of combinations that are being tested. Any false positive association can therefore increase when assessing oligogenic inheritance models. As shown in our analysis, these false positive associations can arise due to a variety of factors, and it is difficult to find criteria that will remove all likely sequencing errors. However, it is important to note that any such filtering of VCF can also lead to the removal of true disease-causing variants, as we have seen in our analysis with the removal of 2 diagnostic oligogenic variants and 11 monogenic diagnostic variants. Different strategies to improve the quality of the variants included in the dataset could be explored in the future. In particular, it could be interesting to try to do joint calling on the variants for the full cohort of samples (patients and controls) from the BAM files, to generate a gVCF file.

Furthermore, we recommend gene enrichment analysis to detect potential differences between case and controls cohorts. In particular, this can also generate interesting insights into the genetics of the disease. While it was not the case here, since we have a relatively small cohort for a very heterogeneous disease, gene enrichment analysis can give interesting monogenic findings in some cases [61]. In any case, this analysis is important in order to understand why certain gene pairs or genes appear to be enriched in the downstream analyses performed on the oligogenic predictions. Indeed, it ensures that the genes that are found to be enriched in the gene pairs are found to be enriched due to their potential oligogenic involvement and not to their over-representation in the patient group.

Analysing general statistics on the combinations prioritized by Hop in patients and controls allowed to observe a few characteristics of Hop predictions. First we noticed that the majority of the combinations prioritized form networks, and involve both genes that are in the original panel as well as novel exome genes, indicating that the tool can make new discoveries. This analysis also confirmed the small bias towards the disease-relevance score noticed in Chapter 5.

We explored three different analyses paths to investigate how to use the predictions of Hop to generate new knowledge on the genetics of male infertility. First we assessed the ability of Hop to discover the previously manually identified causative variants, which validated the performance of the tool on real patient data. Second we investigated how these predictions can be used to identify over-represented gene pairs in patients vs controls as well as in patient subgroups and third, we looked at the potential of the tool to discover new genes that are likely to be relevant for the disease.

In this analysis we also investigated the use of an Inheritance Mode (IM) filter to remove single heterozygous variants in genes that are known or predicted to be autosomal recessive and thus tolerant to these variants. Applying this filter led to a significant reduction in the number of variants to predict and analyse (see Section 6.3) but also to a significant improvement in the ranks of the diagnosed combinations by the predictor (Section 6.6). Moreover, it seemed to lead to a more diverse set of genes in the prioritized combinations (i.e. more genes that were not in the original panel were part of the top combinations). This filter seems relevant for clinical interpretation, since it is based on “rules” that are applied by clinicians when interpreting genomic variants in the context of disease. However, this filter might not be relevant to uncover novel digenic mechanisms, and might thus bias predictions and observations towards cases of dual molecular diagnoses rather than true digenic. It is important to note that most of the results which were described in further details in this chapter were thus obtained with the application of this filter, which is a limitation.

It is important to keep in mind that all variants and genes that we highlight as potentially relevant for male infertility are for now hypothetical, and would need to be further analysed and tested through functional experiments to validate their involvement.

Finally, with the exception of the previously diagnosed cases, we have analysed the results from a cohort perspective, trying to find patterns in the predictions to identify oligogenic signatures that are shared among patients. As shown by the small number of statistically significant gene pairs, this approach would require large cohorts to identify meaningful signatures. However, the Hop tool is still relevant with smaller cohorts as it can predict the potentially causative variant combinations for each patient independently and these can be further analyzed manually. We show that such analysis is likely to be successful for some patients, by finding again the previously diagnosed combinations. Manually analysing the top prioritized network would be time consuming, but might lead to the diagnosis of more patients.

Chapter 7

Discussion and conclusions

In this final chapter, we summarize the research presented in this thesis. First we go through the main scientific contributions, highlighting the novel database created, the computational predictive methods developed based on this database and the analysis pipeline to detect novel oligogenic signatures in whole cohorts. We therefore demonstrate that with our work, oligogenic analysis of whole cohorts is now possible using whole-exome sequencing data directly, without having to filter predictions with a gene panel, which is an important advancement in the field. In a second part, we discuss the limitations of the current work, following three main ideas: first we discuss how knowledge biases, which are found in the OLIDA database, can affect predictive methods and thus influence any result generated using these predictive tools. Secondly, we discuss the main limitations behind ranking approaches and their integration in routine clinical practice, specifically the issue that they do not provide a definite number of instances to examine. Finally, we discuss current issues with validating genetic findings, whether they are from small pedigree studies such as the clinical cases found in OLIDA or from larger cohort studies using prioritization methods such as the male infertility analysis done in this work. Building from these limitations, we introduce ideas for future directions to improve our work. First by discussing how to maintain the created database while reducing its bias, second by discussing how the disease relevance score of our prioritization approach could be improved using deep-learning techniques and third by investigating approaches to bring our analysis to clinicians without compromising data privacy.

Finally, we provide concluding remarks on this thesis work, by highlighting the importance of data quality and state our vision of what is needed for the future of oligogenic disease research.

7.1 Scientific contributions

With the increase in data collection brought forward by next-generation sequencing technologies, the field of medical genetics have undergone important transformations. With WES technologies becoming integrated into routine clinical diagnostic pipelines, there has been an increasing need for computational methods to analyze such data, interpret the genetic variants in relation to disease, and assist clinicians in providing molecular diagnostics. As our understanding of the genetic etiology of diseases has shifted from the "one gene - one disease" paradigm to the idea that more complex patterns of inheritance can cause disease through epistatic interactions, methods tailored towards the identification of such inheritance patterns are becoming essential. Machine learning and predictive algorithms show great promise in this context, but require sufficient amounts of high quality data in order to generate accurate predictions.

In this thesis, we therefore set to improve the state-of-the-art predictors for such task, in order to move towards the detection of oligogenic variant combinations in whole-exome sequencing data. In doing so we achieved three main scientific contributions, which we summarize here. First, we generated a comprehensive repository of data for collecting information on variant combinations reported as causative of diseases in the scientific literature. Second, we used this dataset to train a predictor to prioritize variant combinations in WES data, based on how likely they are to cause the patient's disease. And finally, based on this predictor, we develop an analysis protocol that describes the key steps that should be undertaken when applying this tool to whole cohorts, in order to identify new genetic causes to disease.

7.1.1 Developing a comprehensive repository of data on oligogenic diseases

As the amounts of data linking genetic variants to diseases is continuously increasing in the scientific literature, the collection of this data in an organized format, that can be easily accessed and searched by researchers with an interest in the field is becoming important. Furthermore, assessing the quality of datasets and genetic associations is important in order to be able to have confidence in the collected genetic associations, as well as create high-quality datasets for developing machine learning predictors. While numerous databases have been developed to collect information on variants associated to disease in monogenic cases, the only database aiming at collecting such data for oligogenic diseases presented with limitations. DIDA, published in 2016, was not a comprehensive repository, as it only contained data from curated articles that fulfilled a number of particular criteria, which were not clearly defined.

With the creation of OLIDA, our first main scientific contribution, we achieve two main goals: first, we create the first comprehensive repository collecting all variant combinations which have been reported in the scientific literature as causative of an oligogenic disease; and second, we define a first set of standards for associating an oligogenic variant combination to disease.

In achieving the first goal, we show that the number of diseases associated with oligogenic combinations throughout the years have not increased as fast as the number of genetic variants reported, which thus pinpoints data gaps and biases. OLIDA itself has already grown a lot in the past three years (the number of variant combinations in the database has almost doubled), highlighting the importance of maintaining this resource. The number of publications reporting on oligogenic cases is continuously increasing, and such reports might not have their place in other databases of variant-disease associations such as ClinVar.

In fulfilling our second objective, we observe that the evidence that supports the involvement of a variant combination in a disease is different from the standards that exist for associating monogenic variants to human disease. These considerations are therefore necessary to be put forward to the scientific community, which we did by publishing a commentary article on the subject. Furthermore, we show that applying these standards to current reports demonstrate a very low quality of the reports

In summary, we have created an important repository for oligogenic disease research, which can help biomedical researchers looking for other cases of oligogenic inheritance in their diseases, bioinformaticians in developing tools trained on this dataset, biologists in looking for gene pairs to test for interaction, and clinicians in improving reports on oligogenic combinations by investigating different characteristics of variants (as opposed to monogenic involvement considerations).

7.1.2 Advancing computational methods for the pathogenicity prediction of variant combinations

In the second part of this research, we developed two novel computational methods for assessing the pathogenicity of variant combinations. The first one is an extension of the previously developed Variant Combination Pathogenicity Predictor, but includes an improved training set, new features and a simpler model structure. These three changes lead to predictions of higher accuracy and a faster computational time. Interestingly, the number of positive instances included in the training set did not increase that much compared to the first version of VarCoPP, going from 200 instances from DIDA in VarCoPP to 301 instances from OLIDA in VarCoPP2.0, yet the prediction quality of VarCoPP2.0 is much better. This could be due to the fact that the instances selected were this time of much better quality, and would need to be tested by, for instance, evaluating the performance of a predictor trained on all instances of OLIDA regardless of their confidence score.

In a second step, we integrate the VarCoPP2.0 pathogenicity predictions together with a disease-relevance score for gene pairs in order to develop the High-throughput oligogenic prioritizer (Hop). This novel method is the first prioritization tool directly targeting variant combinations. The only other tool aiming to achieve something similar, OligoPVP [194], relies on monogenic variant pathogenicity scores, which, as we show in Chapter 5, is not sufficient for ranking variant combinations. The advancement brought forward by Hop is thus significant, since, as we show with our comparison of different tools for the task of prioritization, such task can not be performed using current approaches to prioritization.

In addition to developing this novel method, we experiment with using different seed types for the disease-relevance scoring, showing that the combination of information from different sources (in this case gene panels and HPO terms) lead to better prioritization. This is important to know in the case of oligogenic diseases, since the HPO terms have been associated to genes mostly based on monogenic associations through OMIM, and are thus biased towards these genes [71, 79]. We also observe in the top prioritized combinations a small bias towards the disease-relevance score, which is important to keep in mind and understand when analyzing new results generated from Hop (see Chapter 6). This also indicates that Hop can potentially be improved, by using a more sophisticated integration of the disease-relevance and pathogenicity scores. These potential improvements are discussed in more details in Section 7.3.2.

Besides the initial testing in >50,000 synthetic exomes generated by spiking OLIDA combinations in neutral exomes, we also validated the tool in real patient data from the ESTAND cohort. Here again, the predictor showed good performance, ranking all the manually identified combinations in the top 50 when applying a filter based on the inheritance mode, and ranking 5/8 combinations in the top 50 without applying this filter. We furthermore show that the predictor also works well at finding monogenic diagnostic variants in the top combinations, and therefore might also be relevant for the detection of modifiers to these monogenic causes. It is of course important to note that the number of samples in this validation set is extremely small, but WES data of patients for which oligogenic causes have been identified is at the moment extremely rare.

7.1.3 An analysis protocol for detecting oligogenic signatures in large cohorts

Our third and last scientific contribution is the establishment of a first analysis protocol for the detection of oligogenic signatures in case/control cohorts with WES data. This protocol was here defined for a cohort of patients with male infertility recruited at the University of Tartu, but aims to be easily reproducible to be applied in other cohorts of patients with potential oligogenic causes.

The protocol highlights key steps that should be taken to limit false positive genetic associations that are likely to arise when analysing WES data for oligogenic combinations. Indeed, when searching for digenic combinations, the number of instances that we test for association with the disease grows quadratically. It is therefore crucial to control for false-positive variants due to sequencing or calling errors. In this case, we tried to control this number by applying sequentially different filters on the VCF files based on quality criteria, allele frequency of the variants in the cohort or size of the indels. These filters removed several potentially false positive variants but also a certain number of true positive variants which were manually identified as causative. It will therefore be important to refine these variant filtering criteria for each cohort independently in order to avoid removing too many real variant calls. This filtering step is followed by a monogenic gene enrichment analysis, which is important to control for the potential enrichment of false positive variants, but also can lead to novel monogenic findings in the cohort.

Our protocol then highlights three important analyses to discover new findings in case/control cohorts studies. The first is a single patient analysis, where the top prioritized combinations of each patient are visualized as a small network. The second is a gene pair enrichment analysis, which has the potential to identify gene pairs that are significantly enriched in patients compared to controls, and thus potentially represent relevant oligogenic signatures for the disease. The third analysis is a gene within gene pair enrichment analysis, which has the potential to identify novel candidate genes for the disease, but also specific modifier genes that could be part of oligogenic signatures.

The three paths highlighted as analyses need to be further explored to fully understand the significance of the generated findings. However we believe the steps described are important to be taken into account by the biomedical community interested in performing such an analysis. While we are only at an exploratory stage, we were able to show that: (i) the single patient analysis manages to retrieve the manually identified variant combinations, and could thus be used as a way to diagnose patients with oligogenic combinations; (ii) the gene pair enrichment analysis identified pairs that appear to be relevant for the disease, including variants that were already reported in OLIDA as modifiers for similar phenotypes; (iii) the novel genes identified are often good candidates for an involvement in male infertility, as they are expressed in relevant tissues and pathways.

This first analysis also shows the importance of having cohorts of sufficient size to have enough statistical power to make novel discoveries. We also highlighted this point in our commentary article [256], insisting on the need to have large cohorts and even larger control datasets. Another possibility to improve statistical power would be to be able to query large databases of population variation for the frequency of particular variant combinations or gene pairs. Overall, we hope that this protocol will help researchers identify the specific concerns that should be taken into account when aiming to discover new oligogenic combinations in

cohort studies. This is complementary to our published commentary on standards for the reporting of variant combinations causative of disease. Reproducing this analysis in cohorts for other diseases, ideally with larger number of participants, will be essential to better understand the strength and limitations of the proposed methodology.

7.2 Limitations of the current work

In this section, we highlight several of the key limitations linked to our scientific contributions. These open the way to improvements and future directions discussed in the next section. The first limitation is linked to biases in biomedical data that is translated in our database, but also our predictive tools and subsequent predictions. The second limitation is linked with the usage of tools by clinicians, and how to easily interpret the predictions of Hop. The third limitation is linked to the validation of oligogenic, and genetic findings in general, which is essential to produce data of good quality but also in the development of precision medicine strategies.

7.2.1 Knowledge biases can have strong impact on predictions

As highlighted in Section 4.6.3, the content of OLIDA appears to be slightly biased, as the database includes many variant combinations identified in the same diseases and involving the same genes. Furthermore, the curation protocol that we established is likely to contribute to a bias in the number of combinations with higher confidence scores which are found in gene pairs that have direct interactions or are known to be involved in the same pathway. These biases are important to acknowledge and recognize as they have direct effects on predictions generated by tools that are trained on these datasets. In order to further characterize them, it would be important to assess whether VarCoPP2.0 and Hop perform equally well at predicting variant combinations associated with different disease types. In addition to the biases due to OLIDA dataset, Hop might also suffer from biases present in BOCK. Indeed, we have noticed that oligogenic genes (i.e. genes that are known to be associated to an oligogenic disease in OLIDA), and disease-genes in general, are more connected in BOCK than the remaining human genes (Appendix C). This is mostly due to the integration of curated experimental and clinical evidence networks such as the GO and HPO networks, which are heavily biased towards over-studied genes. Moreover, PPI networks, while dense, have also shown to exhibit biases [408, 464]. Highly connected genes are in turn likely to affect the RWR algorithm and thus create false positive associations, as we have seen in the analysis of the ESTAND cohort where several genes appeared to be highly ranked in both the patient and control groups (Chapter 6).

As we are defining protocols for the discovery of novel genetic findings using our tools, these biases need to be recognized as they might push the research even more strongly into a circularity issue, where predictors trained on biased data generate new biased associations which are then further added to the training dataset. In particular, the OLIDA curation criteria include assessing the pathogenicity of variant combinations using *in silico* predictors of pathogenicity, which have, so far, mostly been trained on this dataset.

Furthermore, the evaluation criteria for the genetic evidence defined in our curation protocol currently puts a lot of weight on evidence obtained from pedigree data. Indeed, according to our criteria, the only way to obtain strong genetic evidence is through this type of analysis, which can be difficult to perform as it implies the sequencing and analysis of family members, who are not always available. As we are witnessing a shift towards cohort studies and statistical analyses, this criteria will need to be re-evaluated in the future, in order to allow for strong genetic evidence to be obtained from such analyses. However, since most databases of population variation do not currently allow for obtaining the frequency of combinations of variants, we do believe that such strong evidence can not yet be obtained, as assessing the frequency of variants in combination is essential to assess the significance of oligogenic findings.

7.2.2 From ranking to instance selection

One of the main limitation of Hop is that although it seems to provide in the majority of cases a good ranking of the combinations, placing the known diagnostic combination in the top 50, there is no way of assessing how many combinations are relevant to be considered for further examination by a clinician or in a statistical approach. In Chapter 6, we have investigated using different tops (Top 10 , Top 20 and Top 50), but this only increases the number of possibilities of gene and gene pairs sets to test during further analyses. Furthermore, in this cohort, we could have an estimate of the number of combinations to consider based on the ranks of the manually diagnosed combinations, but this is not always possible as diagnosed cases are not always available.

This is also a known limitation in monogenic variant prioritization tools. To our knowledge, the only methods that have been developed to address this issue are the LIkelihood Ratio Interpretation of Clinical AbnormaLities (LIRICAL) framework [185] and the Phenotype-driven Likelihood Ratio analysis approach (PheLR) [465]. Both approaches use a likelihood ratio paradigm to score and rank variants based on their potential to cause the disease associated with a particular set of HPO terms, which allows to compute a posterior probability of the patient having the specific disease and the variant being causative of the disease. Both approaches rely on the HPO database to assess the probability that a patient has a specific disease depending on whether she presents with a particular symptom. While this is possible for monogenic diseases, for which such information is extensive, digenic

and oligogenic diseases are much less characterized, and can actually present with more diverse phenotypic manifestations. Developing such an approach for oligogenic prioritization is therefore not currently possible, but defining other criteria for selecting instances, such as confidence scores in the ranking or predictions will be important in the future.

Although both Pathogenicity Score (PS) and Disease-relevance Score (DS) are made available by Hop, each score independently is not interpretable and do not provide a systematic way to select instances. This limitation is especially important for the use of the tool in clinics where interpretability is essential, and users need an objective way to select relevant variants, especially if it is for decision making [466, 467].

7.2.3 The problem of validation of genetic findings

As highlighted in Chapter 6, all findings generated by our Hop method and analysis protocol are only supported by statistical evidence and need to be further assessed to be considered as truly disease-causing. While methods such as ARBOCK [374] can assist with identifying relevant pathways and interactions that provide additional support for the involvement of the genes in the disease, some of the graphical explanations generated remain too dense, or only involve high-level processes that do not provide additional knowledge. Furthermore, these explanations need to be further validated by domain experts. With databases such as ClinVar, Franklin and OLIDA, finding other articles reporting on the pathogenicity of identified variants is now more accessible. Nevertheless, since many genetic diseases are rare, the probability to find another case with the same variant remains low. Furthermore, the evidence brought forward by other case reports is once again mostly statistical, as in the majority of cases, the variants presented in these reports have not been functionally validated.

It is also important to keep in mind the possible circularity argument already put forward in Section 7.2.1. As we are now developing tools to detect oligogenic variants based on OLIDA, we need to be careful to not validate the found combinations by looking at whether there were other reports in that same database. For the variants that we have presented in Chapter 6, which were found by Hop and also in OLIDA combinations, we ensured that these variants were not present in our training set. Even so, the genes they were found in might have been included in training and can thus bias these findings. Alternative detection methods (trained on other datasets or unsupervised methods) or alternative independent datasets are thus necessary to ensure proper validation of the findings using such databases.

As we have seen, the study of gene disease associations is entering a new era of cohort studies with a decrease in functional testing of variants (Section 4.6.3). We have stressed the importance of functional analysis in our commentary article on standards for the reporting of oligogenic disease variants [256], but in practice, we are aware that doing such analysis is not straightforward.

Recently, a novel method to functionally assess the joint effect of two variants together has been developed [390]. This functional assay uses a new type of base editing (a genome editing system derived from CRISPR-Cas9 systems, which allow to introduce precise mutations in DNA without causing double strand breaks) to directly introduce two mutations at once. The authors show that a significant proportion of digenic and oligogenic combinations in OLIDA could now be tested using this novel system. The method is actually tested on two OLIDA combinations, validating the synergistic effects of these combinations, and demonstrating its relevance for functional testing of disease variants. This method therefore provides great promise to validate genetic findings using *in vivo* digenic assays. However, such functional experiments are always costly and require experts knowledge which are not always available in clinical research groups. More collaborative efforts are therefore necessary to obtain functional validation of genetic findings.

7.3 Future directions

Starting from the aforementioned limitations, we here highlight several ideas for overcoming these limitations, improving our work and moving forward from the proposed research.

7.3.1 Maintaining and updating OLIDA: using text-mining approaches and reducing bias

In addition to the biases discussed above (Section 7.2.1), one of the main issue linked to OLIDA is the maintenance of the database in terms of time and resources. Indeed, in our curation protocol, each article has to be read by two curators independently, before comparing the responses and annotating the data. While the annotation pipeline is semi-automatic, several issues remained due to the different ways a variant can be defined and referred to. The full curation pipeline is therefore time consuming and with the huge increase in the number of publications reporting on oligogenic variants (see Chapter 4), will not be sustainable in the future.

We envision two strategies for maintaining this database up-to-date, which could be applied independently or collectively, and which should also help in reducing the bias present in the database.

The first idea is to use text mining approaches, which could help identify directly the sections of the article where the relevant information is found but also eventually directly extract such information. These strategies would thus lead to huge gains in time for the curator, as manually searching information in articles was one of the main limiting factor in the curation process. Furthermore, this approach could also reduce biases which are caused by human curation. Indeed, the extraction of information about the involvement of genes in disease through

specific pathways or experiments can sometimes be overlooked by human curation but would be detected by automatic processes. A first step has already been made in this direction, with the creation of the Detection of Unique Variant Ensembles in Literature (DUVEL) corpus [391]. This dataset is the first corpus focusing on relations between two genes and two variants (i.e. digenic variant combinations), providing a potential new benchmark for the detection of such biomedical relations [391]. The growing interest in the field of text mining to assist curation of the biomedical literature [468, 469] is promising for the use of these approaches in the maintenance of OLIDA.

The second idea is to transform OLIDA into a community maintained resource, where entries are submitted by users, removing the need for curation of the literature. A step forward in this direction has already been made, with the design of a submission interface on the OLIDA website. Furthermore, reviews of the evidence associated with particular combinations could also be attributed by the community, as it is done in ClinVar and the Genomics England PanelApp [80, 83]. Confidence scores based on curation could thus be complemented with confidence scores from the biomedical community, which would help improve the level of evidence associated with each entity, but also reduce bias as these scores would be the results of aggregating the knowledge of experts from different fields. There has been an increase in the development of such community databases, allowing for users' data input and reviews (e.g. Varsome, ClinVar and Franklin which are used in this work). Crowdsourcing, i.e. using the voluntary help of large communities to solve problems, has been described to be promising for such purpose in biomedical research [470]. With growing interest in oligogenic diseases underlined by the increasing number of publications, this approach also has considerable potential.

7.3.2 Knowledge graph embeddings for disease-relevance scoring

One of the limitations of Hop that was put forward in Chapter 5 is that its performance appears to plateau at identifying the correct oligogenic variant combination in the top 20 in 70% of the tested exomes. One of the potential reason for this is that the knowledge Hop bases itself onto has been mostly generated by monogenic associations (e.g. HPO to gene links are based on monogenic associations) and that the approach can thus not always identify the digenic links.

However, this might also be caused by the fact that our Hop approach so far makes use of a relatively simple algorithm for knowledge propagation, while novel graph representation techniques are being increasingly used to generate new knowledge. In particular, Knowledge Graph Embedding (KGE) methods, have shown great promise in identifying novel gene-disease associations [471], suggesting that such an approach could also generate new knowledge for digenic genes. The master thesis of Inas Bosch [472] investigated the use of such models for the prediction of the pathogenicity of gene pairs using BOCK, and showed very encouraging results. Using such an approach to compute an improved disease-relevance

score is thus likely to generate better results. In particular, embedding approaches have been shown to perform very well in comparing phenotypes [473]. The application of these methods to monogenic variant or gene prioritization already yielded promising results, outperforming approaches based on random walks [471, 474, 475].

Other ideas to improve rankings of Hop include using approaches of learning-to-rank (see Section 3.4.2) to directly train the predictor to rank the variants instead of using a combination of knowledge propagation and supervised classification, or using a more sophisticated integration of the two scores to generate the final ranking, by training another predictor on these scores using meta-learning approaches.

7.3.3 Bringing oligogenic pipelines to clinicians

While the development of the ORVAL platform was an important step forward in providing access to the first generation of oligogenic predictive tools such as VarCoPP and the digenic effect predictor (see Section 1.3.5), new approaches will need to be developed moving forward. Indeed, as the study of human genetics is moving into the whole exome sequencing era, data privacy concerns are becoming even more crucial as such type of data allows for re-identification of the patients. Furthermore, when performing whole cohort analyses, pipelines need to be available so that they can be easily scaled to run in parallel on a large number of samples.

The "oligopipe" package[1] that has been developed in the group will therefore need to be extended to allow for the analysis of multisample VCFs and the prediction of such data with Hop. This expansion could also include the gene pair enrichment and gene enrichment analyses described in Chapter 6, therefore providing a comprehensive package for oligogenic analyses based on our protocol. For now, this package relies on an annotation database, which is hosted on the Vlaams Supercomputer Centrum (VSC) and contains values of all features that are necessary for prediction using our tools. To remain relevant, this database needs to be updated to include the most recent version of the features, following updates from predictors such as CADD [476], and databases such as Ensembl for gene information [477]. While a pipeline for the construction of the database has been developed, changes in the structure of the datasets will require adaptations of this pipeline, and for each new version of the annotation database, the predictors will need to be retrained. These technical developments have been so far carried out by the combined work of Alexandre Renaux, Emma Verkinderen and Nassim Versbraegen, and were therefore not discussed in this thesis as they are out of scope. However, we believe it is important to discuss them here as they are essential to consider when moving forward with these tools. As our predictive methods are improving in accuracy and gain interest for larger scale analysis, taking into account the technical data structures the predictors rely on is important.

1. https://pypi.org/project/oligopipe/

Another possibility to bring these analyses pipelines to clinicians and other researchers is to use cloud infrastructures. A first private cloud version of the ORVAL platform is currently being investigated in collaboration with the 101 Genomes Foundation. First attempts at running Hop in the google cloud, in order to analyze data from the Epi25 project which is hosted there, have also been done in additional work not reported in this manuscript. Genomics data is being increasingly stored on cloud infrastructure, as it allows for secure sharing of these sensitive datasets [478]. Developing pipelines that can be run in the cloud and thus be brought to the data is therefore likely to be beneficial to the biomedical community [479].

In addition to making the tools available, proper documentation and guidelines as to how the results should be interpreted and analysed will be extremely important to achieve this transition to the clinics. This involves precise description of the biases that are inherent to the tools and that need to be taken into account when interpreting the results.

The availability of these computational pipelines and interpretation protocol, will enable the discovery of oligogenic signatures for new diseases by applying these methods in large cohorts for which data is already available such as the Epi25 project for epilepsy, or the Deciphering Developmental Disorders (DDD) dataset for developmental disorders. This is essential for validating the relevance of the methods described in this thesis.

Finally, it will be important to consider how to adapt these resources and methods to detect oligogenic causes in WGS data, as this is becoming the new standard for genetic analyses. This sequencing not only covers much larger regions of the genome, but is also more accurate for the detection of structural variants and copy-number variants. One of the current solutions simply consists in only predicting variants obtained from WGS which are located in the exons. This has been done when assessing the performance of Hop on the synthetic exomes generated from individuals of the UK10K project (Chapter 5), which have been sequenced using WGS. As more resources emerge for variants located in non-exonic regions, our tools could be extended to include these type of variants, as well as integrate CNVs and structural variants. A first step towards this has already been made by integrating CNVs in OLIDA. However, this still represents a too small training set to create predictors that can directly aim at the detection of variants in WGS data.

7.4 Concluding remarks

In this last section, we provide concluding remarks on two major themes that come out of this thesis: the importance of data quality in bioinformatics and our vision of the future of oligogenic disease research.

7.4.1 The importance of data quality in variant pathogenicity prediction

Throughout the work presented in this thesis we have encountered several times issues linked to data quality, which is why we wanted to provide concluding remarks on this topic, which appears to be common to different fields of bioinformatics.

As the volume of genetic data accumulates, ensuring the quality of this data becomes critical for accurate variant interpretation. Our evaluation of the evidence put forward to associate variant combinations to disease showed a relative decrease in the proportion of reports showing sufficient evidence. Despite the publication of guidelines by American College of Medical Genetics [78], conflicting interpretations of variants remain common [480, 481]. Such discrepancies and inconsistencies in variant classification can have large effects across different levels of human genetics research, as such associations are used to generate disease-gene associations, which in turn influence phenotype-gene associations and so on. Furthermore, correct interpretation is essential for proper integration of these practices in a precision medicine perspective [482].

This issue underlines the necessity for more stringent reporting standards and quality control protocols to ensure that only robust, reproducible findings are published. The many articles describing expansions to the ACMG guidelines [483–486], including our article on guidelines for the reporting of oligogenic variants [256], provide a step forward in this direction, although with a multiplication of guidelines, they become harder to follow.

In the current era of bioinformatics, data quality is essential at every step of the analysis pipeline. This includes data collection, curation, annotation and usage in predictive methods. The initial step of assessing the quality of sequencing data is critical because any errors or biases introduced at this stage can propagate through subsequent analyses. With large scale sequencing projects becoming routine, it is important to correct for the many biases that can arise, including but not limited to batch effects, ethnicity biases and sequencing kits artefacts. This requires meticulous data analysis and quality control measures, which can oftentimes be overlooked.

Next, the collected data must be carefully curated and annotated. Proper phenotype annotation, for instance, is essential for the accurate interpretation of genetic variants in relation to disease (see Section 5.5.4). Accurate annotation and curation of datasets has a direct impact on the performance, reproducibility and generalizability of predictive methods. With higher-quality datasets, predictive models will have higher performance but also will produce more reproducible results, as they will not learn from biases in the datasets. If the training data is of poor quality, the models will likely produce inaccurate predictions, which can lead to errors in subsequent results and clinical decisions. This is the case for many variant effect predictors for example, which were shown to have very different performance across datasets, in part due to data circularity [487].

In conclusion, the importance of data quality in variant interpretation, and bioinformatics in general, can not be understated. As genomic data continues to accumulate, and is predicted to generate one of the largest “big data” problems [488], the need for rigorous quality control measures at every step of the research process becomes essential. Ensuring the accuracy and reliability of genetic data from collection to interpretation is essential for advancing our understanding of the genetic basis of diseases and translating these findings into meaningful clinical applications. It is thus by prioritizing data quality that we can improve the reliability of gene-disease associations, improve predictive models, and eventually, contribute to the advancement of precision medicine [482].

7.4.2 The future of oligogenic disease research

Overall, with this thesis work, we bring oligogenic disease research forward in different ways. First we provide new resources to assist clinicians in validating or making new discoveries, whether it is in whole cohorts or in single patients. With OLIDA, we provide the means for finding other cases where the same variants, variant combinations, genes or gene pairs were found to be causative of the disease. While this can rarely provide complete validation of the involvement of a specific variant combination in a disease phenotype, it can help build up the evidence for the pathogenic mechanism. With Hop and VarCoPP2.0, we provide the means to assess the pathogenicity of variant combinations using *in silico* tools, as well as to identify other potential variant combinations in a patient’s exome. Finally, with our protocol, we define the key steps necessary to analyse case/control cohorts with these tools and highlight the important considerations that should be taken into account when analysing the results.

As the set of resources for analysing oligogenic inheritance models is growing, it is important to keep in mind the current limitations of the approaches, which we summarize here. First, our tools remain limited to the predictions of SNVs and small indels, although CNVs are also increasingly reported. Although we have started collecting oligogenic combinations involving this type of variants in OLIDA, ML tools to predict the pathogenicity of such variants are still scarce [489, 490]. In addition to not covering CNVs, our tools are also limited to WES data. Here again, this is mostly due to limitations in available training data, which has been so far mostly generated using panels or WES technologies. As more data becomes available, we envision our oligogenic tools to be relatively easily translatable to cover both CNVs and variants in other regions of the genome, which could be made possible by the addition of new features to characterize these types of variants. Finally, it will be important to consider how to move further in the oligogenic spectrum than the digenic predictions presented here. One possibility is simply to combine these digenic combinations into small networks, as it was done in Chapter 6. It will also be important to characterize whether this approach can provide interesting insights into the mechanisms underlying more complex polygenic diseases.

We have already highlighted a few ideas on how to improve the performance of the tools, which should also gain from an increase in the amount of training data available. The main challenge that lies ahead is thus to bring these prediction and analysis pipelines to clinicians. This process not only involves the development of user-friendly interfaces and tools to generate the predictions, but mostly requires the design of careful guidelines and recommendations on how to interpret the generated results, in light of the potential biases inherent to such predictive methods.

To end on a positive note, we believe that the work presented here provides strong foundations to help geneticists, bioinformaticians and clinical researchers better understand the associations between variant combinations and human diseases. Moving forward, the datasets, methods, guidelines and pipelines defined in our work will enable novel discoveries in the field, which will in turn help refine these methods.

Appendix A

OLIDA decision trees

A.1 Functional score

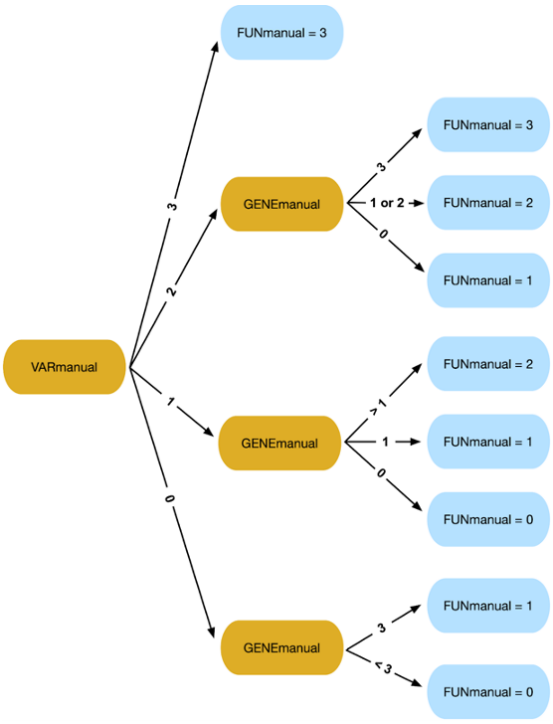


Figure A.1: Decision tree for defining the FUNmanual and FUNmeta scores of a variant combination (where, in the second scenario, the VARmanual, GENEmanual and FUNmanual scores are replaced by VARmeta, GENEmeta and FUNmeta scores accordingly). The orange nodes represent an evidence score and the edges the value of this evidence score, while blue nodes represent decision nodes where the aggregated functional score is defined. The variant combination evidence is considered as the starting point (root) of the decision tree and, along with the information from the gene combination evidence, is used to define the final aggregated functional score. As the evidence at the variant level is very important in asserting whether a specific variant combination is functionally responsible for the observed phenotype, its absence (VARmanual = 0) can only be compensated by a very strong synergistic evidence at the gene level (GENEmanual = 3).

A.2 Final score

Figure A.2: Decision tree for defining the FINALmanual and FINAlmeta score of a variant combination (where, in the second scenario, all scores, apart from the FAMmanual, are replaced by their corresponding metascores). The orange nodes represent an evidence score and the edges the value of this evidence score, while blue nodes represent decision nodes where the final score is defined. The familial genetic score is considered as the starting point (root) of the decision tree and, along with the statistical score and functional score (whose value is defined by the decision tree in Figure A.1) is used to define the final score. Based on this decision tree, there is always a combination of genetic and functional evidence that is required to obtain a FINAL manual/meta score of 1, and the familial evidence (FAMmanual) is considered to be primarily important for this decision. A more lenient scenario is found in the case where the FAMmanual score is strong (=3), meaning that there is strong evidence of the pathogenicity of the studied variant combination based on the segregation of the involved variants in a large pedigree, as a variant combination can still be accepted with a final score of 1, if there is at least some minimum functional evidence at the gene level (GENEmeta/manual = 1).

Appendix B

OLIDA schema

Following here are the entity-relation schemas of OLIDA. Each table represents an entity with the name in purple and below its attributes, on the left the type of data and on the right their name. Entities are connected with different types of relations, with different arrows on each side representing the cardinality of their side of the relationship :

- two parallel straight lines represents exactly one,
- a round and a parallel line represents zero or one,
- a round followed by a fork represents zero or more,
- a fork with one straight line represents at least one or more, i.e. many,

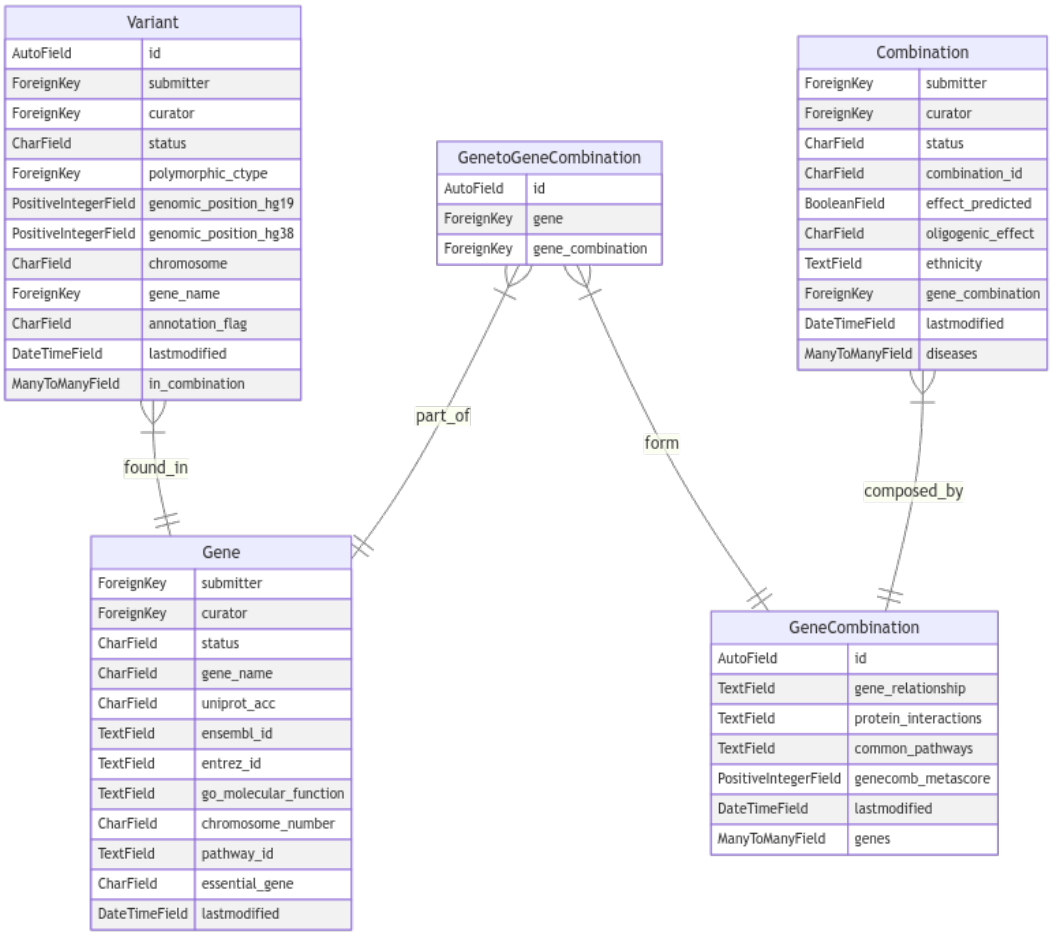


Figure B.1: Entity relation diagram for the `Gene` and related entities of OLIDA.

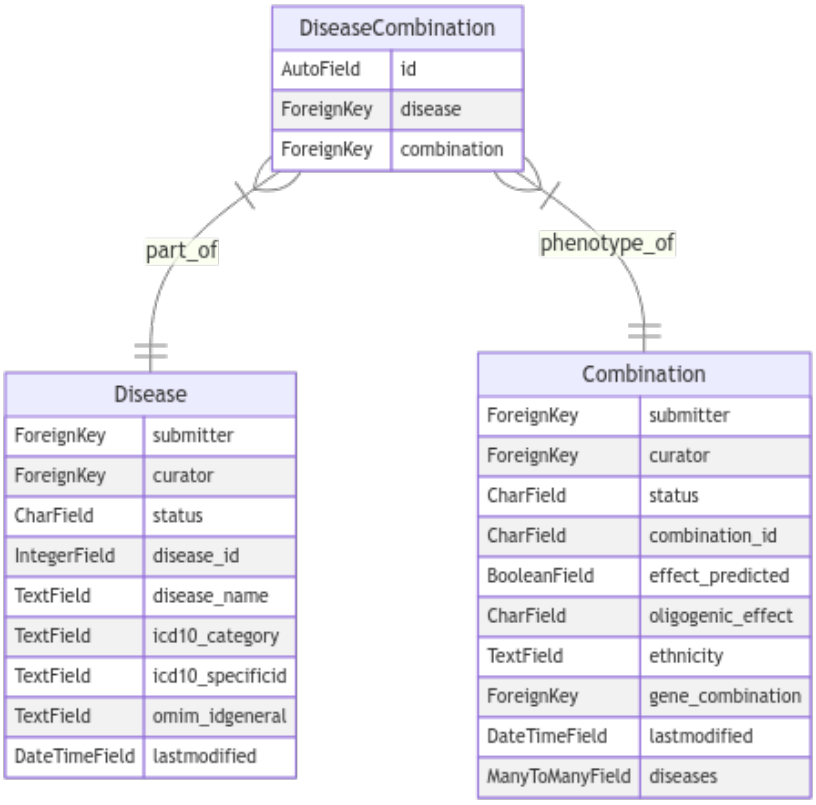


Figure B.2: Entity relation diagram for the `Disease` and related entities of the OLIDA database.

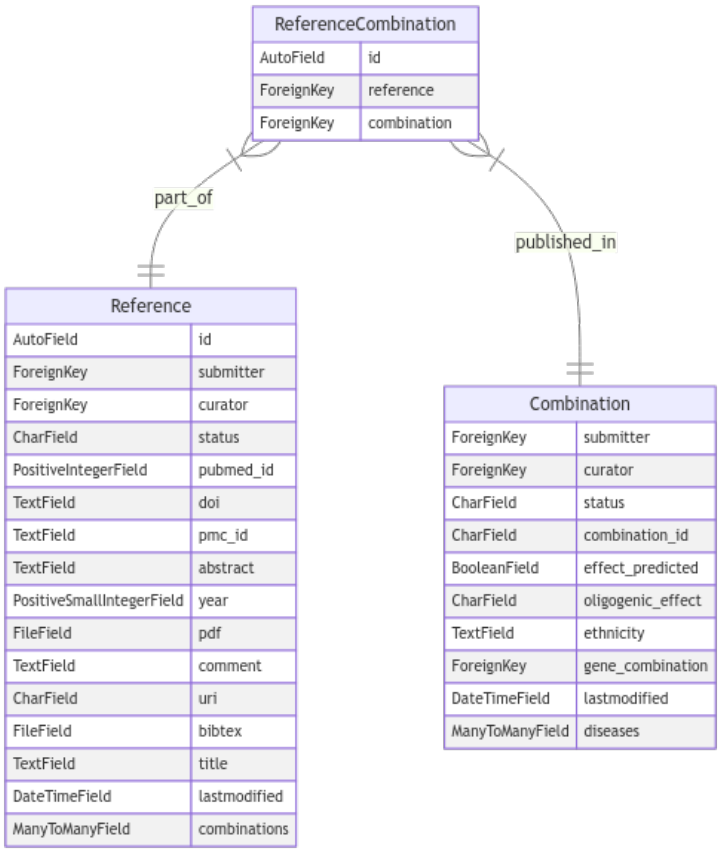


Figure B.3: Entity relation diagram for the `Reference` and related entities of the OLIDA database.

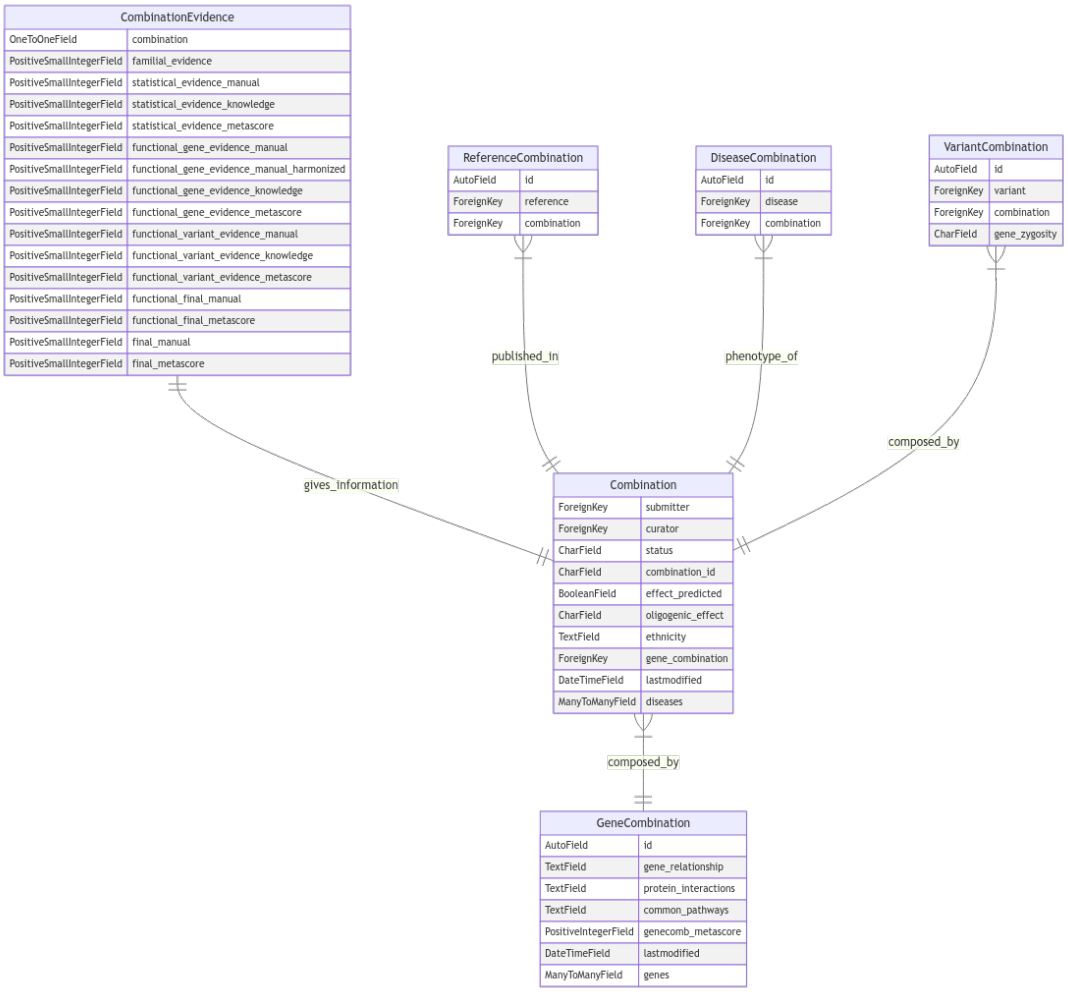


Figure B.4: Entity relation diagram for the `Combination` and related entities of the OLIDA database.

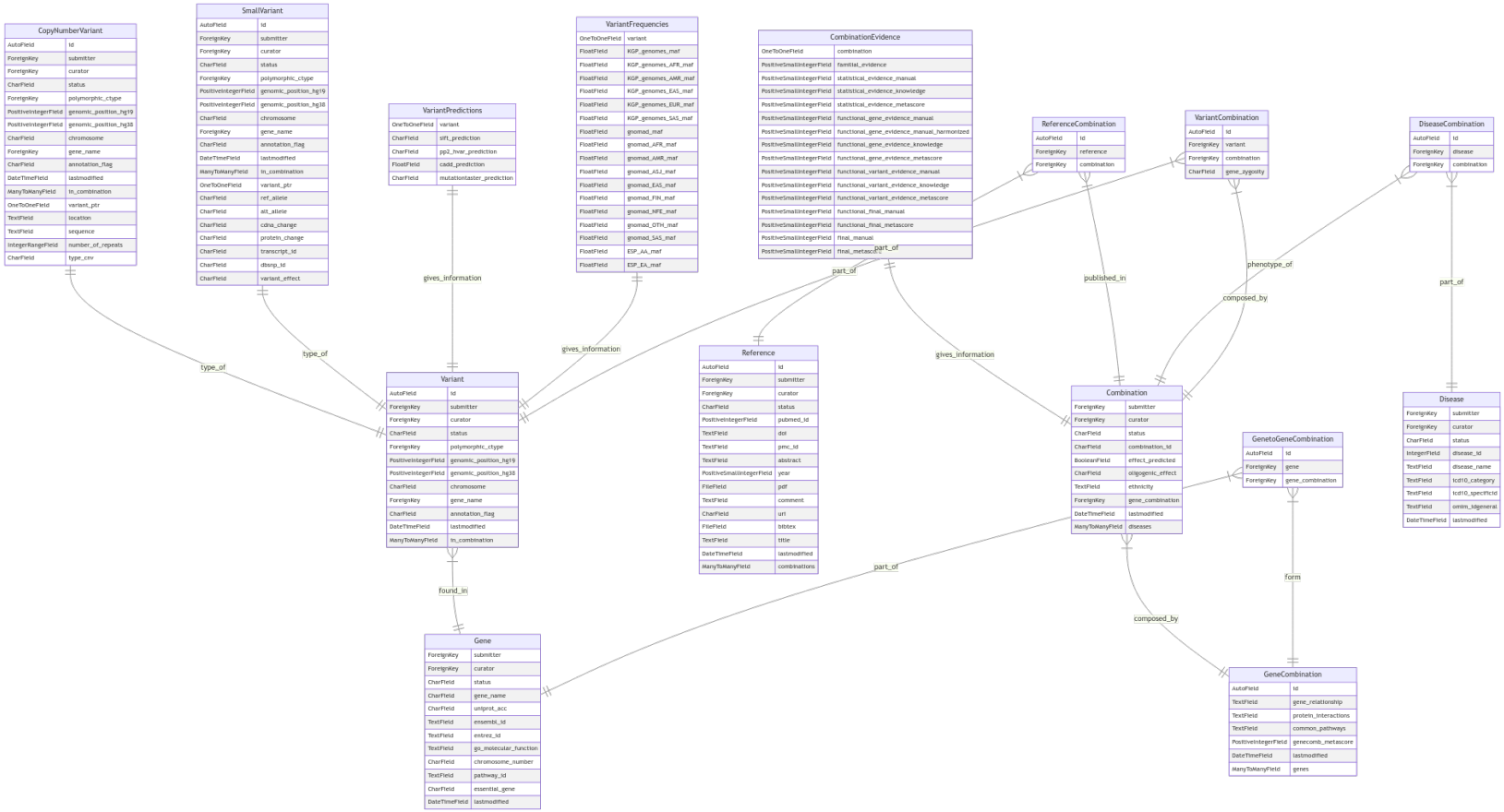


Figure B.5: Entity relation diagram for the complete OLIDA database.

Appendix C

Analysis of gene set degrees in BOCK and knowledge induced biases

In order to better quantify the potential biases implied by using a knowledge-graph approach, we here assess whether certain sets of genes are more likely to be found by the random walk algorithm. To do this, we investigate **the degree of a node**, which assesses the number of nodes a node is connected to. We computed this value for each node in the network and then visualize and quantify the difference in distribution between 4 sets of genes:

1. **Digenic genes used in training (N=323):** All genes that are known to be involved in a digenic disease with sufficient confidence, and which have been used in the training set of Hop.
2. **Oligogenic genes (N=762):** All genes that are found to be associated with an oligogenic disease in OLIDAv2.
3. **Disease genes (N=4487):** All genes which are associated with a disease in the Orphanet database.
4. **All human genes (N=20725):** All human genes which are connected in the graph.

The distributions of degrees are shown in Figure C.1. In order to test for significance in the difference between the distributions, we used the Kolgomorov-Smirnov statistical test [491], and used Bonferoni correction [492] to adjust the p-values for the 6 pairwise comparisons. Only comparisons with p-values < 0.05 are shown in Figure C.1.

We observe that gene sets containing genes involved in disease (oligogenic or monogenic, blue, orange and green boxplots in the Figure C.1) all show significant difference with all human genes (Red boxplot), and that they typically present with a higher median degree. However, the distribution of degrees in oligogenic genes does not differ significantly from the distribution of degrees in monogenic genes. This highlights the fact that the bias present in BOCK is probably due to study bias, and it is not specific for oligogenic genes.

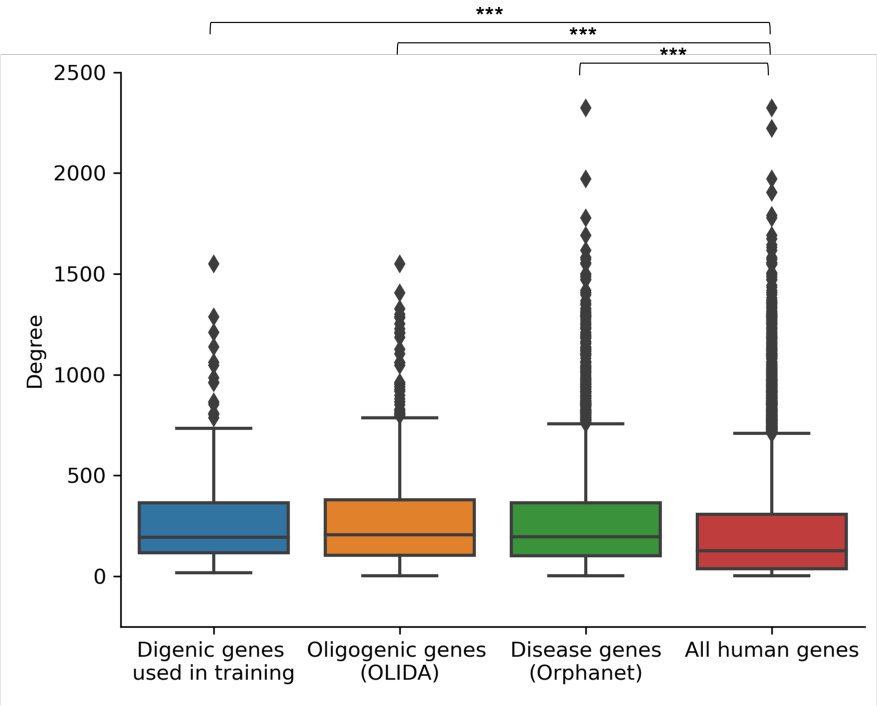


Figure C.1: Distribution of the degrees of different gene sets in BOCK.

Appendix D

Additional information on VCF filtering and gene enrichment analysis

D.1 VCF filtering including the GEMINI patients

In this section, we report on the number of variants in each cohort after different filtering steps when considering the GEMINI patients dataset. We only show the results for the “Routine pipeline”, ’“Sufficient coverage” and “Cohort allele frequency” filters, since after these filtering, the results appear to be too discordant between the GEMINI patients and the two other cohorts, and we thus decided to exclude the GEMINI patients. The number of variants for the GEMINI controls and FIMM patient sets also differs slightly starting from the sufficient coverage filter, since the intersection of the positions selected by this filter was done on the three sets in this case.

		GEMINI controls (N=322)		GEMINI patients (N=81)		FIMM patients (N=429)	
Filter type	Description	Global count	Sample count	Global count	Sample count	Global count	Sample count
Routine pipeline	Remove variants with DP>=10, GQ>=20	692,322	Min: 33,525 Avg: 72,595.5 Med: 70,948 Max: 115,258	390,728	Min: 57,330 Avg: 70,838.7 Med: 70,681 Max: 82,811	910,225	Min: 68,126 Avg: 112,329.5 Med: 111,040 Max: 146,794
Correct for sequencing kits	Remove sites that do not have 10X coverage in at least 90% of the samples	160,057 (23.12%)	Min: 15,250 Avg: 20,254.4 Med: 20,323 Max: 27,704	103,678 (26.53%)	Min: 19,442 Avg: 20,431.0 Med: 20,388 Max: 25,091	180,304 (19.81%)	Min: 21,326 Avg: 22,256.9 Med: 22,253.5 Max: 24,037
Cohort allele frequency 2%	Remove variants with AF>2% in the cohort	96,627 (13.96%)	Min: 437 Avg: 548.51 Med: 537 Max: 1,014	38,435 (9.84%)	Min: 186 Avg: 485.9 Med: 402 Max: 6,473	119,254 (13.1%)	Min: 440 Avg: 538.84 Med: 525.5 Max: 2,398

Table D.1: Variant counts in the VCF files after various filters. The global counts represent all the variants in the VCF file while the sample count show the minimum (Min), average (Avg), median (Med) and maximum (Max) variant counts per sample for each set. The percentage under the global counts show the percentage of variants of the raw files which are retained after applying the filters listed in the rows above and the filter in the current rows. Each filter is shortly described in the description column.

D.2 Gene enrichment plots at different filtering stages

In this section, we show the gene enrichment plots at the different stages of the VCF filtering process between the male infertility patients (referred to as FIMM patients) and the controls (referred to as GEMINI controls).

We observe a huge number of enriched genes when not applying the cohort allele frequency filter (Figure D.1).

The amount of genes that appear to be significantly enriched remains relatively high when removing variants present in more than 5% of the samples in the cohort (Figure D.2).

In both the enrichment plots for the files filtered with 2% frequency in the cohort and 3% frequency in the cohort, the PABPC1 and FMN2 gene appeared to be enriched when considering rare synonymous variants. When assessing the variants present within these genes, we observed that they were mostly long indels around the same genomic locations, which were potentially sequencing errors. This motivated the implementation of the long indels filter, which was the last filter applied before running the Hop analysis.

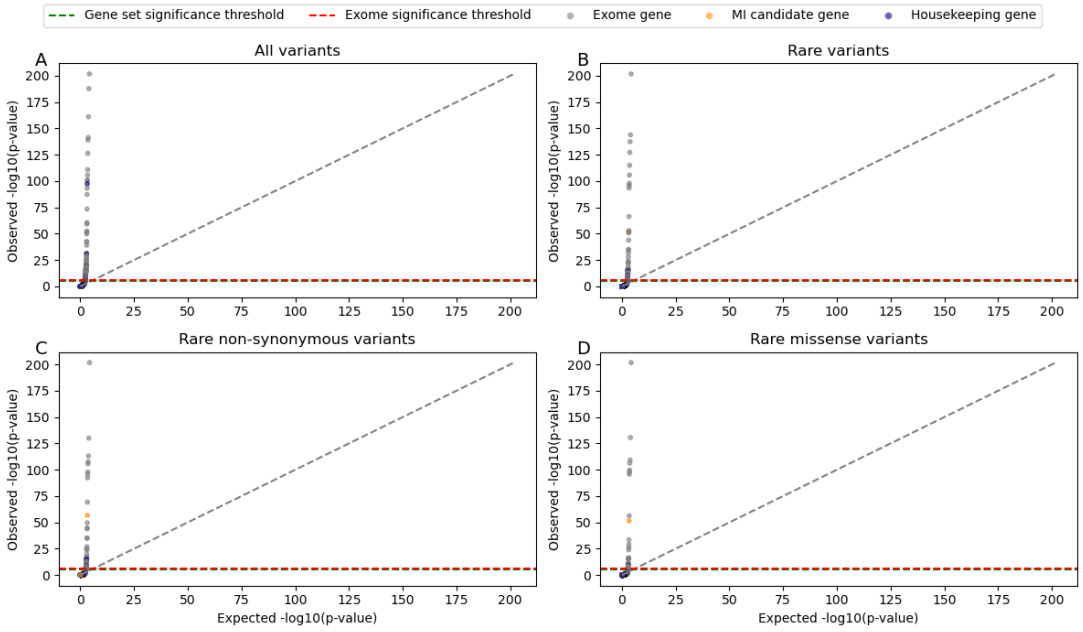


Figure D.1: QQplot of the gene enrichment in patients vs controls **after applying the routine pipeline and the sufficient coverage filter** for different types of qualifying variants : all variants (A), rare variants (B), rare non-synonymous variants (C) and rare missense variants (D).

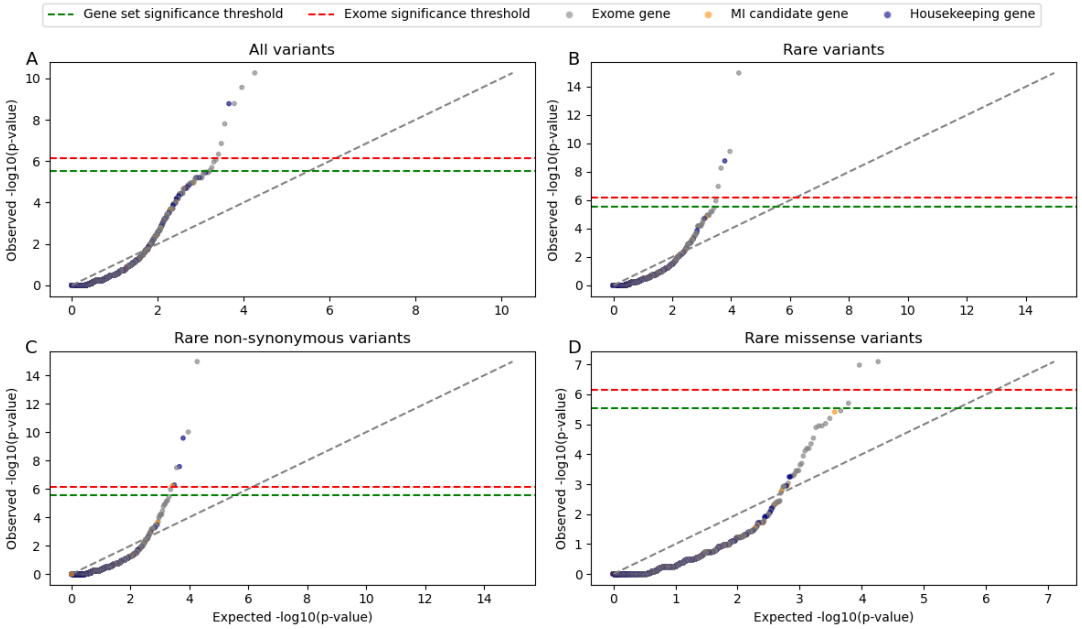


Figure D.2: QQplot of the gene enrichment in patients vs controls **after applying the routine pipeline, the sufficient coverage filter and the cohort allele frequency filter for variants present in more than 5% of samples** for different types of qualifying variants : all variants (A), rare variants (B), rare non-synonymous variants (C) and rare missense variants (D).

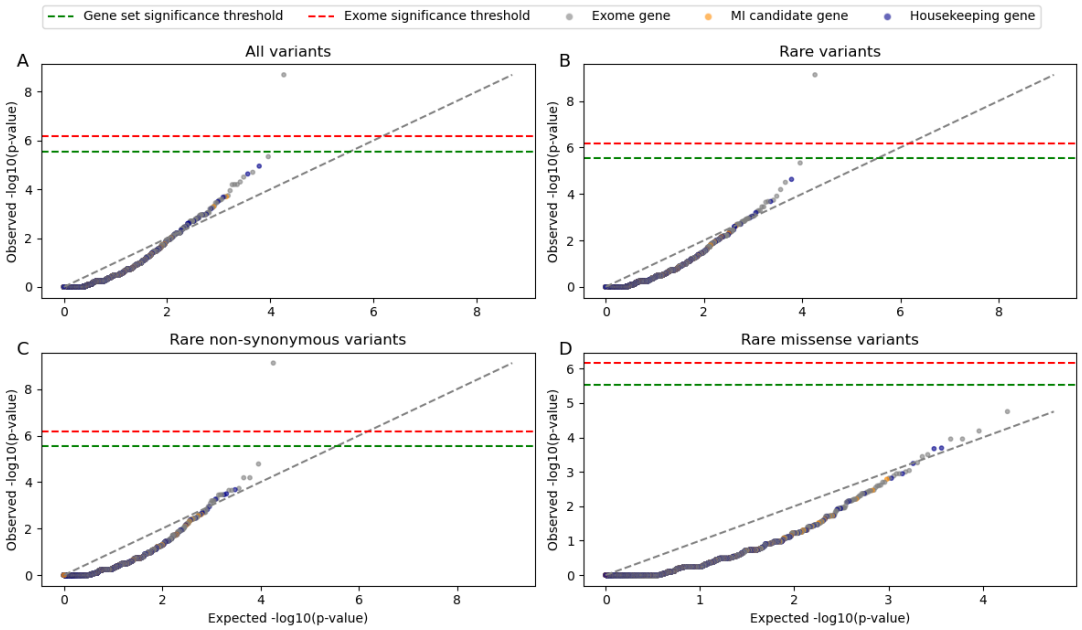


Figure D.3: QQplot of the gene enrichment in patients vs controls **after applying the routine pipeline, the sufficient coverage filter and the cohort allele frequency filter for variants present in more than 3% of samples** for different types of qualifying variants : all variants (A), rare variants (B), rare non-synonymous variants (C) and rare missense variants (D).

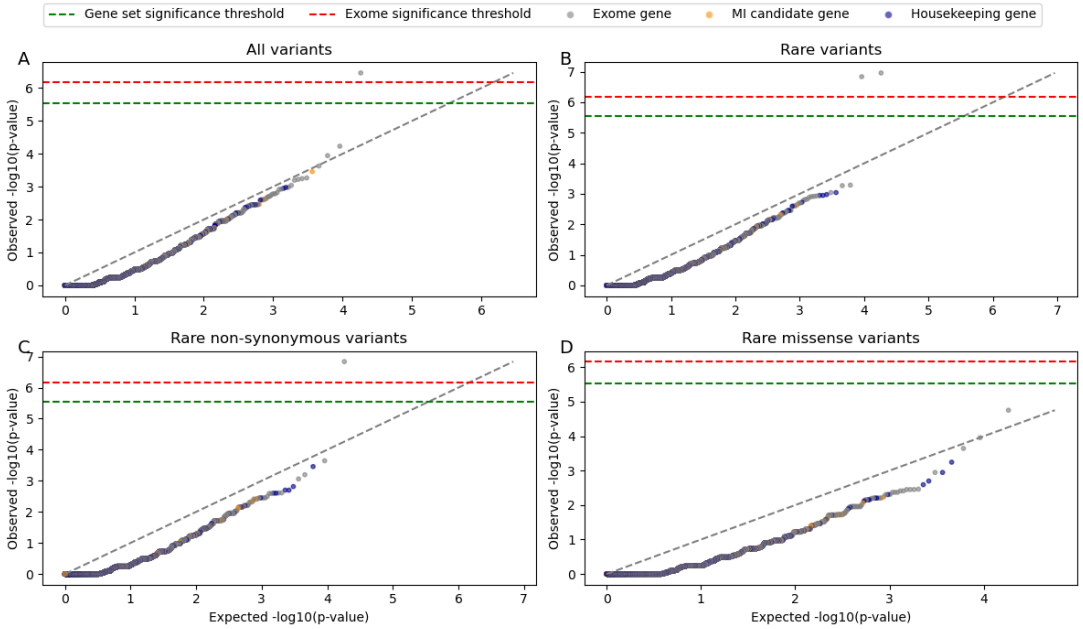


Figure D.4: QQplot of the gene enrichment in patients vs controls **after applying the routine pipeline, the sufficient coverage filter and the cohort allele frequency filter for variants present in more than 2% of samples** for different types of qualifying variants : all variants (A), rare variants (B), rare non-synonymous variants (C) and rare missense variants (D).

Appendix E

Data on diagnostic variants in the male infertility cohort

E.1 Oligogenic diagnostic variants

Cohort	Sample_ID	Gene pair	Variants	Filter
FIMM	Sample_119	TP63;SPRED1	3:189872929:C:T:Heterozygous 15:38351302:C:T:Heterozygous	- -
FIMM	Sample_203	HUWE1;DYRK1A DYRK1A;DHX37 HUWE1;DHX37	X:53580915:C:A:Hemizygous 21:37480708:G:A:Heterozygous 21:37493086:G:T:Heterozygous 12:124971337:C:T:Heterozygous	- - - -
FIMM	Sample_233	SOS2;KIF7 SOS2;CHEK1 CHEK1;KIF7	14:50231258:T:C:Heterozygous 15:89649836:T:G:Heterozygous 11:125644203:C:T:Heterozygous	SufCov - -
FIMM	Sample_274	FANCM;PROKR2	14:45189123:C:T:Homozygous 20:5302327:G:A:Heterozygous	- -
GEMINI	Sample_454	KMT2D;PROK2	12:49024579:G:A:Heterozygous 3:71772801:G:A:Heterozygous 3:71781567:C:T:Heterozygous	- - -
GEMINI	Sample_463	CHEK2;WT1	22:28725099:A:G:Heterozygous 11:32434910:A:C:Heterozygous	- SufCov
GEMINI	Sample_487	NR5A1;PROP1	5:177994145:ACT:A:Heterozygous 9:124491229:C:G:Heterozygous	- IM filter
GEMINI	Sample_491	BRIP1;OTX2	17:61859862:G:C:Heterozygous 14:56802204:G:C:Heterozygous	- -

Table E.1: List of samples and their oligogenic diagnostic variants. The filter column indicates which filter removed the diagnostic variant. SufCov refers to the Sufficient coverage filter and Indel refers to the filter on indels longer than 3bps, IM filter refers to the Inheritance mode filter that removes single heterozygous variants in AR genes. A '-' means that the variant was not filtered. Variants are listed in the same order as the genes.

E.2 Monogenic diagnostic variants

Table E.2: List of samples and their monogenic diagnostic variants. The filter column indicates which filter removed the diagnostic variant. SufCov refers to the Sufficient coverage filter and Indel refers to the filter on indels longer than 3bps.

Cohort	Sample_ID	Gene	Variant	Filter
FIMM	Sample_10	LZTR1	22:20991684:G:A	-
FIMM	Sample_18	ROS1	6:117324366:T:TC:Heterozygous 6:117359831:A:T:Heterozygous	SufCov -
FIMM	Sample_35	TEX14	17:58613423:G:A:Homozygous	-
FIMM	Sample_44	MCMDC2	8:66896253:C:T:Heterozygous 8:66901346:T:A:Heterozygous	- -
FIMM	Sample_62	GLI2	2:120975045:A:G:Heterozygous	-
FIMM	Sample_85	ASZ1	7:117385790:T:C:Homozygous	-
FIMM	Sample_88	GATA4	8:11708799:C:T	SufCov
FIMM	Sample_97		12:112489086:A:G	-
FIMM	Sample_101	FGF8	10:101774779:A:G:Heterozygous	-
FIMM	Sample_104	TCF12	15:57091880:C:A:Heterozygous	-
FIMM	Sample_114	GLUD2	X:121048096:C:T:Hemizygous	-
FIMM	Sample_115	HUWE1	X:53580915:C:A:Hemizygous	-
FIMM	Sample_123	GLI2	2:120975045:A:G:Heterozygous	-
FIMM	Sample_131	DMRT1	9:967999:A:G:Heterozygous	-
FIMM	Sample_133	DHX37	12:124957115:C:CG:Heterozygous	-
FIMM	Sample_135	SYCP2	20:59915200:C:T	SufCov
FIMM	Sample_137	ATRX	X:77682714:C:A	SufCov
FIMM	Sample_147	TGIF2LY	Y:3579775:GCC:G:Hemizygous	-
FIMM	Sample_152	DMRT1	9:847030:C:T:Heterozygous	-
FIMM	Sample_172	SEMA3A	7:84001957:G:A:Heterozygous	-

FIMM	Sample_173	DMRT1	9:847000:G:A:Heterozygous	-
FIMM	Sample_174	AR	X:67643362:C:G:Hemizygous	-
FIMM	Sample_177	NSMF	9:137455628:C:T:Heterozygous	-
FIMM	Sample_186	KLB	4:39447401:TC:T:Heterozygous	-
FIMM	Sample_194	WT1	11:32396277:A:T:Heterozygous	-
FIMM	Sample_200	TUBB3	16:89935404:G:A	-
FIMM	Sample_205		17:31259047:G:T	-
FIMM	Sample_209	NR5A1	9:124483004:A:G:Heterozygous	-
FIMM	Sample_220	NR5A1	9:124500626:G:A:Heterozygous	-
FIMM	Sample_222	TCF12	15:57263126:C:T:Heterozygous	-
FIMM	Sample_236	M1AP	2:74581766:C:CA:Heterozygous	-
FIMM	Sample_239	GATA4	8:11757012:G:A:Heterozygous	-
FIMM	Sample_246	MCMDC2	8:66901346:T:A:Homozygous	-
FIMM	Sample_295	RBM5	3:50093753:G:A:Heterozygous	-
FIMM	Sample_298	PROKR2	20:5314116:C:T:Heterozygous	-
FIMM	Sample_303		15:66481821:G:C	-
FIMM	Sample_305	FGF8	10:101774779:A:G:Heterozygous	-
FIMM	Sample_334	GREB1L	18:21384359:C:T:Heterozygous	-
FIMM	Sample_339	GREB1L	18:21383541:A:C:Heterozygous	-
FIMM	Sample_343	TEX14	17:58613423:G:A:Homozygous	-
FIMM	Sample_353	SOS1	2:39054692:T:G	SufCov
FIMM	Sample_356	AR	X:67546252:T:C:Hemizygous	-
FIMM	Sample_366	NR5A1	9:124500497:CG:C:Heterozygous	-
FIMM	Sample_378	LEO1	15:51965956:G:A:Heterozygous	-
FIMM	Sample_386	NR5A1	9:124500367:G:A:Heterozygous	-
FIMM	Sample_390	ACTRT1	X:128051659:A:AT:Hemizygous	-

FIMM	Sample_399	BNC1	15:83264613:TGACTGCAGCTCTCGATG:T	Indel
FIMM	Sample_405	PROKR2	20:5314116:C:T:Heterozygous	-
FIMM	Sample_424	PROK2	3:71781525:AT:A:Heterozygous	-
GEMINI	Sample_436	FANCM	14:45159189:C:CA 14:45183764:A:G	- SufCov
GEMINI	Sample_455	DHX37	12:124952504:T:C:Heterozygous	-
GEMINI	Sample_464	DHX37	12:124952568:T:C:Heterozygous	-
GEMINI	Sample_477	FANCM	14:45159189:C:CA 14:45183764:A:G	- SufCov
GEMINI	Sample_478	DCAF12L1	X:126552201:CAA:C:Hemizygous	-
GEMINI	Sample_480	AR	X:67546432:C:A	SufCov
GEMINI	Sample_509	DDX3Y	Y:12915947:CTG:C	SufCov
GEMINI	Sample_515	SMCHD1	18:2724951:C:T	SufCov

Appendix F

Details on enriched gene pairs within patient groups

We here show the detailed variant combinations found within each enriched gene pair from Section 6.7. For the “All patients” group, we only show the gene pairs without the IM filter, since the other gene pairs are already present in the main text. For the other groups, we show first in Table format the detail for the gene pairs found to be enriched without applying the IM filter and then in Figures the gene pairs found to be enriched when appluing the IM filter.

F.1 All patients

We here show the detailed variant combinations found in the significantly frequent gene pairs found when not applying the IM filter in all patients vs controls.

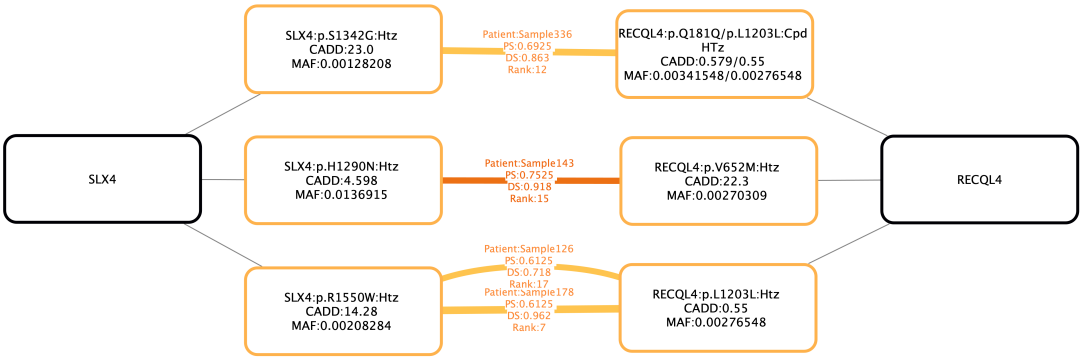


Figure F.1: Variant combinations identified in the gene pair SLX4;RECQL4 in the top 20 of 4 patients of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

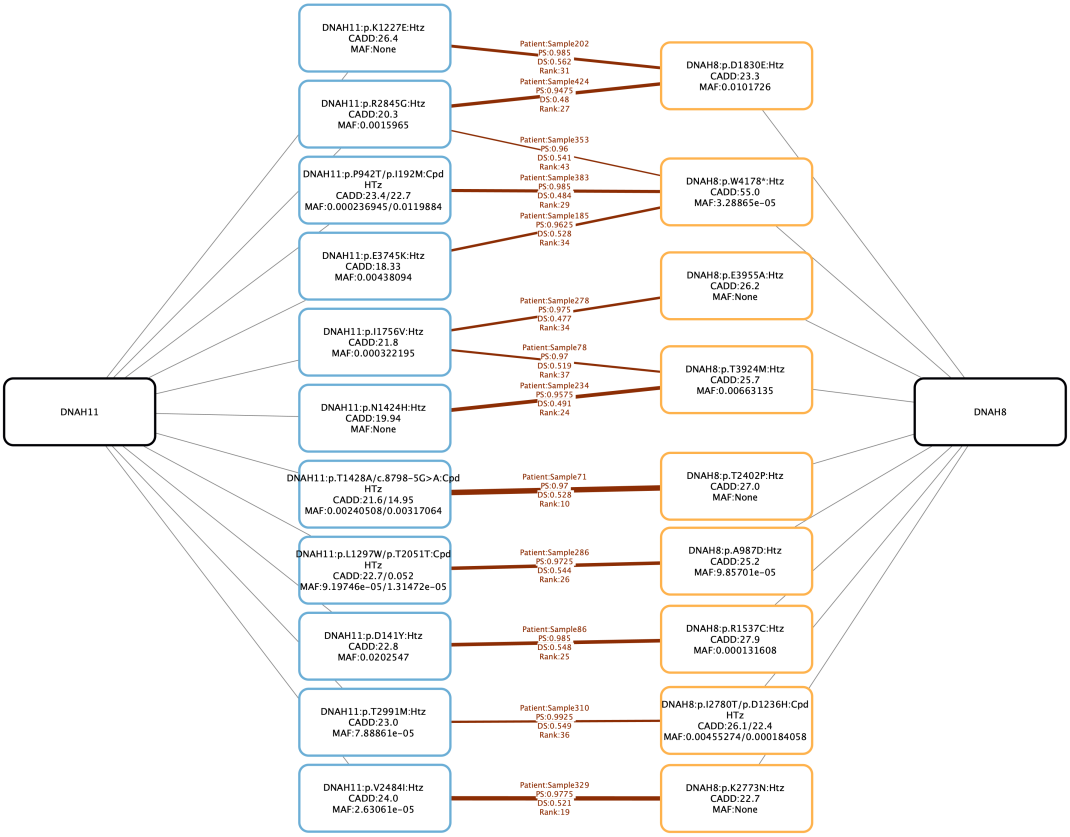


Figure F.2: Variant combinations identified in the gene pair DNAH11;DNAH8 in the top 50 of 13 patients of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

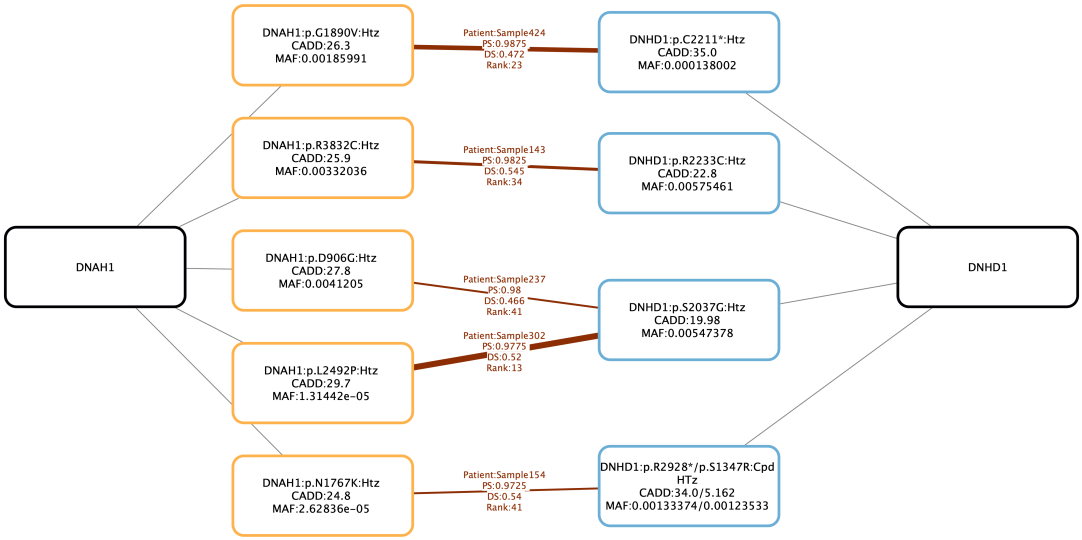


Figure F.3: Variant combinations identified in the gene pair DNAH1;DNHD1 in the top 50 of 5 patients of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

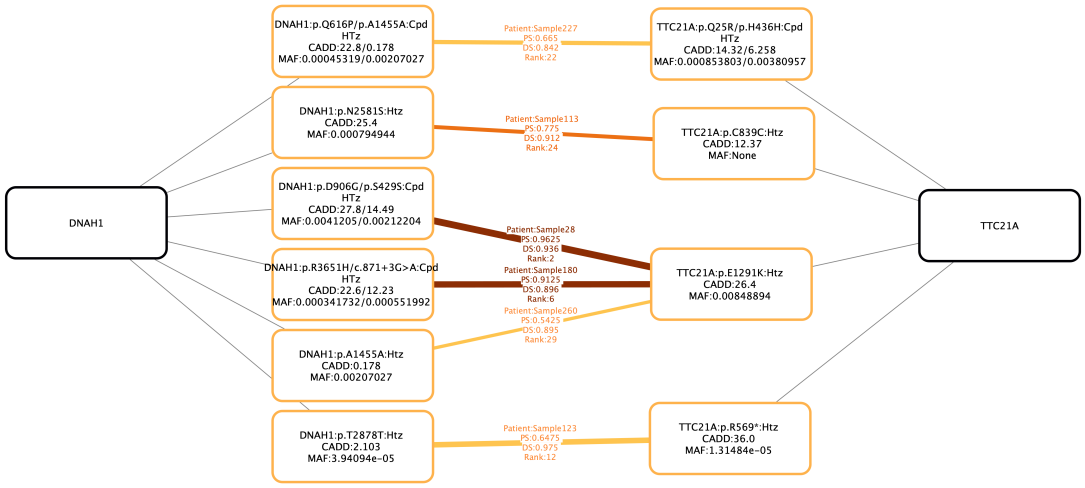


Figure F.4: Variant combinations identified in the gene pair DNAH1;TTC21A in the top 50 of 6 patients of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

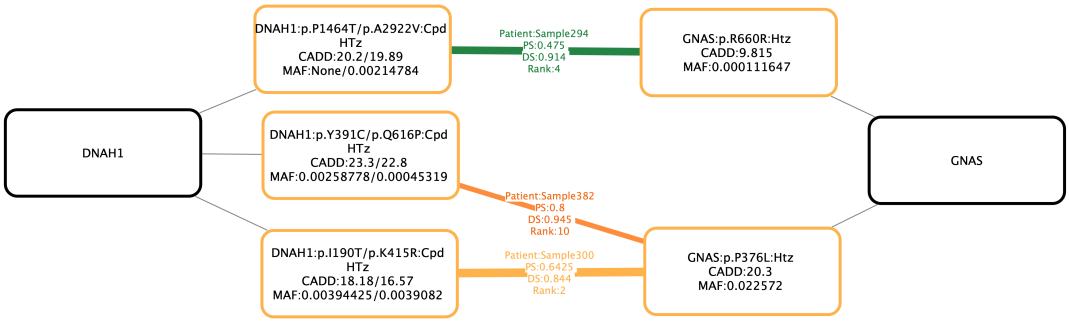


Figure F.5: Variant combinations identified in the gene pair GNAS;DNAH1 in the top 50 of 7 patients of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

F.2 Patients with NOA

F.2.1 Enriched gene pairs with no filters

GenePair	Sample_ID	VariantA	VariantB	CADD A	CADD B	MAF_A	MAF_B	VarCoPP pred.	Rank
BRCA2;SLX4	Sample238	13:32338100:G:A:Htz	16:3591029:G:A:Htz	23.7	8.876	0.0	0.0	Disease-Causing-99% confidence	9
BRCA2;SLX4	Sample240	13:32340810:C:A:Htz	16:3590229:T:C:Htz	17.77	2.406	0.0	0.0	Disease-Causing	5
BRCA2;SLX4	Sample369	13:32379413:G:A:Htz 13:32331021:G:C:Htz	16:3589283:C:G:Htz	26.2 1.936	24.3	0.008 0.0	0.0	Disease-Causing-99.9% confidence	2
BRCA2;FANCA	Sample365	13:32337239:C:T:Htz	16:89765073:CAG:C:Htz	11.92	9.985	0.0	0.0	Disease-Causing	1
BRCA2;FANCA	Sample258	13:32332863:A:G:Htz	16:89740069:C:T:Htz	16.41	0.002	0.0	0.026	Disease-Causing	19
BRCA2;FANCA	Sample369	13:32379413:G:A:Htz 13:32331021:G:C:Htz	16:89740831:G:A:Htz	26.2 1.936	4.403	0.008 0.0	0.0	Disease-Causing-99% confidence	12
ATP2B4;DCC	Sample258	1:203727420:C:T:Htz	18:53157503:G:A:Htz	20.8	27.4	0.008	0.003	Disease-Causing-99.9% confidence	4
ATP2B4;DCC	Sample264	1:203727420:C:T:Htz	18:53066161:A:G:Htz	20.8	22.3	0.008	0.003	Disease-Causing-99% confidence	18
ATP2B4;DCC	Sample327	1:203727420:C:T:Htz	18:53322098:A:G:Htz	20.8	25.4	0.008	0.003	Disease-Causing-99.9% confidence	4
DNAI1;DNAH8	Sample329	9:34514693:C:T:Htz	6:38886850:G:C:Htz	26.9	22.7	0.0	None	Disease-Causing-99.9% confidence	17
DNAI1;DNAH8	Sample354	9:34490011:G:C:Htz	6:38848787:A:C:Htz	26.6	27.5	0.0	0.0	Disease-Causing-99.9% confidence	32
DNAI1;DNAH8	Sample71	9:34513145:T:C:Htz	6:38872749:A:C:Htz	21.7	27.0	0.0	None	Disease-Causing-99.9% confidence	7
CREBBP;AKAP9	Sample261	16:3729247:A:G:Htz	7:92085625:T:A:Htz	22.5	23.3	0.0	None	Disease-Causing-99.9% confidence	37
CREBBP;AKAP9	Sample71	16:3781229:G:T:Htz	7:92098173:A:G:Htz	22.5	18.41	0.008	0.001	Disease-Causing-99% confidence	31

CREBBP;AKAP9	Sample82	16:3770772:G:A:Htz	7:92070953:T:C:Htz	23.7	20.5	0.001	0.002	Disease-Causing-99.9% confidence	9

F.2.2 Enriched gene pairs with Inheritance Mode filter

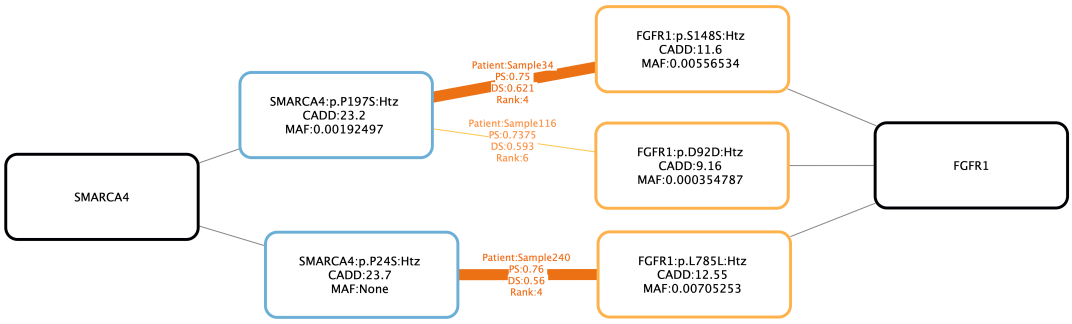


Figure F.6: Variant combinations identified in the gene pair SMARCA4;FGFR1 in the top 50 of 3 patients with NOA of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

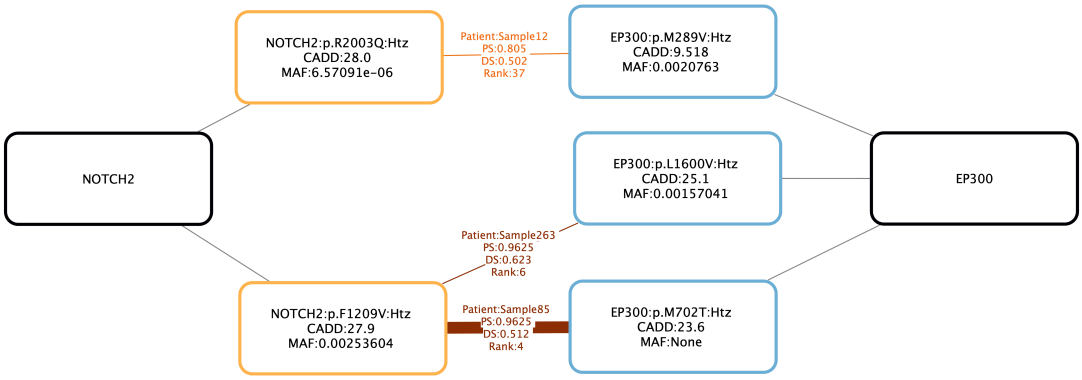


Figure F.7: Variant combinations identified in the gene pair NOTCH2;EP300 in the top 50 of 3 patients with NOA of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

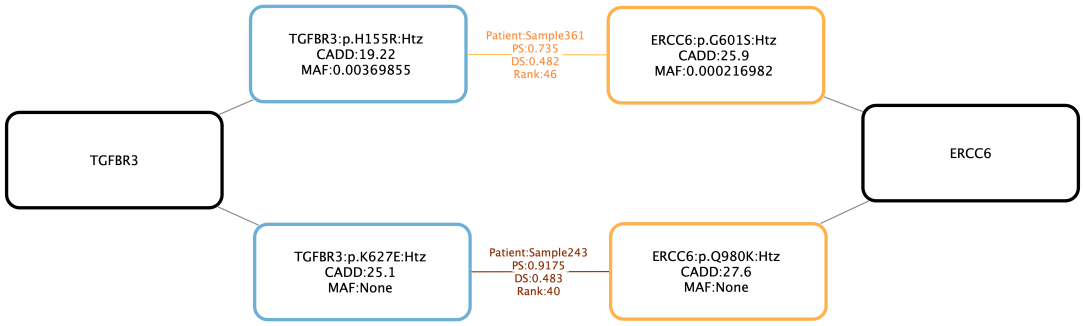


Figure F.8: Variant combinations identified in the gene pair TGFBR3;ERCC6 in the top 50 of 2 patients with NOA of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

F.3 Patients with OZ

F.3.1 Enriched gene pairs with no filter

GenePair	Sample_ID	VariantA	VariantB	CADD A	CADD B	MAF_A	MAF_B	VarCoPP pred.	Rank
CHD7;RECQL4	Sample76	8:60741739:T:A:Htz	8:144513256:C:T:Htz	22.9	22.4	0.013	0.0	Disease-Causing-99.9% confidence	3
CHD7;RECQL4	Sample192	8:60865619:C:T:Htz	8:144513412:G:A:Htz	25.1	39.0	0.0	0.0	Disease-Causing-99.9% confidence	1
CHD7;RECQL4	Sample143	8:60781400:C:T:Htz	8:144514032:C:T:Htz	25.0	22.3	0.0	0.003	Disease-Causing-99.9% confidence	2
SETX;BLM	Sample304	9:132349370:C:T:Htz	15:90747403:T:C:Htz	3.279	23.6	0.006	0.001	Disease-Causing-99% confidence	5
SETX;BLM	Sample86	9:132329848:G:C:Htz	15:90763011:G:A:Htz	18.66	8.089	0.0	0.003	Disease-Causing-99% confidence	9
SETX;BLM	Sample404	9:132349370:C:T:Htz	15:90754819:A:G:Htz	3.279	22.0	0.006	0.001	Disease-Causing-99% confidence	8
SBF1;BMPR1B	Sample241	22:50466679:T:G:Htz 22:50466645:G:A:Htz	4:95129947:G:A:Htz	26.9 1.495	25.8	None 0.002	0.001	Disease-Causing-99.9% confidence	3
SBF1;BMPR1B	Sample199	22:50459254:C:T:Htz	4:95148783:G:A:Htz	34.0	28.3	None	0.001	Disease-Causing-99.9% confidence	10
SBF1;BMPR1B	Sample399	22:50447434:C:T:Htz	4:95129947:G:A:Htz	25.1	25.8	0.0	0.001	Disease-Causing-99.9% confidence	2
ATM;RECQL4	Sample182	11:108244065:C:T:Htz	8:144514199:C:T:Htz	7.609	24.7	0.002	0.001	Disease-Causing-99% confidence	16
ATM;RECQL4	Sample120	11:108272729:C:G:Htz 11:108267276:T:C:Htz	8:144511436:G:A:Htz	22.8 14.93	22.2	0.017 0.008	0.001	Disease-Causing-99.9% confidence	1
ATM;RECQL4	Sample143	11:108289005:C:T:Htz	8:144514032:C:T:Htz	9.085	22.3	0.021	0.003	Disease-Causing	17
ATM;RECQL4	Sample306	11:108289034:A:G:Htz	8:144514032:C:T:Htz	7.977	22.3	0.0	0.003	Disease-Causing	6
ATM;RECQL4	Sample211	11:108227865:T:C:Htz	8:144515323:C:T:Htz	3.535	1.438	0.002	0.002	Neutral	49
ATM;RECQL4	Sample204	11:108227849:C:G:Htz	8:144511449:G:A:Htz	23.9	0.55	0.007	0.003	Disease-Causing	40

ATM;RECQL4	Sample32	11:108326110:G:C:Htz	8:144512483:G:C:Htz	21.4	7.404	0.0	None	Disease-Causing-99% confidence	7
RECQL4;MCM9	Sample129	8:144514032:C:T:Htz	6:118829055:T:A:Htz 6:118931629:C:G:Htz	22.3	24.5 4.615	0.003	0.004 0.0	Disease-Causing-99.9% confidence	3
RECQL4;MCM9	Sample55	8:144514032:C:T:Htz	6:118856543:G:A:Htz	22.3	13.57	0.003	0.0	Disease-Causing-99% confidence	7
RECQL4;MCM9	Sample430	8:144516772:G:A:Htz	6:118913414:T:C:Htz 6:118856424:C:T:Htz	0.744	25.7 9.434	0.0	0.003 0.005	Disease-Causing 99% confidence	13
RECQL4;MCM9	Sample171	8:144511990:C:T:Htz	6:118816260:G:A:Htz	10.7	9.727	0.004	0.006	Disease-Causing	22
CHD7;GLI2	Sample320	8:60862233:C:T:Htz	2:120990682:G:A:Htz	26.9	0.092	None	0.0	Disease-Causing	5
CHD7;GLI2	Sample206	8:60845246:C:T:Htz	2:120989815:G:A:Htz	8.019	22.3	0.006	None	Disease-Causing	10
CHD7;GLI2	Sample291	8:60742905:C:T:Htz	2:120990472:G:A:Htz 2:120989968:A:G:Htz	10.64	28.6 5.846	0.0	0.009 0.009	Disease-Causing-99% confidence	17
CHD7;GLI2	Sample200	8:60865611:A:G:Htz	2:120955394:G:A:Htz	19.35	2.376	0.0	0.001	Disease-Causing	31
SPTBN2;DCC	Sample406	11:66708270:G:A:Htz	18:53530568:A:G:Htz	0.926	27.8	0.002	None	Disease-Causing	24
SPTBN2;DCC	Sample13	11:66701195:C:T:Htz	18:53066161:A:G:Htz	0.533	22.3	0.001	0.003	Neutral	46
SPTBN2;DCC	Sample62-409	11:66696533:C:T:Htz	18:53305725:T:C:Htz	16.12	22.0	0.0	None	Disease-Causing	16
TOP2A;RECQL4	Sample204	17:40400903:C:G:Htz	8:144511449:G:A:Htz	25.6	0.55	None	0.003	Disease-Causing	32
TOP2A;RECQL4	Sample32	17:40398936:A:G:Htz	8:144512483:G:C:Htz	17.04	7.404	0.003	None	Disease-Causing	9
TOP2A;RECQL4	Sample72	17:40395479:T:C:Htz	8:144516576:C:T:Htz	22.3	0.579	None	0.003	Disease-Causing	40
TOP2A;RECQL4	Sample302	17:40398936:A:G:Htz	8:144512237:T:C:Htz	17.04	0.044	0.003	0.0	Disease-Causing	9
KLC2;DNAH1	Sample429	11:66261880:A:G:Htz	3:52381677:A:G:Htz	27.4	25.1	0.0	0.0	Disease-Causing-99.9% confidence	36
KLC2;DNAH1	Sample32	11:66265923:C:T:Htz	3:52347945:G:A:Htz	23.5	21.7	0.001	None	Disease-Causing-99.9% confidence	27
KLC2;DNAH1	Sample368	11:66262174:G:A:Htz	3:52370168:A:C:Htz	26.8	25.0	0.001	None	Disease-Causing-99.9% confidence	9
CSMD1;SH3TC2	Sample321	8:3409570:G:A:Htz	5:149028434:G:A:Htz	25.3	22.7	0.0	0.001	Disease-Causing	33

CSMD1;SH3TC2	Sample128	8:4637444:C:T:Htz	5:149028434:G:A:Htz	24.2	22.7	0.0	0.001	Disease-Causing	40
CSMD1;SH3TC2	Sample52	8:2942497:A:T:Htz	5:149010302:G:A:Htz	23.4	30.0	0.0	0.0	Disease-Causing	2
NOTCH2;ATM	Sample301	1:119915647:G:C:Htz	11:108289789:A:G:Htz	14.64	22.2	0.005	0.001	Disease-Causing-99% confidence	10
NOTCH2;ATM	Sample328	1:120005564:T:C:Htz	11:108326110:G:C:Htz	7.551	21.4	0.0	0.0	Disease-Causing	25
NOTCH2;ATM	Sample187	1:120005472:C:A:Htz	11:108252828:A:G:Htz	22.1	18.54	0.009	0.0	Disease-Causing-99.9% confidence	8
NOTCH2;ATM	Sample211	1:119926524:T:C:Htz	11:108227865:T:C:Htz	22.6	3.535	0.006	0.002	Disease-Causing	12
NOTCH2;ATM	Sample143	1:119922384:T:A:Htz 1:119941711:A:C:Htz	11:108289005:C:T:Htz	19.01 8.116	9.085	0.001 0.0	0.021	Disease-Causing	23
NOTCH2;ATM	Sample58	1:120005472:C:A:Htz	11:108326110:G:C:Htz	22.1	21.4	0.009	0.0	Disease-Causing-99.9% confidence	4
NOTCH2;ATM	Sample57	1:120005472:C:A:Htz	11:108272729:C:G:Htz 11:108267276:T:C:Htz	22.1	22.8 14.93	0.009	0.017 0.008	Disease-Causing-99.9% confidence	4
EPHB1;SETX	Sample282	3:134951923:A:G:Htz	9:132330224:A:C:Htz	22.4	23.0	0.0	0.0	Disease-Causing	23
EPHB1;SETX	Sample86	3:135166066:G:A:Htz	9:132329848:G:C:Htz	21.2	18.66	0.002	0.0	Disease-Causing	33
EPHB1;SETX	Sample57	3:135201495:G:A:Htz	9:132326938:A:C:Htz	25.2	25.0	0.0	0.003	Disease-Causing-99% confidence	16
NOTCH2;EIF2B2	Sample321	1:120005472:C:A:Htz	14:75003066:G:C:Htz	22.1	22.9	0.009	0.002	Disease-Causing-99% confidence	7
NOTCH2;EIF2B2	Sample278	1:119926524:T:C:Htz	14:75003066:G:C:Htz	22.6	22.9	0.006	0.002	Disease-Causing-99% confidence	14
NOTCH2;EIF2B2	Sample57	1:120005472:C:A:Htz	14:75003066:G:C:Htz	22.1	22.9	0.009	0.002	Disease-Causing-99% confidence	24
NOTCH2;TTC21A	Sample171	1:119915647:G:C:Htz 1:119929116:C:T:Htz	3:39129310:G:A:Htz	14.64 12.36	32.0	0.005 0.001	0.002	Disease-Causing	39
NOTCH2;TTC21A	Sample321	1:120005472:C:A:Htz	3:39129310:G:A:Htz	22.1	32.0	0.009	0.002	Disease-Causing-99% confidence	5

NOTCH2;TTC21A	Sample57	1:120005472:C:A:Htz	3:39137690:C:T:Htz	22.1	25.1	0.009	0.0	Disease-Causing-99% confidence	28
ATM;ME1	Sample109	11:108299779:A:C:Htz	6:83237739:T:G:Htz	17.59	28.7	0.002	0.0	Disease-Causing	33
ATM;ME1	Sample58	11:108326110:G:C:Htz	6:83430935:C:CG:Htz	21.4	25.0	0.0	0.0	Disease-Causing-99% confidence	9
ATM;ME1	Sample120	11:108272729:C:G:Htz 11:108267276:T:C:Htz	6:83253698:T:C:Htz	22.8 14.93	25.5	0.017 0.008	0.0	Disease-Causing-99% confidence	4
CHD7;MCM8	Sample328	8:60741739:T:A:Htz	20:5986117:A:G:Htz	22.9	23.7	0.013	0.008	Disease-Causing-99.9% confidence	3
CHD7;MCM8	Sample58	8:60742450:A:G:Htz	20:5986117:A:G:Htz	15.57	23.7	0.005	0.008	Disease-Causing	34
CHD7;MCM8	Sample413	8:60742450:A:G:Htz	20:5955229:G:A:Htz 20:5955186:A:G:Htz	15.57	26.5 9.812	0.005	0.011 0.001	Disease-Causing	3
DCC;POTEJ	Sample429	18:53066161:A:G:Htz	2:130656831:C:T:Htz	22.3	22.9	0.003	0.007	Disease-Causing	46
DCC;POTEJ	Sample406	18:53530568:A:G:Htz	2:130656781:A:T:Htz	27.8	19.05	None	0.006	Disease-Causing	25
DCC;POTEJ	Sample62-409	18:53305725:T:C:Htz	2:130656831:C:T:Hmz	22.0	22.9	None	0.007	Disease-Causing	26
EPHB1;DCC	Sample429	3:135248421:C:A:Htz	18:53066161:A:G:Htz	9.02	22.3	0.008	0.003	Disease-Causing	23
EPHB1;DCC	Sample401	3:135166066:G:A:Htz	18:52906285:C:A:Htz	21.2	11.08	0.002	0.0	Disease-Causing-99% confidence	3
EPHB1;DCC	Sample62-409	3:135179912:C:T:Htz	18:53305725:T:C:Htz	0.281	22.0	0.0	None	Disease-Causing	39
ROS1;USP42	Sample304	6:117356626:T:C:Htz	7:6149652:G:A:Htz	16.89	25.5	0.001	0.0	Disease-Causing	24
ROS1;USP42	Sample275	6:117300987:A:C:Htz	7:6154748:G:C:Htz	22.8	23.3	0.02	0.012	Disease-Causing-99% confidence	28
ROS1;USP42	Sample70	6:117394218:A:G:Htz	7:6149620:A:G:Htz	24.1	24.8	0.002	None	Disease-Causing-99% confidence	3
DNAH11;DNAH8	Sample286	7:21615151:T:G:Htz 7:21698186:G:T:Htz	6:38803237:C:A:Htz	22.7 0.052	25.2	0.0 0.0	0.0	Disease-Causing-99.9% confidence	25
DNAH11;DNAH8	Sample78	7:21658969:A:G:Htz 7:21571797:T:C:Htz	6:38938181:C:T:Htz	21.8 17.4	25.7	0.0 0.006	0.007	Disease-Causing-99.9% confidence	30

DNAH11;DNAH8	Sample86	7:21545075:G:T:Htz	6:38842667:C:T:Htz	22.8	27.9	0.02	0.0	Disease-Causing-99.9% confidence	24
DNAH11;DNAH8	Sample278	7:21658969:A:G:Htz	6:38938845:A:C:Htz	21.8	26.2	0.0	None	Disease-Causing-99.9% confidence	33
DNAH11;DNAH8	Sample202	7:21606456:A:G:Htz	6:38852717:C:A:Htz 6:38931867:T:A:Htz	26.4	23.3 0.379	None	0.01 0.011	Disease-Causing-99.9% confidence	30
DNAH11;DNAH8	Sample185	7:21861883:G:A:Htz	6:38973669:G:A:Htz 6:38852717:C:A:Htz	18.33	55.0 23.3	0.004	0.0 0.01	Disease-Causing-99.9% confidence	33
CFAP61;DNAH1	Sample241	20:20074378:G:A:Htz	3:52392627:G:A:Htz	34.0	24.7	0.0	0.002	Disease-Causing-99.9% confidence	33
CFAP61;DNAH1	Sample285	20:20290308:C:A:Htz	3:52395043:G:A:Htz 3:52328017:G:A:Htz	25.4	22.6 12.23	0.008	0.0 0.001	Disease-Causing-99.9% confidence	39
CFAP61;DNAH1	Sample302	20:20228327:G:A:Htz	3:52380002:T:C:Htz	23.9	29.7	0.02	0.0	Disease-Causing-99.9% confidence	11
CFAP61;DNAH1	Sample395	20:20196636:G:A:Htz	3:52388836:C:T:Htz	23.2	28.9	0.001	0.0	Disease-Causing-99.9% confidence	23
USP42;PRDM9	Sample419	7:6154748:G:C:Htz	5:23527533:T:TGG:Htz 5:23527530:CAA:C:Htz	23.3	23.5 14.75	0.012	0.004 0.004	Disease-Causing-99% confidence	9
USP42;PRDM9	Sample275	7:6154748:G:C:Htz	5:23527732:C:T:Htz	23.3	22.7	0.012	0.0	Disease-Causing-99% confidence	22
USP42;PRDM9	Sample416	7:6135895:C:T:Htz	5:23527321:G:A:Htz	22.9	24.6	0.003	0.001	Disease-Causing-99% confidence	13

F.3.2 Enriched gene pairs with Inheritance Mode filter

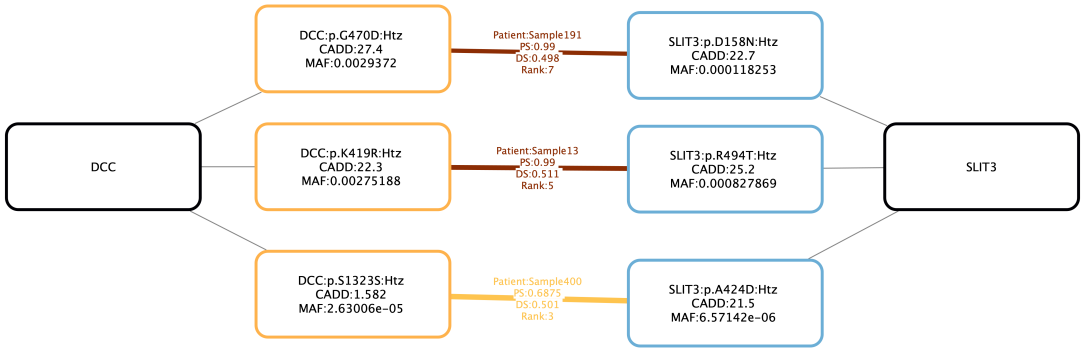


Figure F.9: Variant combinations identified in the gene pair SLIT3;DCC in the top 50 of 3 patients with OZ of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

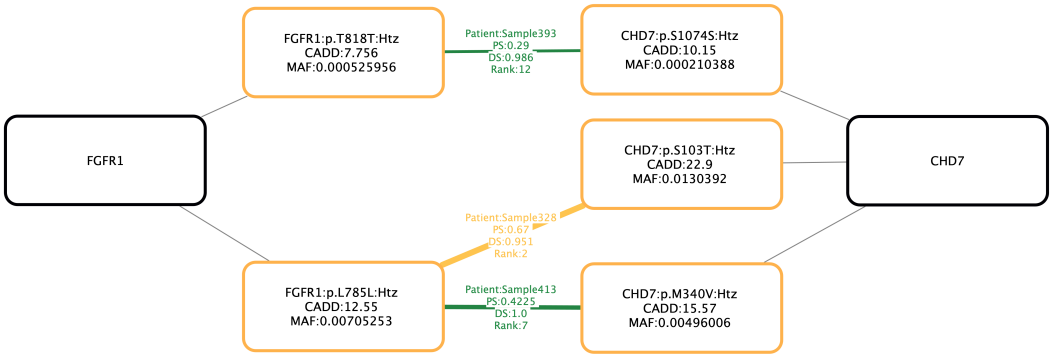


Figure F.10: Variant combinations identified in the gene pair CHD7;FGFR1 in the top 50 of 3 patients with OZ of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

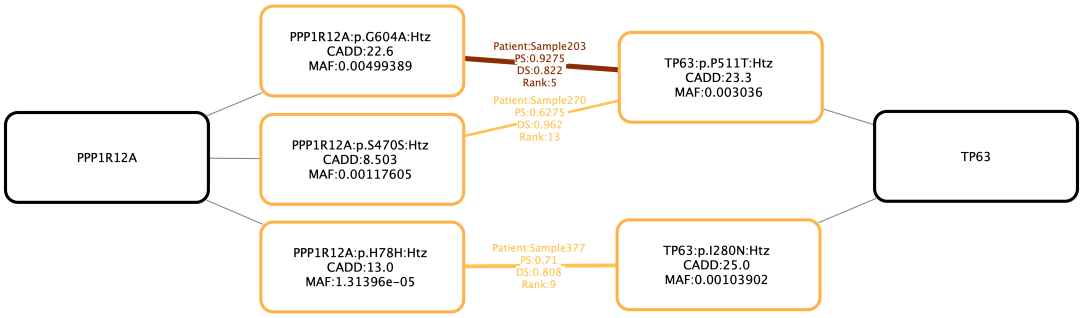


Figure F.11: Variant combinations identified in the gene pair TP63;PPP1R12A in the top 50 of 3 patients with OZ of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

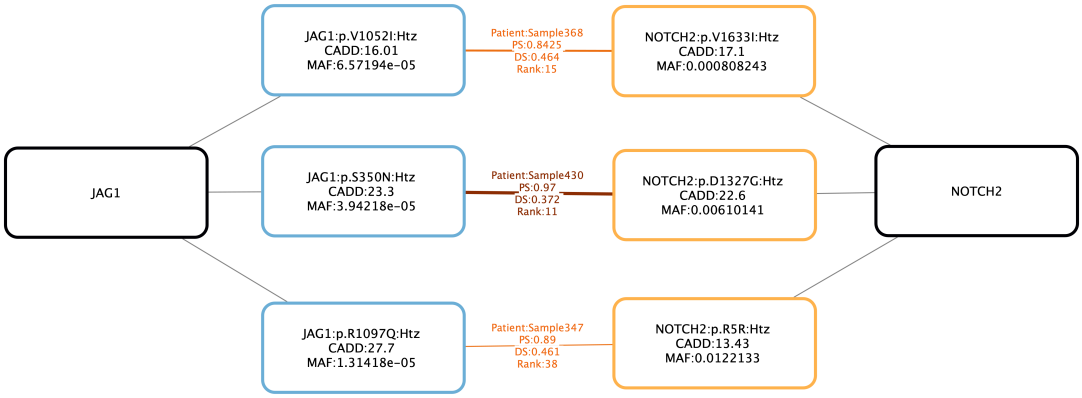


Figure F.12: Variant combinations identified in the gene pair NOTCH2;JAG1 in the top 50 of 3 patients with OZ of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

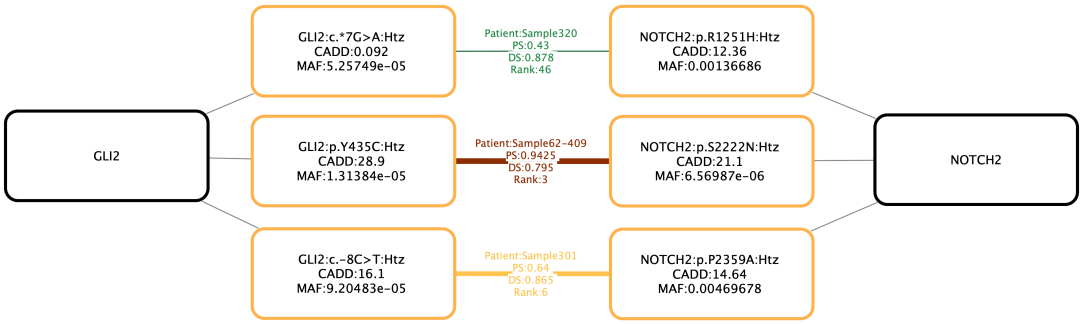


Figure F.13: Variant combinations identified in the gene pair NOTCH2;GLI2 in the top 50 of 3 patients with OZ of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

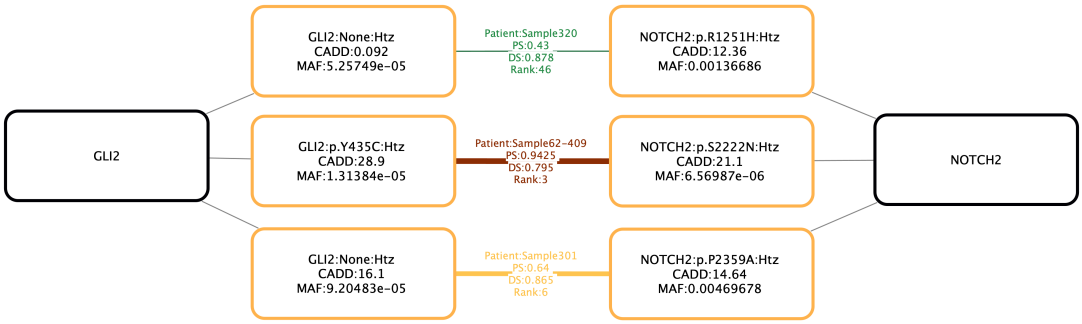


Figure F.14: Variant combinations identified in the gene pair NOTCH2;MCM2 in the top 50 of 3 patients with OZ of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

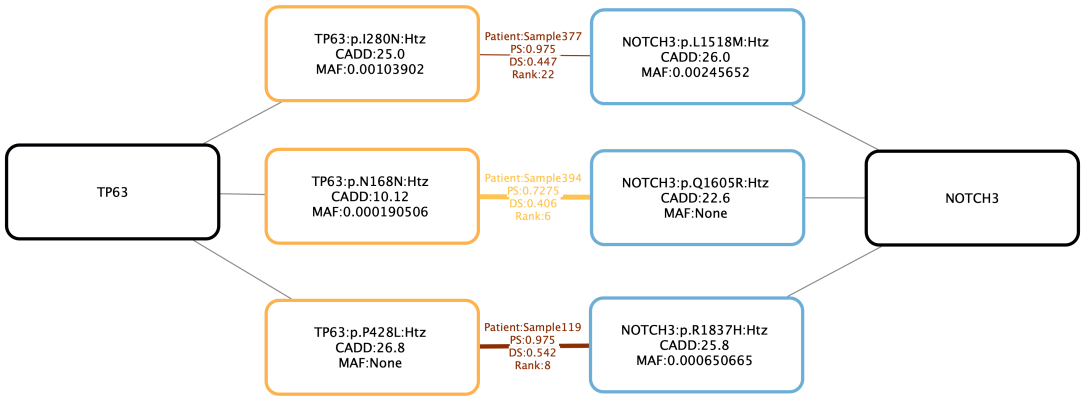


Figure F.15: Variant combinations identified in the gene pair TP63;NOTCH3 in the top 50 of 3 patients with OZ of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

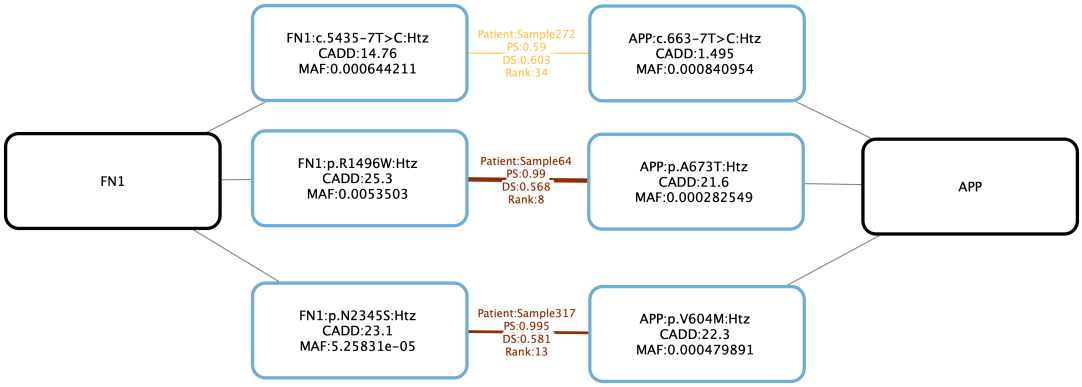


Figure F.16: Variant combinations identified in the gene pair APP;FN1 in the top 50 of 3 patients with OZ of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

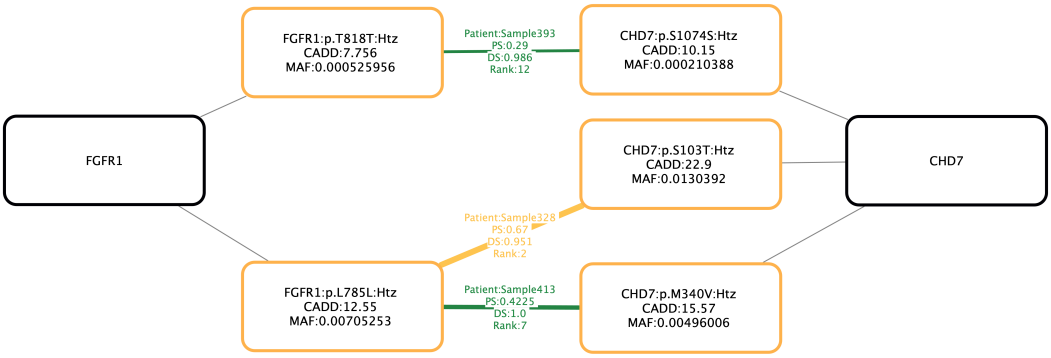


Figure F.17: Variant combinations identified in the gene pair CHD7;FGFR1 in the top 50 of 3 patients with OZ of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

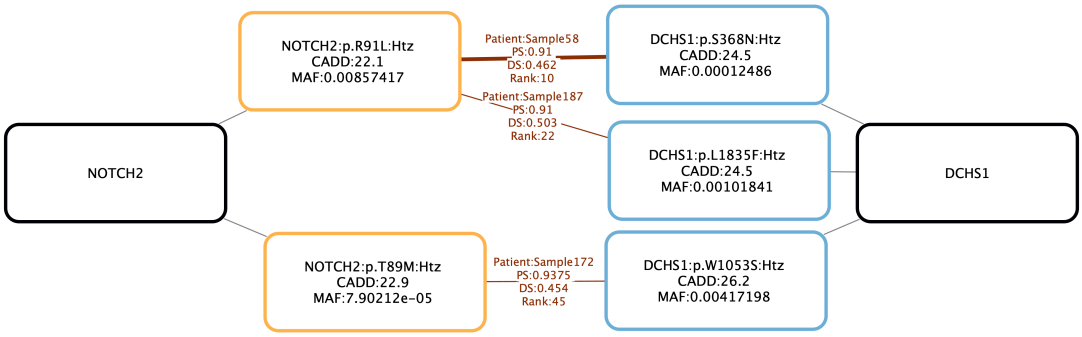


Figure F.18: Variant combinations identified in the gene pair NOTCH2;DCHS1 in the top 50 of 3 patients with OZ of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

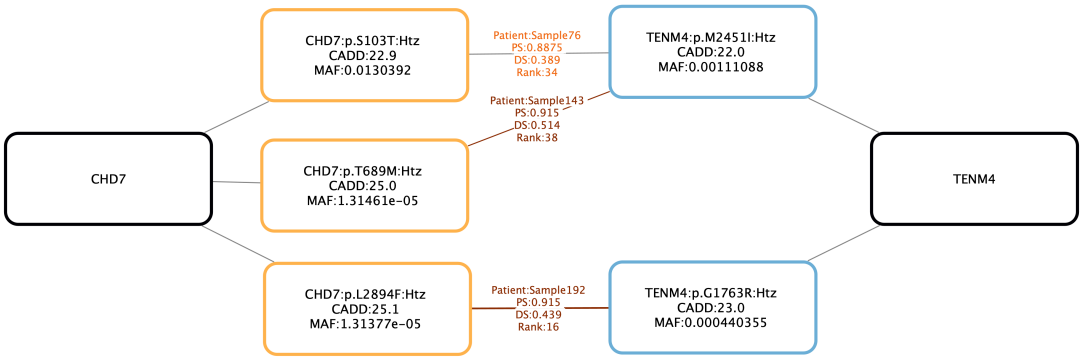


Figure F.19: Variant combinations identified in the gene pair CHD7;TENM4 in the top 50 of 3 patients with OZ of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

F.4 Patients with cryptorchidism

F.4.1 Enriched gene pairs with no filters

GenePair	Sample_ID	VariantA	VariantB	CADD A	CADD B	MAF_A	MAF_B	VarCoPP pred.	Rank
ROS1;DNAH1	Sample41	6:117300987:A:C:Htz	3:52326879:C:G:Htz 3:52332280:A:G:Htz	22.8	32.0 23.3	0.02	0.0 0.003	Disease-Causing- 99% confidence	3
ROS1;DNAH1	Sample113	6:117318198:G:T:Htz	3:52381773:A:G:Htz	24.4	25.4	0.0	0.001	Disease-Causing- 99% confidence	8
ROS1;DNAH1	Sample170	6:117300974:C:T:Htz	3:52375988:G:A:Htz	24.6	23.7	0.001	0.002	Disease-Causing- 99% confidence	6
POLR3A;SETX	Sample75	10:78021605:G:A:Htz	9:132342716:A:C:Htz	28.6	23.4	None	0.003	Disease-Causing- 99.9% confidence	1
POLR3A;SETX	Sample223	10:77986123:T:C:Htz	9:132349370:C:T:Htz	22.6	3.279	0.006	0.006	Disease-Causing	2
POLR3A;SETX	Sample90	10:78022289:C:T:Htz	9:132327789:G:A:Htz	2.72	23.4	0.005	0.001	Disease-Causing	18
WDR11;SEMA5A	Sample333	10:120904077:G:A:Htz	5:9237823:C:T:Htz	30.0	17.85	0.001	0.0	Disease-Causing	15
WDR11;SEMA5A	Sample83	10:120904077:G:A:Htz	5:9044470:G:A:Htz	30.0	23.0	0.001	0.001	Disease-Causing- 99% confidence	15
WDR11;SEMA5A	Sample140	10:120904077:G:A:Htz	5:9066553:C:T:Htz	30.0	19.74	0.001	0.0	Disease-Causing	13
TTC21A;DNAH1	Sample227	3:39109131:A:G:Htz 3:39125424:C:T:Htz	3:52346662:A:C:Htz 3:52359344:C:T:Htz	14.32 6.258	22.8 0.178	0.001 0.004	0.0 0.002	Disease-Causing	10
TTC21A;DNAH1	Sample180	3:39138609:G:A:Htz 3:39130250:C:T:Htz	3:52395043:G:A:Htz 3:52328017:G:A:Htz	26.4 0.294	22.6 12.23	0.008 0.0	0.0 0.001	Disease-Causing- 99.9% confidence	5
TTC21A;DNAH1	Sample113	3:39131029:C:T:Htz	3:52381773:A:G:Htz	12.37	25.4	None	0.001	Disease-Causing- 99% confidence	13
SLX4;RECQL4	Sample178	16:3584860:G:A:Htz	8:144511449:G:A:Htz	14.28	0.55	0.002	0.003	Disease-Causing	6
SLX4;RECQL4	Sample126	16:3584860:G:A:Htz	8:144511449:G:A:Htz	14.28	0.55	0.002	0.003	Disease-Causing	16
SLX4;RECQL4	Sample336	16:3589614:T:C:Htz	8:144516576:C:T:Htz 8:144511449:G:A:Htz	23.0	0.579 0.55	0.001	0.003 0.003	Disease-Causing	11
SBF1;SEMA5A	Sample83	22:50447434:C:T:Htz	5:9044470:G:A:Htz	25.1	23.0	0.0	0.001	Disease-Causing	23
SBF1;SEMA5A	Sample3	22:50466179:C:T:Htz	5:9226869:C:T:Htz	23.2	23.3	0.002	0.0	Disease-Causing	34

SBF1;SEMA5A	Sample21	22:50464824:T:C:Htz	5:9226869:C:T:Htz	23.6	23.3	0.001	0.0	Disease-Causing	14
GNAS;DNAH1	Sample294	20:58855245:C:T:Htz	3:52359369:C:A:Htz 3:52386299:C:T:Htz	9.815	20.2 19.89	0.0	None 0.002	Neutral	36
GNAS;DNAH1	Sample174	20:58854392:C:T:Htz	3:52392627:G:A:Htz	20.3	24.7	0.023	0.002	Disease-Causing-99% confidence	17
GNAS;DNAH1	Sample300	20:58854392:C:T:Htz	3:52326302:T:C:Htz 3:52332352:A:G:Htz	20.3	18.18 16.57	0.023	0.004 0.004	Disease-Causing	50
GNAS;DNAH1	Sample29	20:58853749:A:G:Htz	3:52352606:G:A:Htz	16.47	22.9	0.002	0.0	Disease-Causing	31
GNAS;DNAH1	Sample382	20:58854392:C:T:Htz	3:52332280:A:G:Htz 3:52346662:A:C:Htz	20.3	23.3 22.8	0.023	0.003 0.0	Disease-Causing-99% confidence	16
DNAH1;DNHD1	Sample227	3:52346662:A:C:Htz 3:52359344:C:T:Htz	11:6498469:A:T:Htz	22.8 0.178	18.98	0.0 0.002	None	Disease-Causing-99.9% confidence	21
DNAH1;DNHD1	Sample29	3:52352606:G:A:Htz	11:6568195:C:T:Htz	22.9	23.3	0.0	0.001	Disease-Causing-99.9% confidence	47
DNAH1;DNHD1	Sample424	3:52366791:G:T:Htz	11:6547572:T:A:Htz	26.3	35.0	0.002	0.0	Disease-Causing-99.9% confidence	22
NOTCH2;VWA3A	Sample137	1:120005472:C:A:Htz	16:22152554:G:A:Htz	22.1	25.7	0.009	0.001	Disease-Causing-99% confidence	30
NOTCH2;VWA3A	Sample366	1:120005472:C:A:Htz	16:22121092:A:G:Htz	22.1	31.0	0.009	0.0	Disease-Causing-99% confidence	15
NOTCH2;VWA3A	Sample74	1:119926524:T:C:Htz	16:22149834:C:T:Htz	22.6	17.98	0.006	0.0	Disease-Causing	30
RNF213;ZFHX2	Sample118	17:80374528:G:T:Htz	14:23522108:G:A:Htz	26.8	23.2	0.0	0.0	Disease-Causing	40
RNF213;ZFHX2	Sample425	17:80373040:G:A:Htz	14:23527630:C:T:Htz	23.4	23.8	0.003	0.0	Disease-Causing	45
RNF213;ZFHX2	Sample93	17:80337649:C:T:Htz	14:23522137:C:T:Htz	23.9	24.2	0.0	0.0	Disease-Causing	29
DNAH10;DNAH6	Sample93	12:123848820:G:A:Htz	2:84593994:C:A:Htz	26.7	17.26	0.008	None	Disease-Causing-99.9% confidence	41
DNAH10;DNAH6	Sample167	12:123928562:T:C:Htz	2:84707547:G:A:Htz 2:84517989:C:T:Htz	27.6	25.1 15.45	0.001	0.003 0.007	Disease-Causing-99.9% confidence	23

DNAH10;DNAH6	Sample331	12:123870407:C:T:Htz 12:123875490:C:T:Htz	2:84529097:AT:A:Htz	22.8 20.7	23.1	0.014 0.015	0.0	Disease-Causing- 99.9% confidence	22
DNAH10;DNAH6	Sample217	12:123787868:A:C:Htz	2:84707547:G:A:Htz 2:84517989:C:T:Htz	20.5	25.1 15.45	0.0	0.003 0.007	Disease-Causing- 99.9% confidence	16
DNAH11;DNAH8	Sample383	7:21599943:C:A:Htz 7:21558882:A:G:Htz	6:38973669:G:A:Htz	23.4 22.7	55.0	0.0 0.012	0.0	Disease-Causing- 99.9% confidence	28
DNAH11;DNAH8	Sample234	7:21619115:A:C:Htz	6:38938181:C:T:Htz	19.94	25.7	None	0.007	Disease-Causing- 99.9% confidence	23
DNAH11;DNAH8	Sample310	7:21765459:C:T:Htz	6:38886870:T:C:Htz 6:38823020:G:C:Htz	23.0	26.1 22.4	0.0	0.005 0.0	Disease-Causing- 99.9% confidence	35
DNAH11;DNAH8	Sample353	7:21748602:C:G:Htz	6:38973669:G:A:Htz	20.3	55.0	0.002	0.0	Disease-Causing- 99.9% confidence	42
DNAH11;DNAH8	Sample424	7:21748602:C:G:Htz	6:38852717:C:A:Htz	20.3	23.3	0.002	0.01	Disease-Causing- 99.9% confidence	26
CENPE;DNAH1	Sample100	4:103145116:T:G:Htz	3:52332280:A:G:Htz	5.19	23.3	0.004	0.003	Disease-Causing	26
CENPE;DNAH1	Sample151	4:103180359:T:C:Htz	3:52350578:A:G:Htz 3:52370124:C:T:Htz	8.67	27.8 2.146	0.0	0.004 0.0	Disease-Causing- 99% confidence	18
CENPE;DNAH1	Sample345	4:103159060:A:C:Htz 4:103148884:G:A:Htz	3:52372030:A:G:Htz	19.02 0.136	10.43	0.011 0.011	0.001	Disease-Causing	21
SOHLH1;MCM8	Sample134	9:135697628:C:T:Htz	20:5985980:G:T:Htz	23.2	28.2	0.002	0.0	Disease-Causing- 99% confidence	16
SOHLH1;MCM8	Sample110	9:135697628:C:T:Htz	20:5955247:A:G:Htz	23.2	25.7	0.002	0.001	Disease-Causing- 99% confidence	24
SOHLH1;MCM8	Sample310	9:135698370:G:A:Htz	20:5986117:A:G:Htz	22.2	23.7	None	0.008	Disease-Causing	25
CDC14A;DNAH1	Sample151	1:100499218:G:A:Htz	3:52350578:A:G:Htz 3:52370124:C:T:Htz	25.2	27.8 2.146	0.002	0.004 0.0	Disease-Causing- 99.9% confidence	3
CDC14A;DNAH1	Sample194	1:100499218:G:A:Htz	3:52383885:C:T:Htz	25.2	28.8	0.002	None	Disease-Causing- 99.9% confidence	8

CDC14A;DNAH1	Sample142	1:100499218:G:A:Htz	3:52326302:T:C:Htz 3:52332352:A:G:Htz	25.2	18.18 16.57	0.002	0.004 0.004	Disease-Causing	27
CFAP65;TMEM247	Sample309	2:219013946:T:C:Htz	2:46480456:G:C:Htz	27.2	23.1	0.003	0.003	Disease-Causing-99% confidence	8
CFAP65;TMEM247	Sample355	2:219027915:G:A:Htz	2:46480712:G:A:Htz	23.9	18.19	0.0	0.0	Disease-Causing	48
CFAP65;TMEM247	Sample159	2:219013946:T:C:Htz	2:46480712:G:A:Htz	27.2	18.19	0.003	0.0	Disease-Causing	16
HCN4;ATP2B4	Sample194	15:73322743:G:A:Htz	1:203727420:C:T:Htz 1:203683341:G:A:Htz	22.4	20.8 8.76	0.002	0.008 0.003	Disease-Causing-99.9% confidence	20
HCN4;ATP2B4	Sample407	15:73322516:C:G:Htz	1:203722694:G:C:Htz	20.9	22.4	0.001	0.001	Disease-Causing-99.9% confidence	3
HCN4;ATP2B4	Sample314	15:73322788:C:T:Htz	1:203722694:G:C:Htz	16.64	22.4	0.0	0.001	Disease-Causing	22
FAT4;CELSR3	Sample228	4:125319813:A:T:Htz	3:48645795:G:A:Htz 3:48659715:C:A:Htz	22.7	32.0 22.0	0.001	0.001 0.0	Disease-Causing-99.9% confidence	4
FAT4;CELSR3	Sample222	4:125320069:T:A:Htz	3:48656847:G:A:Htz	23.5	23.1	0.003	None	Disease-Causing-99.9% confidence	33
FAT4;CELSR3	Sample229	4:125468676:C:T:Htz 4:125487377:T:C:Htz	3:48660755:C:T:Htz	23.3 6.69	27.0	0.003 0.004	0.0	Disease-Causing-99.9% confidence	4

F.4.2 Enriched gene pairs with Inheritance Mode filter

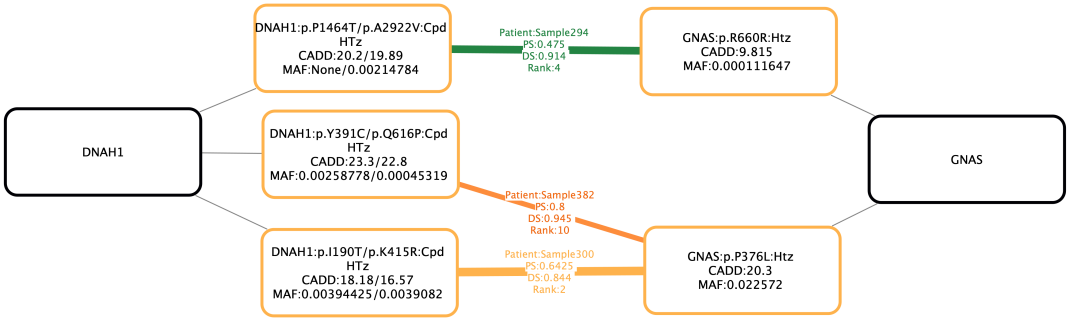


Figure F.20: Variant combinations identified in the gene pair GNAS;DNAH1 in the top 50 of 3 patients with cryptorchidism of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

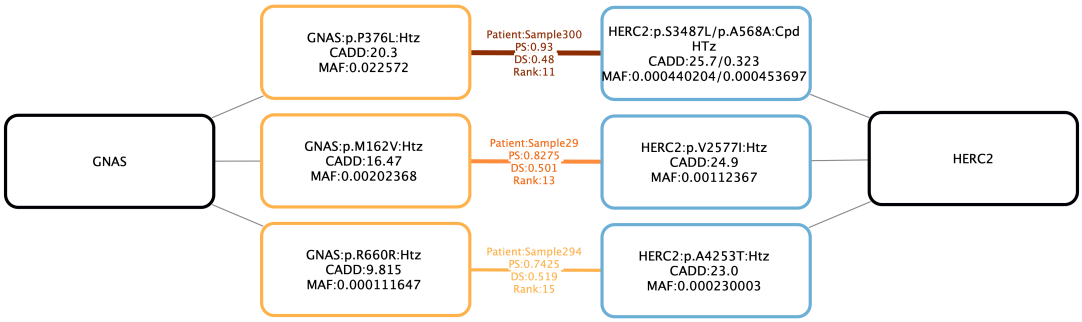


Figure F.21: Variant combinations identified in the gene pair HERC2;GNAS in the top 50 of 3 patients with cryptorchidism of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

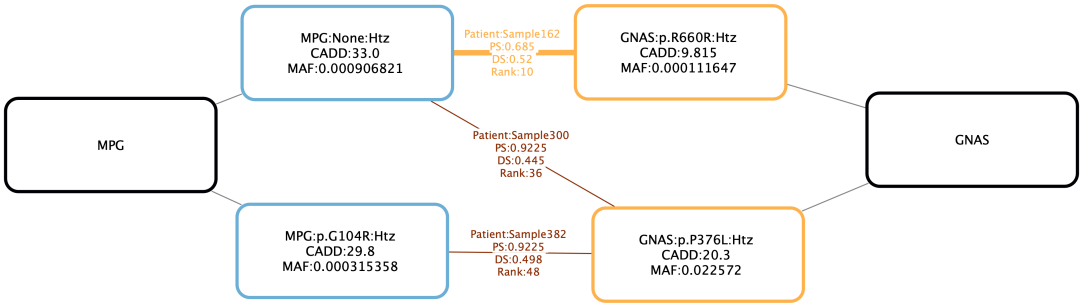


Figure F.22: Variant combinations identified in the gene pair GNAS;MPG in the top 50 of 3 patients with cryptorchidism of the ESTAND cohort. The black nodes represent the genes in which the variants are located. The colored edges represent the predicted variant combinations, with the ID of the patient, the PS and DS. Edges are colored based on the confidence interval of the VarCoPP2.0 score (the darker the color the higher the confidence interval). For each variant the protein change is provided if available, as well as the scaled CADD score and MAF in the GnomADv3.1 database.

Appendix G

Tables of enriched genes within gene pairs

G.1 Patients with NOA

Top 10

Gene	$p-value$	Num. patients carriers	Num. Control carriers	Gene	$p-value$	Num. patients carriers	Num. Control carriers
CBX2	0.001	8	2	ESR1	0.014	5	3
CSMD1	0.03	10	12	SMARCA4	0.017	8	7
ROS1	0.043	13	19	DHX37	0.036	6	6
HS6ST1	0.049	6	6	PTK2B	0.044	5	4

Top 20

Gene	$p-value$	Num. patients carriers	Num. Control carriers	Gene	$p-value$	Num. patients carriers	Num. Control carriers
CBX2	0.002	8	4	ESR1	0.005	6	3
CCDC146	0.006	6	3	SMARCA4	0.022	9	9
CEP250	0.008	12	11	NOTCH2	0.029	17	25
SMARCA4	0.008	7	4	COL7A1	0.041	6	6
				TP53BP1	0.042	6	6
				DHX37	0.047	6	6

Top 50

Gene	$p-value$	Num. patients carriers	Num. Control carriers	Gene	$p-value$	Num. patients carriers	Num. Control carriers
CBX2	0.006	10	10	TNRC18	0.003	10	7
CCDC146	0.008	6	3	ESR1	0.008	6	4
COL7A1	0.008	7	5	PTPRT	0.009	7	4
MYO6	0.008	6	3	NOTCH2	0.02	18	27
PABPC1	0.008	7	4	SMARCA4	0.022	9	9
CEP250	0.009	12	12	COL7A1	0.023	11	14
SMARCA4	0.026	9	9	GORASP1	0.023	7	7

NOTCH2	0.039	15	23	DHX37	0.027	8	7
RAB3GAP2	0.049	6	6	DNAJC13	0.044	6	6
				PABPC1	0.045	8	9
				NOS1	0.046	6	6

Table G.1: Significantly frequent gene pairs in the top 10, top 20 and top 50 ranking of the patients with cryptorchidism.

G.2 Patients with OZ

Top 10

Gene	$p-value$	Num. patients carriers	Num. Control carriers	Gene	$p-value$	Num. patients carriers	Num. Control carriers
GLI2	0.024	11	7	ACTRT1	0.001	7	0
EPHB1	0.049	10	7	TUB	0.004	7	1
				VCX3B	0.004	7	1
				GLI2	0.012	15	10
				PIAS2	0.012	6	1
				ANK3	0.03	8	4
				NOTCH2	0.036	25	24

Top 20

Gene	$p-value$	Num. patients carriers	Num. Control carriers	Gene	$p-value$	Num. patients carriers	Num. Control carriers
ACTRT1	0.001	7	0	ACTRT1	0.001	7	0
PPM1J	0.002	12	5	FAF1	0.002	7	1
NOTCH2	0.024	24	21	VCX3B	0.004	7	1
				TEX13C	0.012	6	1
				TUB	0.013	7	2
				MAGI1	0.015	8	3
				GLI2	0.02	15	10
				PIAS2	0.023	7	3

Top 50

Gene	$p-value$	Num. patients carriers	Num. Control carriers	Gene	$p-value$	Num. patients carriers	Num. Control carriers
ACTRT1	0.001	7	0	ACTRT1	0.001	7	0
MAGI1	0.004	9	2	MGA	0.003	12	5
PPM1J	0.006	12	6	FAF1	0.004	10	2
ASZ1	0.007	6	1	SOX30	0.004	7	1
VCX3B	0.008	7	1	TEX13C	0.006	7	1
TUB	0.009	6	1	VCX3B	0.007	8	2
SORL1	0.021	9	4	ADGRL3	0.009	8	3

GLI2	0.026	13	8	CASZ1	0.011	8	3
DCAF17	0.027	6	2	IQGAP2	0.015	13	8
RIF1	0.027	6	2	PTPRT	0.016	10	4
POLR1B	0.028	6	2	MAGI1	0.018	11	6
NOTCH2	0.031	25	23	PIAS2	0.023	7	3
CHD7	0.042	20	18	NCOR1	0.026	12	8
XRRA1	0.044	7	4	POLR1B	0.028	7	3
SMARCAL1	0.049	6	3	VCX	0.033	8	5
				DNAJC13	0.036	10	6
				HNF1B	0.036	7	3
				PABPC1	0.038	13	9
				SF3B3	0.04	8	5
				CIT	0.042	10	6
				NOTCH2	0.045	27	27

Table G.2: Significantly frequent gene pairs in the top 10, top 20 and top 50 ranking of the patients with cryptorchidism.

G.3 Patients with cryptorchidism

Top 10

Gene	$p-value$	Num. patients carriers	Num. Control carriers	Gene	$p-value$	Num. patients carriers	Num. Control carriers
RNF213	0.016	14	12	BMPR1A	0.011	7	2
SBF1	0.022	11	8	LEO1	0.035	6	3
DNAH1	0.035	31	45	GLI2	0.043	10	10

Top 20

Gene	$p-value$	Num. patients carriers	Num. Control carriers	Gene	$p-value$	Num. patients carriers	Num. Control carriers
TUBA3C	0.018	6	2	BCORL1	0.003	6	1
RNF213	0.024	17	18	PTPRT	0.004	5	0
SBF1	0.026	12	11	BMPR1A	0.011	7	2
CDC14A	0.035	6	3	AP2A2	0.012	5	1
DNAH6	0.048	18	22	FAT4	0.013	5	1
				DGKZ	0.019	5	1
				SNAPC4	0.021	6	2
				LEO1	0.035	6	3
				GLI2	0.043	10	10
				IGSF1	0.045	7	5
				NOTCH2	0.046	20	25

Top 50

Gene	$p-value$	Num. patients carriers	Num. Control carriers	Gene	$p-value$	Num. patients carriers	Num. Control carriers
BCORL1	0.004	6	1	BCORL1	0.003	6	1
FLNB	0.01	7	3	XAB2	0.003	13	7
POTEJ	0.012	11	8	PTPRT	0.004	10	4
KPNA2	0.015	6	2	FAT4	0.008	13	9
IFT74	0.016	5	1	CACNA1B	0.012	7	3
RNF213	0.016	22	25	DNHD1	0.016	8	5
BMPR1A	0.02	7	3	KPNA2	0.016	7	3
MMRN1	0.024	9	7	BMPR1A	0.017	7	3
NOTCH2	0.03	20	23	NCOR1	0.022	10	8
DNAH6	0.032	21	25	PPRC1	0.023	9	6
NCOR1	0.032	10	8	DGKZ	0.034	6	3
GLI2	0.033	10	8	DHX37	0.036	9	7
SLC26A8	0.039	8	6	RBM5	0.041	6	4
TUBA3C	0.04	6	3	COL7A1	0.044	13	14

Table G.3: Significantly frequent gene pairs in the top 10, top 20 and top 50 ranking of the patients with cryptorchidism.

Bibliography

[1] A. A. Durmaz, E. Karaca, U. Demkow, G. Toruner, J. Schoumans, and O. Cogulu, "Evolution of genetic techniques: past, present, and beyond.," *BioMed research international*, vol. 2015, p. 461524, 2015.

[2] J. Gayon, "From Mendel to epigenetics: History of genetics," *Comptes Rendus Biologies*, vol. 339, no. 7, pp. 225–230, 2016.

[3] T. H. Morgan, A. H. Sturtevant, H. J. Muller, and C. B. Bridges, *The mechanism of Mendelian heredity*. H. Holt and Company, 1923.

[4] O. T. Avery, C. M. MacLeod, and M. McCarty, "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii," in *Die Entdeckung der Doppelhelix: Die grundlegenden Arbeiten von Watson, Crick und anderen*, pp. 97–120, Springer, 2017.

[5] G. K. Hunter, "Phoebus levene and the tetranucleotide structure of nucleic acids," *Ambix*, vol. 46, no. 2, pp. 73–103, 1999.

[6] J. D. WATSON and F. H. CRICK, "The structure of DNA," *Cold Spring Harbor symposia on quantitative biology*, vol. 18, pp. 123–131, 1953.

[7] J. Marshall, "The genetic code.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, p. 5760, apr 2014.

[8] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[9] J. Fraser, I. Williamson, W. A. Bickmore, and J. Dostie, "An Overview of Genome Organization and How We Got There: from FISH to Hi-C.," *Microbiology and molecular biology reviews : MMBR*, vol. 79, pp. 347–372, sep 2015.

[10] F. Crick, "Central Dogma of Molecular Biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[11] A. B. Rose, "Introns as gene regulators: A brick on the accelerator," *Frontiers in Genetics*, vol. 10, no. FEB, pp. 1–6, 2019.

[12] S. Pujar, N. A. O'Leary, C. M. Farrell, J. E. Loveland, J. M. Mudge, C. Wallin, *et al.*, "Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation," *Nucleic acids research*, vol. 46, pp. D221–D228, jan 2018.

[13] S. R. Eddy, "Non–coding RNA genes and the modern RNA world ," *Nature Reviews Genetics*, vol. 2, no. 12, pp. 919–929, 2001.

[14] A. Warr, C. Robert, D. Hume, A. Archibald, N. Deeb, and M. Watson, "Exome Sequencing: Current and Future Perspectives.," *G3 (Bethesda, Md.)*, vol. 5, pp. 1543–1550, jul 2015.

[15] D. O. Hessen, *Impact of Noncoding DNA on Phenotype Evolution*, pp. 1–6. John Wiley & Sons, Ltd, 2017.

[16] E. Perenthaler, S. Yousefi, E. Niggl, and T. S. Barakat, "Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development," *Frontiers in Cellular Neuroscience*, vol. 13, no. July, pp. 1–21, 2019.

[17] E. Medico-Salsench, F. Karkala, K. Lanko, and T. S. Barakat, "The non-coding genome in genetic brain disorders: new targets for therapy?," *Essays in biochemistry*, vol. 65, pp. 671–683, oct 2021.

[18] J. M. Heather and B. Chain, "The sequence of sequencers: The history of sequencing DNA.," *Genomics*, vol. 107, pp. 1–8, jan 2016.

[19] C. Cheng, Z. Fei, and P. Xiao, "Methods to improve the accuracy of next-generation sequencing.," *Frontiers in bioengineering and biotechnology*, vol. 11, p. 982111, 2023.

[20] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, pp. 5463–5467, dec 1977.

[21] M. Frank, A. Prenzler, R. Eils, and J.-M. Graf von der Schulenburg, "Genome sequencing: a systematic review of health economic evidence.," *Health economics review*, vol. 3, p. 29, dec 2013.

[22] M. L. Metzker, "Sequencing technologies the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[23] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333–351, 2016.

[24] M. Kircher and J. Kelso, "High-throughput DNA sequencing – concepts and limitations," *BioEssays*, vol. 32, pp. 524–536, jun 2010.

[25] E. E. Schadt, S. Turner, and A. Kasarskis, "A window into third-generation sequencing," *Human Molecular Genetics*, vol. 19, pp. R227–R240, oct 2010.

[26] T. Xiao and W. Zhou, "The third generation sequencing: the advanced approach to genetic diseases.," *Translational pediatrics*, vol. 9, pp. 163–173, apr 2020.

[27] M. C. Lucas and E. M. Novoa, "Long-read sequencing in the era of epigenomics and epitranscriptomics," *nature methods*, vol. 20, no. 1, pp. 25–29, 2023.

[28] Y. Sun, C. A. Ruivenkamp, M. J. Hoffer, T. Vrijenhoek, M. Kriek, C. J. van Asperen, *et al.*, "Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome?," *Human Mutation*, vol. 36, pp. 648–655, jun 2015.

[29] D. C. Koboldt, "Best practices for variant calling in clinical sequencing," *Genome Medicine*, vol. 12, no. 1, p. 91, 2020.

[30] D. M. Church, V. A. Schneider, K. M. Steinberg, M. C. Schatz, A. R. Quinlan, C.-S. Chin, *et al.*, "Extending reference assembly models.," *Genome biology*, vol. 16, p. 13, jan 2015.

[31] V. A. Schneider, T. Graves-Lindsay, K. Howe, N. Bouk, H.-C. Chen, P. A. Kitts, *et al.*, "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly.," *Genome research*, vol. 27, pp. 849–864, may 2017.

[32] S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, *et al.*, "The complete sequence of a human genome," *Science*, vol. 376, pp. 44–53, apr 2022.

[33] W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, *et al.*, "A draft human pangenome reference," *Nature*, vol. 617, no. 7960, pp. 312–324, 2023.

[34] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol, "Human genetic variation and its contribution to complex traits," *Nature Reviews Genetics*, vol. 10, no. 4, pp. 241–251, 2009.

[35] J. Huddleston and E. E. Eichler, "An incomplete understanding of human genetic variation," *Genetics*, vol. 202, no. 4, pp. 1251–1254, 2016.

[36] . G. P. Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, *et al.*, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, 10 2010.

[37] A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, *et al.*, "A global reference for human genetic variation.," *Nature*, vol. 526, pp. 68–74, 10 2015.

[38] P. L. Auer, A. P. Reiner, G. Wang, H. M. Kang, G. R. Abecasis, D. Altshuler, *et al.*, "Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project," *American journal of human genetics*, vol. 99, pp. 791–801, 10 2016.

[39] A. F. Wright, "Genetic Variation: Polymorphisms and Mutations," *Encyclopedia of Life Sciences*, pp. 1–10, 2005.

[40] M. Jackson, L. Marks, G. H. W. May, and J. B. Wilson, "The genetic basis of disease.," *Essays in biochemistry*, vol. 62, pp. 643–723, dec 2018.

[41] C. Alkan, B. P. Coe, and E. E. Eichler, "Genome structural variation discovery and genotyping.," *Nature reviews. Genetics*, vol. 12, pp. 363–376, may 2011.

[42] P. Balachandran and C. R. Beck, "Structural variant identification and characterization," *Chromosome Research*, vol. 28, no. 1, pp. 31–47, 2020.

[43] J. P. Venables, "Downstream intronic splicing enhancers," *FEBS letters*, vol. 581, p. 4127—4131, 9 2007.

[44] Z. Zeng and Y. Bromberg, "Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives," *Frontiers in Genetics*, vol. 10, p. 914, 2019.

[45] J. T. den Dunnen, R. Dalgleish, D. R. Maglott, R. K. Hart, M. S. Greenblatt, J. McGowan-Jordan, *et al.*, "HGVS Recommendations for the Description of Sequence Variants: 2016 Update," *Human Mutation*, vol. 37, pp. 564–569, jun 2016.

[46] P. J. Freeman, R. K. Hart, L. J. Gretton, A. J. Brookes, and R. Dalgleish, "VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions," *Human Mutation*, vol. 39, pp. 61–68, jan 2018.

[47] M. Wang, K. M. Callenberg, R. Dalgleish, A. Fedtsov, N. K. Fox, P. J. Freeman, *et al.*, "hgvs: A Python package for manipulating sequence variants using HGVS nomenclature: 2018 Update.," *Human mutation*, vol. 39, pp. 1803–1813, dec 2018.

[48] M. Wildeman, E. van Ophuizen, J. T. den Dunnen, and P. E. M. Taschner, "Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker.," *Human mutation*, vol. 29, pp. 6–13, jan 2008.

[49] M. Haendel, N. Vasilevsky, D. Unni, C. Bologa, N. Harris, H. Rehm, *et al.*, "How many rare diseases are there?," 2 2020.

[50] C. I. E. Smith, P. Bergman, and D. W. Hagey, "Estimating the number of diseases - the concept of rare, ultra-rare, and hyper-rare.," *iScience*, vol. 25, p. 104698, aug 2022.

[51] E. Union, "Regulation (EC) Nˆ141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products.," 2000.

[52] S. Nguengang Wakap, D. M. Lambert, A. Olry, C. Rodwell, C. Gueydan, V. Lanneau, *et al.*, "Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database," *European Journal of Human Genetics*, vol. 28, no. 2, pp. 165–173, 2020.

[53] A. Schieppati, J.-I. Henter, E. Daina, and A. Aperia, "Why rare diseases are an important medical and social issue," *The Lancet*, vol. 371, pp. 2039–2041, 6 2008.

[54] J. K. Stoller, "The Challenge of Rare Diseases," *CHEST*, vol. 153, pp. 1309–1314, jun 2018.

[55] A. A. Mitani and S. Haneuse, "Small Data Challenges of Studying Rare Diseases," *JAMA Network Open*, vol. 3, pp. e201965–e201965, mar 2020.

[56] Eurordis, "Rare Diseases: Understanding this Public Health Priority," tech. rep., Eurordis, 2005.

[57] D. Julkowska, C. P. Austin, C. M. Cutillo, D. Gancberg, C. Hager, J. Halftermeyer, *et al.*, "The importance of international collaboration for rare diseases research: a European perspective.," *Gene therapy*, vol. 24, pp. 562–571, sep 2017.

[58] P. Sernadela, L. González-Castro, C. Carta, E. van der Horst, P. Lopes, R. Kaliyaperumal, *et al.*, "Linked Registries: Connecting Rare Diseases Patient Registries through a Semantic Web Layer.," *BioMed research international*, vol. 2017, p. 8327980, 2017.

[59] C. A. Brownstein, I. A. Holm, R. Ramoni, and D. B. Goldstein, "Data sharing in the undiagnosed diseases network.," *Human mutation*, vol. 36, pp. 985–988, oct 2015.

[60] C. Montano, T. Cassini, S. G. Ziegler, M. Boehm, E.-R. Nicoli, J. A. Mindell, *et al.*, "Diagnosis and discovery: Insights from the NIH Undiagnosed Diseases Program," *Journal of Inherited Metabolic Disease*, vol. 45, pp. 907–918, sep 2022.

[61] T. E. Consortium, "Sub-genic intolerance, ClinVar, and the epilepsies: A whole-exome sequencing study of 29,165 individuals.," *American journal of human genetics*, vol. 108, pp. 965–982, jun 2021.

[62] D. B. Callaghan, S. Rogic, P. P. C. Tan, K. Calli, Y. Qiao, R. Baldwin, *et al.*, "Whole genome sequencing and variant discovery in the ASPIRE autism spectrum disorder cohort," *Clinical Genetics*, vol. 96, pp. 199–206, sep 2019.

[63] H. V. Firth and C. F. Wright, "The Deciphering Developmental Disorders (DDD) study.," *Developmental medicine and child neurology*, vol. 53, pp. 702–703, aug 2011.

[64] A. Rath, A. Olry, F. Dhombres, M. M. Brandt, B. Urbero, and S. Ayme, "Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users.," *Human mutation*, vol. 33, pp. 803–808, may 2012.

[65] K. J. Karczewski and M. P. Snyder, "Integrative omics for health and disease.," *Nature reviews. Genetics*, vol. 19, pp. 299–310, may 2018.

[66] S. Köhler, S. C. Doelken, C. J. Mungall, S. Bauer, H. V. Firth, I. Bailleul-Forestier, *et al.*, "The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data," *Nucleic Acids Research*, vol. 42, pp. D966–D974, jan 2014.

[67] S. Köhler, L. Carmody, N. Vasilevsky, and E. Al, "Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources," *Nucleic Acids Research*, vol. 47, pp. D1018–D1027, 11 2018.

[68] S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, *et al.*, "Clinical diagnostics in human genetics with semantic similarity searches in ontologies.," *American journal of human genetics*, vol. 85, pp. 457–464, oct 2009.

[69] T. Zemojtel, S. Köhler, L. Mackenroth, M. Jäger, J. Hecht, P. Krawitz, *et al.*, "Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome.," *Science translational medicine*, vol. 6, p. 252ra123, sep 2014.

[70] O. J. Buske, M. Girdea, S. Dumitriu, B. Gallinger, T. Hartley, H. Trang, *et al.*, "PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases.," *Human mutation*, vol. 36, pp. 931–940, oct 2015.

[71] S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, *et al.*, "The Human Phenotype Ontology in 2021," *Nucleic acids research*, vol. 49, pp. D1207–D1217, 1 2021.

[72] C. M. Delude, "Deep phenotyping: The details of disease," *Nature*, vol. 527, no. 7576, pp. S14–S15, 2015.

[73] J. T. Wright and M. C. Herzberg, "Science for the Next Century: Deep Phenotyping.," *Journal of dental research*, vol. 100, pp. 785–789, jul 2021.

[74] P. M. Visscher, L. Yengo, N. J. Cox, and N. R. Wray, "Discovery and implications of polygenicity of common diseases.," *Science (New York, N.Y.)*, vol. 373, pp. 1468–1473, sep 2021.

[75] J. A. Veltman and H. G. Brunner, "De novo mutations in human genetic disease.," *Nature reviews. Genetics*, vol. 13, pp. 565–575, 7 2012.

[76] R. A. Veitia, S. Caburet, and J. A. Birchler, "Mechanisms of Mendelian dominance," *Clinical Genetics*, vol. 93, pp. 419–428, mar 2018.

[77] S. Chen and G. Parmigiani, "Meta-analysis of BRCA1 and BRCA2 penetrance.," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 25, pp. 1329–1333, apr 2007.

[78] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, *et al.*, "Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology," *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 17, p. 405, 5 2015.

[79] M. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, "Online Mendelian Inheritance in Man, OMIM.," 2022.

[80] A. R. Martin, E. Williams, R. E. Foulger, S. Leigh, L. C. Daugherty, O. Niblock, *et al.*, "PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels," *Nature genetics*, 2019.

[81] Z. Stark, R. E. Foulger, E. Williams, B. A. Thompson, C. Patel, S. Lunke, *et al.*, "Scaling national and international improvement in virtual gene panel curation via a collaborative approach to discordance resolution," *The American Journal of Human Genetics*, vol. 108, no. 9, pp. 1551–1557, 2021.

[82] M. J. Landrum and B. L. Kattman, "ClinVar at five years: Delivering on the promise.," *Human mutation*, vol. 39, pp. 1623–1630, nov 2018.

[83] M. J. Landrum, S. Chitipiralla, G. R. Brown, C. Chen, B. Gu, J. Hart, *et al.*, "ClinVar: improvements to accessing data.," *Nucleic acids research*, vol. 48, pp. D835–D844, 1 2020.

[84] J. L. Badano and N. Katsanis, "Beyond Mendel: an evolving view of human genetic disease transmission," *Nature Reviews Genetics*, vol. 3, no. 10, pp. 779–789, 2002.

[85] M. A. Kennedy, "Mendelian Genetic Disorders," 9 2005.

[86] K. J. Mitchell, "What is complex about complex disorders?," *Genome biology*, vol. 13, p. 237, 1 2012.

[87] V. van Heyningen and P. L. Yeyati, "Mechanisms of non-Mendelian inheritance in genetic disease," *Human Molecular Genetics*, vol. 13, no. REV. ISS. 2, pp. 225–233, 2004.

[88] C. Deltas, "Digenic inheritance and genetic modifiers.," *Clinical genetics*, 2017.

[89] B. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, *et al.*, "Identification of the cystic fibrosis gene: genetic analysis.," *Science (New York, N.Y.)*, vol. 245, pp. 1073–1080, 9 1989.

[90] G. R. Cutting, "Modifier genetics: cystic fibrosis.," *Annual review of genomics and human genetics*, vol. 6, pp. 237–260, 2005.

[91] M. L. Drumm, A. G. Ziady, and P. B. Davis, "Genetic variation and clinical heterogeneity in cystic fibrosis.," *Annual review of pathology*, vol. 7, pp. 267–282, 2012.

[92] J. F. Robinson and N. Katsanis, "Oligogenic Disease," in *Vogel and Motulsky's Human Genetics: Problems and Approaches* (M.R. Speicher et al., ed.), ch. 7, pp. 243–262, Springer-Verlag Berlin Heiderlberg, 4th editio ed., 2010.

[93] N. Katsanis, "The continuum of causality in human genetic disorders," *Genome Biology*, vol. 17, no. 1, pp. 233–237, 2016.

[94] J. R. Giudicessi, A. A. M. Wilde, and M. J. Ackerman, "The genetic architecture of long QT syndrome: A critical reappraisal," *Trends in cardiovascular medicine*, vol. 28, pp. 453–464, 10 2018.

[95] S. E. Antonarakis and J. S. Beckmann, "Mendelian disorders deserve more attention," *Nature Reviews Genetics*, vol. 7, no. 4, pp. 277–282, 2006.

[96] A. A. Schaffer, "Digenic inheritance in medical genetics," *Journal of Medical Genetics*, vol. 50, no. 10, pp. 641–652, 2013.

[97] N. Katsanis, S. J. Ansley, J. L. Badano, E. R. Eichers, R. A. Lewis, B. E. Hoskins, *et al.*, "Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder.," *Science (New York, N.Y.)*, vol. 293, pp. 2256–2259, 9 2001.

[98] P. L. Beales, J. L. Badano, A. J. Ross, S. J. Ansley, B. E. Hoskins, B. Kirsten, *et al.*, "Genetic interaction of BBS1 mutations with alleles at other BBS loci can result in non-Mendelian Bardet-Biedl syndrome.," *American journal of human genetics*, vol. 72, pp. 1187–1199, 5 2003.

[99] O. M'hamdi, I. Ouertani, and H. Chaabouni-Bouhamed, "Update on the genetics of bardet-biedl syndrome.," *Molecular syndromology*, vol. 5, pp. 51–56, 2 2014.

[100] M. Cerrone, C. A. Remme, R. Tadros, C. R. Bezzina, and M. Delmar, "Beyond the One Gene–One Disease Paradigm," *Circulation*, vol. 140, pp. 595–610, 8 2019.

[101] R. A. Fisher, "Xv.—the correlation between relatives on the supposition of mendelian inheritance.," *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, vol. 52, no. 2, pp. 399–433, 1919.

[102] B. Lehner, "Molecular mechanisms of epistasis within and between genes," *Trends in Genetics*, vol. 27, no. 8, pp. 323–331, 2011.

[103] J. Domingo, P. Baeza-Centurion, and B. Lehner, "The Causes and Consequences of Genetic Interactions (Epistasis).," *Annual review of genomics and human genetics*, vol. 20, pp. 433–460, aug 2019.

[104] A. M. Gazzo, D. Daneels, E. Cilia, M. Bonduelle, M. Abramowicz, S. Van Dooren, *et al.*, "DIDA: A curated and annotated digenic diseases database," *Nucleic Acids Research*, vol. 44, no. D1, pp. D900–D907, 2015.

[105] S. Papadimitriou, A. Gazzo, N. Versbraegen, C. Nachtegael, J. Aerts, Y. Moreau, *et al.*, "Predicting disease-causing variant combinations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 116, no. 24, pp. 11878–11887, 2019.

[106] S. Marwaha, J. W. Knowles, and E. A. Ashley, "A guide for the diagnosis of rare and undiagnosed disease: beyond the exome," *Genome Medicine*, vol. 14, pp. 1–22, 12 2022.

[107] S. L. Sawyer, T. Hartley, D. A. Dyment, C. L. Beaulieu, J. Schwartzentruber, A. Smith, *et al.*, "Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care," *Clinical genetics*, vol. 89, pp. 275–284, 3 2016.

[108] A. H. Graungaard and L. Skov, "Why do we need a diagnosis? A qualitative study of parents' experiences, coping and needs, when the newborn child is severely disabled.," *Child: care, health and development*, vol. 33, pp. 296–307, 5 2007.

[109] V. Shashi, A. McConkie-Rosell, B. Rosell, K. Schoch, K. Vellore, M. McDonald, *et al.*, "The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders.," *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 16, pp. 176–182, 2 2014.

[110] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, *et al.*, "Exome sequencing identifies the cause of a mendelian disorder.," *Nature genetics*, vol. 42, pp. 30–35, 1 2010.

[111] Y. Yang, D. M. Muzny, J. G. Reid, M. N. Bainbridge, A. Willis, P. A. Ward, *et al.*, "Clinical whole-exome sequencing for the diagnosis of mendelian disorders.," *The New England journal of medicine*, vol. 369, pp. 1502–1511, 10 2013.

[112] M. N. Bainbridge, W. Wiszniewski, D. R. Murdock, J. Friedman, C. Gonzaga-Jauregui, I. Newsham, *et al.*, "Whole-genome sequencing for optimized patient management.," *Science translational medicine*, vol. 3, p. 87re3, 6 2011.

[113] X. Zhu, S. Petrovski, P. Xie, E. K. Ruzzo, Y.-F. Lu, K. M. McSweeney, *et al.*, "Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios.," *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 17, pp. 774–781, 10 2015.

[114] K. Splinter, D. R. Adams, C. A. Bacino, H. J. Bellen, J. A. Bernstein, A. M. Cheatle-Jarvela, *et al.*, "Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease," *New England Journal of Medicine*, vol. 379, pp. 2131–2139, 10 2018.

[115] N. Malod-Dognin, J. Petschnigg, and N. Pržulj, "Precision medicine - A promising, yet challenging road lies ahead," *Current Opinion in Systems Biology*, vol. 7, pp. 1–7, 2018.

[116] K. Divaris, "Fundamentals of Precision Medicine.," *Compendium of continuing education in dentistry (Jamesburg, N.J. : 1995)*, vol. 38, no. 8 Suppl, pp. 30–32, 2017.

[117] E. A. Ashley, "Towards precision medicine," *Nature Reviews Genetics*, vol. 17, no. 9, pp. 507–522, 2016.

[118] L. G. Biesecker, "Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: lessons from the ClinSeq project," *Genetics in Medicine*, vol. 14, no. 4, pp. 393–398, 2012.

[119] A. M. McInerney-Leo and E. L. Duncan, "Massively Parallel Sequencing for Rare Genetic Disorders: Potential and Pitfalls.," *Frontiers in endocrinology*, vol. 11, p. 628946, 2020.

[120] C. Di Resta, S. Galbiati, P. Carrera, and M. Ferrari, "Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities," *EJIFCC*, vol. 29, p. 4, 4 2018.

[121] D. G. MacArthur, T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, *et al.*, "Guidelines for investigating causality of sequence variants in human disease," *Nature 2014 508:7497*, vol. 508, pp. 469–476, 4 2014.

[122] R. R. Fabsitz, A. McGuire, R. R. Sharp, M. Puggal, L. M. Beskow, L. G. Biesecker, *et al.*, "Ethical and Practical Guidelines for Reporting Genetic Research Results to Study Participants," *Circulation: Cardiovascular Genetics*, vol. 3, pp. 574–580, 12 2010.

[123] Y. Moreau and L. C. Tranchevent, "Computational tools for prioritizing candidate genes: Boosting disease gene discovery," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 523–536, 2012.

[124] O. Zolotareva and M. Kleine, "A Survey of Gene Prioritization Tools for Mendelian and Complex Human Diseases," *Journal of integrative bioinformatics*, vol. 16, no. 4, pp. 1–20, 2019.

[125] X. Yuan, J. Wang, B. Dai, Y. Sun, K. Zhang, F. Chen, *et al.*, "Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases," *Briefings in Bioinformatics*, vol. 23, 3 2022.

[126] G. M. Cooper and J. Shendure, "Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.," *Nature reviews. Genetics*, vol. 12, pp. 628–640, 8 2011.

[127] D. S. W. Ho, W. Schierding, M. Wake, R. Saffery, and J. O'Sullivan, "Machine learning SNP based prediction for precision medicine," *Frontiers in Genetics*, vol. 10, no. MAR, pp. 1–10, 2019.

[128] M. Kumaran, U. Subramanian, and B. Devarajan, "Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data," *BMC Bioinformatics*, vol. 20, no. 1, p. 342, 2019.

[129] H. Ye, J. Meehan, W. Tong, and H. Hong, "Alignment of Short Reads: A Crucial Step for Application of Next-Generation Sequencing Data in Precision Medicine," 2015.

[130] M. Alser, J. Rotman, D. Deshpande, K. Taraszka, H. Shi, P. I. Baykal, *et al.*, "Technology dictates algorithms: recent developments in read alignment," *Genome Biology*, vol. 22, no. 1, p. 249, 2021.

[131] J. Casper, A. S. Zweig, C. Villarreal, C. Tyner, M. L. Speir, K. R. Rosenbloom, *et al.*, "The UCSC Genome Browser database: 2018 update," *Nucleic Acids Research*, vol. 46, pp. D762–D769, jan 2018.

[132] H. Li, M. Dawood, M. M. Khayat, J. R. Farek, S. N. Jhangiani, Z. M. Khan, *et al.*, "Exome variant discrepancies due to reference-genome differences.," *American journal of human genetics*, vol. 108, pp. 1239–1250, jul 2021.

[133] B. Pan, R. Kusko, W. Xiao, Y. Zheng, Z. Liu, C. Xiao, *et al.*, "Similarities and differences between variants called with human reference genome HG19 or HG38," *BMC Bioinformatics*, vol. 20, no. 2, p. 101, 2019.

[134] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1754–1760, jul 2009.

[135] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2.," *Nature methods*, vol. 9, pp. 357–359, mar 2012.

[136] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype," *Nature Biotechnology*, vol. 37, no. 8, pp. 907–915, 2019.

[137] C. Trapnell and S. L. Salzberg, "How to map billions of short reads onto genomes," *Nature Biotechnology*, vol. 27, no. 5, pp. 455–457, 2009.

[138] X. Yu, K. Guda, J. Willis, M. Veigl, Z. Wang, S. Markowitz, *et al.*, "How do alignment programs perform on sequencing data with varying qualities and from repetitive regions?," *BioData mining*, vol. 5, pp. 1–12, 2012.

[139] V. Phan, S. Gao, Q. Tran, and N. S. Vo, "How genome complexity can explain the difficulty of aligning reads to genomes.," *BMC bioinformatics*, vol. 16 Suppl 17, no. Suppl 17, p. S3, 2015.

[140] T. Günther and C. Nettelblad, "The presence and impact of reference bias on population genomic studies of prehistoric human populations," *PLoS genetics*, vol. 15, no. 7, p. e1008302, 2019.

[141] N.-C. Chen, B. Solomon, T. Mun, S. Iyer, and B. Langmead, "Reference flow: reducing reference bias using multiple population genomes.," *Genome biology*, vol. 22, p. 8, jan 2021.

[142] M.-J. Lin, S. Iyer, N.-C. Chen, and B. Langmead, "Measuring, visualizing and diagnosing reference bias with biastools," *bioRxiv*, 2024.

[143] V. Narasimhan, P. Danecek, A. Scally, Y. Xue, C. Tyler-Smith, and R. Durbin, "BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data," *Bioinformatics*, vol. 32, pp. 1749–1751, jun 2016.

[144] E. Garrison and G. Marth, "Haplotype-based variant detection from short-read sequencing," 2012.

[145] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, *et al.*, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.," *Genome research*, vol. 20, pp. 1297–1303, sep 2010.

[146] A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg, A. O. M. Wilkie, *et al.*, "Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications.," *Nature genetics*, vol. 46, pp. 912–918, aug 2014.

[147] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, *et al.*, "Machine learning in bioinformatics," *Briefings in Bioinformatics*, vol. 7, pp. 86–112, 3 2006.

[148] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. 8, 2012.

[149] J. Banerjee, J. N. Taroni, R. J. Allaway, D. V. Prasad, J. Guinney, and C. Greene, "Machine learning in rare disease," *Nature Methods*, vol. 20, no. 6, pp. 803–814, 2023.

[150] S. J. MacEachern and N. D. Forkert, "Machine learning for precision medicine," *Genome*, vol. 64, pp. 416–425, oct 2020.

[151] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 3, pp. 1–21, 2021.

[152] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[153] Z. C. Lipton, "The Mythos of Model Interpretability," *Queue*, vol. 16, p. 31–57, 6 2018.

[154] C. B. Azodi, J. Tang, and S. H. Shiu, "Opening the Black Box: Interpretable Machine Learning for Geneticists," *Trends in Genetics*, vol. 36, no. 6, pp. 442–455, 2020.

[155] J. Mitros and B. Mac Namee, "A Categorisation of Post-hoc Explanations for Predictive Models," *arXiv*, 4 2019.

[156] M. A. Ahmad, C. Eckert, A. Teredesai, G. Mckelvey, A. Teredsai, and G. Mckelvey, "Interpretable Machine Learning in Healthcare," *IEEE Intelligent Informatics Bulletin*, vol. 19, no. 1, pp. 1–7, 2018.

[157] R. Elshawi, M. H. Al-Mallah, and S. Sakr, "On the interpretability of machine learning-based model for predicting hypertension," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, 2019.

[158] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function.," *Nucleic acids research*, vol. 31, pp. 3812–3814, 7 2003.

[159] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, "Predicting functional effect of human missense mutations using PolyPhen-2," *Current protocols in human genetics*, vol. Chapter 7, pp. Unit7.20–Unit7.20, 1 2013.

[160] K. Eilbeck, A. Quinlan, and M. Yandell, "Settling the score: Variant prioritization and Mendelian disease," *Nature Reviews Genetics*, vol. 18, no. 10, pp. 599–612, 2017.

[161] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Research*, vol. 47, pp. D886–D894, 10 2018.

[162] N. M. Ioannidis, J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, *et al.*, "REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants," *The American Journal of Human Genetics*, vol. 99, no. 4, pp. 877–885, 2016.

[163] D. G. Grimm, C.-A. Azencott, F. Aicheler, U. Gieraths, D. G. MacArthur, K. E. Samocha, *et al.*, "The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity," *Human mutation*, vol. 36, no. 5, pp. 513–523, 2015.

[164] J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, *et al.*, "Disease variant prediction with deep generative models of evolutionary data," *Nature*, vol. 599, no. 7883, pp. 91–95, 2021.

[165] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[166] J. Cheng, G. Novati, J. Pan, C. Bycroft, A. Žemgulytė, T. Applebaum, *et al.*, "Accurate proteome-wide missense variant effect prediction with AlphaMissense," *Science*, vol. 381, p. eadg7492, may 2024.

[167] S. Mukherjee, J. D. Cogan, J. H. Newman, J. A. Phillips III, R. Hamid, U. D. Network, *et al.*, "Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network," *The American Journal of Human Genetics*, vol. 108, no. 10, pp. 1946–1963, 2021.

[168] R. M. Raj and A. Sreeja, "Analysis of computational gene prioritization approaches," *Procedia Computer Science*, vol. 143, pp. 395–410, 2018.

[169] D. Smedley, J. O. Jacobsen, M. Jäger, S. Köhler, M. Holtgrewe, M. Schubach, *et al.*, "Next-generation diagnostics and disease-gene discovery with the Exomiser," *Nature Protocols*, vol. 10, no. 12, pp. 2004–2015, 2015.

[170] P. N. Robinson, S. Köhler, A. Oellrich, S. M. Genetics, K. Wang, C. J. Mungall, *et al.*, "Improved exome prioritization of disease genes through cross-species phenotype comparison," *Genome research*, vol. 24, pp. 340–348, 2 2014.

[171] M. V. Singleton, S. L. Guthery, K. V. Voelkerding, K. Chen, B. Kennedy, R. L. Margraf, *et al.*, "Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families.," *American journal of human genetics*, vol. 94, pp. 599–610, 4 2014.

[172] A. Sifrim, D. Popovic, L.-C. Tranchevent, A. Ardeshirdavani, R. Sakai, P. Konings, *et al.*, "eXtasy: variant prioritization by genomic data fusion.," *Nature methods*, vol. 10, pp. 1083–1084, 11 2013.

[173] C. Wu, B. Devkota, P. Evans, X. Zhao, S. W. Baker, R. Niazi, *et al.*, "Rapid and accurate interpretation of clinical exomes using Phenoxome: a computational phenotype-driven approach," *European Journal of Human Genetics*, vol. 27, no. 4, pp. 612–620, 2019.

[174] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the Interactome for Prioritization of Candidate Disease Genes," *The American Journal of Human Genetics*, vol. 82, pp. 949–958, 4 2008.

[175] I. Boudellioua, R. B. Mahamad Razali, M. Kulmanov, Y. Hashish, V. B. Bajic, E. Goncalves-Serra, *et al.*, "Semantic prioritization of novel causative genomic variants," *PLoS Computational Biology*, vol. 13, no. 4, pp. 1–21, 2017.

[176] I. Boudellioua, M. Kulmanov, P. N. Schofield, G. V. Gkoutos, and R. Hoehndorf, "DeepPVP: Phenotype-based prioritization of causative variants using deep learning," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–8, 2019.

[177] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "PhenomeNET: a whole-phenome approach to disease gene discovery.," *Nucleic acids research*, vol. 39, p. e119, 10 2011.

[178] H. Zhang, A. Ferguson, G. Robertson, M. Jiang, T. Zhang, C. Sudlow, *et al.*, "Benchmarking network-based gene prioritization methods for cerebral small vessel disease," *Briefings in Bioinformatics*, vol. 22, p. bbab006, sep 2021.

[179] P. Buphamalai, T. Kokotovic, V. Nagy, and J. Menche, "Network analysis reveals rare disease signatures across multiple levels of biological organization.," *Nature communications*, vol. 12, p. 6306, nov 2021.

[180] M. Agrawal, M. Zitnik, and J. Leskovec, "Large-scale analysis of disease pathways in the human interactome," *Pacific Symposium on Biocomputing*, vol. 0, no. 212669, pp. 111–122, 2018.

[181] J. Peng, K. Bai, X. Shang, G. Wang, H. Xue, S. Jin, *et al.*, "Predicting disease-related genes using integrated biomedical networks," *BMC genomics*, vol. 18, 1 2017.

[182] Y. Zhang, J. Liu, X. Liu, X. Fan, Y. Hong, Y. Wang, *et al.*, "Prioritizing disease genes with an improved dual label propagation framework," *BMC bioinformatics*, vol. 19, 2 2018.

[183] A. Lysenko, K. A. Boroevich, and T. Tsunoda, "Arete - candidate gene prioritization using biological network topology with additional evidence types," *BioData mining*, vol. 10, 7 2017.

[184] J. Birgmeier, M. Haeussler, C. A. Deisseroth, E. H. Steinberg, K. A. Jagadeesh, A. J. Ratner, *et al.*, "AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature," *Science Translational Medicine*, vol. 12, no. 544, 2020.

[185] P. N. Robinson, V. Ravanmehr, J. O. Jacobsen, D. Danis, X. A. Zhang, L. C. Carmody, *et al.*, "Interpretable Clinical Genomics with a Likelihood Ratio Paradigm," *American Journal of Human Genetics*, vol. 107, no. 3, pp. 403–417, 2020.

[186] Q. Li, K. Zhao, C. D. Bustamante, X. Ma, and W. H. Wong, "Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis," *Genetics in Medicine*, vol. 21, no. 9, pp. 2126–2134, 2019.

[187] N. Versbraegen, A. Fouché, C. Nachtegael, S. Papadimitriou, A. Gazzo, G. Smits, *et al.*, "Using game theory and decision decomposition to effectively discern and characterise bi-locus diseases," *Artificial Intelligence in Medicine*, vol. 99, no. June, p. 101690, 2019.

[188] A. Renaux, S. Papadimitriou, N. Versbraegen, C. Nachtegael, S. Boutry, A. Nowé, *et al.*, "ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations," *Nucleic Acids Research*, vol. 47, no. W1, pp. W93–W98, 2019.

[189] A. Costantini, H. Valta, A. M. Suomi, O. Mäkitie, and F. Taylan, "Oligogenic Inheritance of Monoallelic TRIP11, FKBP10, NEK1, TBX5, and NBAS Variants Leading to a Phenotype Similar to Odontochondrodysplasia," *Frontiers in Genetics*, vol. 12, no. June, 2021.

[190] M. Laan, L. Kasak, K. Timinskas, M. Grigorova, Č. Venclovas, A. Renaux, *et al.*, "NR5A1 c.991-1G > C splice-site variant causes familial 46,XY partial gonadal dysgenesis with incomplete penetrance.," *Clinical endocrinology*, vol. 94, pp. 656–666, apr 2021.

[191] M. Brancaccio, C. Mennitti, A. Cesaro, E. Monda, V. D'Argenio, G. Casaburi, *et al.*, "Multidisciplinary In-Depth Investigation in a Young Athlete Suffering from Syncope Caused by Myocardial Bridge.," *Diagnostics (Basel, Switzerland)*, vol. 11, nov 2021.

[192] H. Dallali, N. Kherijid, W. Kammoun, M. Mrad, M. Soltani, H. Trabelsi, *et al.*, "Multiallelic rare variants in BBS genes support an oligogenic ciliopathy in a non-obese juvenile-onset syndromic diabetic patient: a case report," *Frontiers in Genetics*, 2021.

[193] I. Martinez de Lapiscina, C. Kouri, J. Aurrekoetxea, M. Sanchez, R. Naamneh Elzenaty, K.-S. Sauter, *et al.*, "Genetic reanalysis of patients with a difference of sex development carrying the NR5A1/SF-1 variant p.Gly146Ala has discovered other likely disease-causing variations.," *PloS one*, vol. 18, no. 7, p. e0287515, 2023.

[194] I. Boudellioua, M. Kulmanov, P. N. Schofield, G. V. Gkoutos, and R. Hoehndorf, "OligoPVP: Phenotype-driven analysis of individual genomic information to prioritize oligogenic disease variants," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018.

[195] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, *et al.*, "Genome-wide association studies," *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 59, 2021.

[196] S. Lee, G. R. Abecasis, M. Boehnke, and X. Lin, "Rare-variant association analysis: study designs and statistical tests.," *American journal of human genetics*, vol. 95, pp. 5–23, jul 2014.

[197] P. L. Auer and G. Lettre, "Rare variant association studies: considerations, challenges and opportunities.," *Genome medicine*, vol. 7, no. 1, p. 16, 2015.

[198] D. L. Nicolae, "Association Tests for Rare Variants.," *Annual review of genomics and human genetics*, vol. 17, pp. 117–130, aug 2016.

[199] F. Rajabli and B. W. Kunkle, "Strategies in Aggregation Tests for Rare Variants," *Current Protocols*, vol. 3, p. e931, nov 2023.

[200] B. Li and S. M. Leal, "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.," *American journal of human genetics*, vol. 83, pp. 311–321, sep 2008.

[201] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test.," *American journal of human genetics*, vol. 89, pp. 82–93, jul 2011.

[202] S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, *et al.*, "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies.," *American journal of human genetics*, vol. 91, pp. 224–237, aug 2012.

[203] R. Fan and S.-H. Lo, "A robust model-free approach for rare variants association studies incorporating gene-gene and gene-environmental interactions.," *PloS one*, vol. 8, no. 12, p. e83057, 2013.

[204] J. Zhao, Y. Zhu, and M. Xiong, "Genome-wide gene-gene interaction analysis for next-generation sequencing.," *European journal of human genetics : EJHG*, vol. 24, pp. 421–428, mar 2016.

[205] M. Kwon, S. Leem, J. Yoon, and T. Park, "GxGrare: gene-gene interaction analysis method for rare variants from high-throughput sequencing data.," *BMC systems biology*, vol. 12, p. 19, mar 2018.

[206] A. T. Timberlake, J. Choi, S. Zaidi, Q. Lu, C. Nelson-Williams, E. D. Brooks, *et al.*, "Two locus inheritance of non-syndromic midline craniosynostosis via rare SMAD6 and common BMP2 alleles.," *eLife*, vol. 5, sep 2016.

[207] G. Kerner, M. Bouaziz, A. Cobat, B. Bigio, A. T. Timberlake, J. Bustamante, *et al.*, "A genome-wide case-only test for the detection of digenic inheritance in human exomes," *Proceedings of the National Academy of Sciences*, vol. 117, pp. 19367–19375, aug 2020.

[208] N. Borumandnia, H. A. Majd, N. Khadembashi, and H. Alaii, "Assessing the Trend of Infertility Rate in 198 Countries and Territories in Last Decades.," *Iranian journal of public health*, vol. 50, pp. 1735–1737, aug 2021.

[209] A. Agarwal, S. Baskaran, N. Parekh, C.-L. Cho, R. Henkel, S. Vij, *et al.*, "Male infertility.," *Lancet (London, England)*, vol. 397, pp. 319–333, jan 2021.

[210] C. Krausz, "Male infertility: Pathogenesis and clinical diagnosis," *Best Practice & Research Clinical Endocrinology & Metabolism*, vol. 25, no. 2, pp. 271–285, 2011.

[211] H. Tournaye, C. Krausz, and R. D. Oates, "Novel concepts in the aetiology of male reproductive impairment.," *The lancet. Diabetes & endocrinology*, vol. 5, pp. 544–553, jul 2017.

[212] E. Chung and G. B. Brock, "Cryptorchidism and its impact on male fertility: a state of art review of current literature.," *Canadian Urological Association journal = Journal de l'Association des urologues du Canada*, vol. 5, pp. 210–214, jun 2011.

[213] M. Cocuzza, C. Alvarenga, and R. Pagani, "The epidemiology and etiology of azoospermia.," *Clinics (Sao Paulo, Brazil)*, vol. 68 Suppl 1, no. Suppl 1, pp. 15–26, 2013.

[214] R. Fraietta, D. S. Zylberstejn, and S. C. Esteves, "Hypogonadotropic hypogonadism revisited.," *Clinics (Sao Paulo, Brazil)*, vol. 68 Suppl 1, no. Suppl 1, pp. 81–88, 2013.

[215] P. Asero, A. E. Calogero, R. A. Condorelli, L. Mongioi', E. Vicari, F. Lanzafame, *et al.*, "Relevance of genetic investigation in male infertility.," *Journal of endocrinological investigation*, vol. 37, pp. 415–427, may 2014.

[216] M. Punab, O. Poolamets, P. Paju, V. Vihljajev, K. Pomm, R. Ladva, *et al.*, "Causes of male infertility: a 9-year prospective monocentre study on 1737 patients with reduced total sperm counts.," *Human reproduction (Oxford, England)*, vol. 32, pp. 18–31, jan 2017.

[217] F. Tüttelmann, C. Ruckert, and A. Röpke, "Perspectives for novel genetic diagnostics after 20 years of unchanged routine," *Medizinische Genetik*, vol. 30, no. 1, pp. 12–20, 2018.

[218] M. Laan, L. Kasak, and M. Punab, "Translational aspects of novel findings in genetics of male infertility—status quo 2021," *British Medical Bulletin*, vol. 140, pp. 5–22, dec 2021.

[219] F. Lanfranco, A. Kamischke, M. Zitzmann, and E. Nieschlag, "Klinefelter's syndrome," *The Lancet*, vol. 364, pp. 273–283, jul 2004.

[220] L. Tiepolo and O. Zuffardi, "Localization of factors controlling spermatogenesis in the nonfluorescent portion of the human y chromosome long arm," *Human Genetics*, vol. 34, no. 2, pp. 119–124, 1976.

[221] P. H. Vog, A. Edelmann, S. Kirsch, O. Henegariu, P. Hirschmann, F. Kiesewetter, *et al.*, "Human Y Chromosome Azoospermia Factors (AZF) Mapped to Different Subregions in Yq11," *Human Molecular Genetics*, vol. 5, pp. 933–943, jul 1996.

[222] R. Reijo, R. K. Alagappan, D. C. Page, and P. Patrizio, "Severe oligozoospermia resulting from deletions of azoospermia factor gene on Y chromosome," *The Lancet*, vol. 347, pp. 1290–1293, may 1996.

[223] C. V. Hopps, A. Mielnik, M. Goldstein, G. D. Palermo, Z. Rosenwaks, and P. N. Schlegel, "Detection of sperm in men with Y chromosome microdeletions of the AZFa, AZFb and AZFc regions," *Human Reproduction*, vol. 18, pp. 1660–1665, aug 2003.

[224] T. Kuroda-Kawaguchi, H. Skaletsky, L. G. Brown, P. J. Minx, H. S. Cordum, R. H. Waterston, *et al.*, "The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men," *Nature Genetics*, vol. 29, no. 3, pp. 279–286, 2001.

[225] M. S. Oud, L. Volozonoka, R. M. Smits, L. E. Vissers, L. Ramos, and J. A. Veltman, "A systematic review and standardized clinical validity assessment of male infertility genes," *Human Reproduction*, vol. 34, no. 5, pp. 932–941, 2019.

[226] B. J. Houston, A. Riera-Escamilla, M. J. Wyrwoll, A. Salas-Huetos, M. J. Xavier, L. Nagirnaja, *et al.*, "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*, vol. 28, pp. 15–29, dec 2021.

[227] K. Lillepea, A.-G. Juchnewitsch, L. Kasak, A. Valkna, A. Dutta, K. Pomm, *et al.*, "Toward clinical exomes in diagnostics and management of male infertility," *The American Journal of Human Genetics*, 2024.

[228] J. M. Bieniek, C. D. Lapin, and K. A. Jarvi, "Genetics of cftr and male infertility," *Translational andrology and urology*, vol. 10, no. 3, p. 1391, 2021.

[229] E. Bieth, S. M. Hamdi, and R. Mieusset, "Genetics of the congenital absence of the vas deferens," *Human Genetics*, vol. 140, no. 1, pp. 59–76, 2021.

[230] M. Schwanzel-Fukuda, D. Bick, and D. W. Pfaff, "Luteinizing hormone-releasing hormone (LHRH)-expressing cells do not migrate normally in an inherited hypogonadal (Kallmann) syndrome.," *Brain research. Molecular brain research*, vol. 6, pp. 311–326, dec 1989.

[231] H.-J. Cho, Y. Shan, N. C. Whittington, and S. Wray, "Nasal Placode Development, GnRH Neuronal Migration and Kallmann Syndrome.," *Frontiers in cell and developmental biology*, vol. 7, p. 121, 2019.

[232] M. Kałużna, B. Budny, M. Rabijewski, J. Kałużny, A. Dubiel, M. Trofimiuk-Müldner, *et al.*, "Defects in GnRH Neuron Migration/Development and Hypothalamic-Pituitary Signaling Impact Clinical Variability of Kallmann Syndrome.," *Genes*, vol. 12, jun 2021.

[233] S. D. Quaynor, H.-G. Kim, E. M. Cappello, T. Williams, L. P. Chorich, D. P. Bick, *et al.*, "The prevalence of digenic mutations in patients with normosmic hypogonadotropic hypogonadism and Kallmann syndrome.," *Fertility and sterility*, vol. 96, pp. 1424–1430.e6, dec 2011.

[234] A. Gach, I. Pinkier, U. Wysocka, K. Sałacińska, D. Salachna, M. Szarras-Czapnik, *et al.*, "New findings in oligogenic inheritance of congenital hypogonadotropic hypogonadism.," *Archives of medical science : AMS*, vol. 18, no. 2, pp. 353–364, 2022.

[235] N. Alhathal, S. Maddirevula, S. Coskun, H. Alali, M. Assoum, T. Morris, *et al.*, "A genomics approach to male infertility," *Genetics in Medicine*, vol. 22, no. 12, pp. 1967–1975, 2020.

[236] L. Nagirnaja, A. M. Lopes, W.-L. Charng, B. Miller, R. Stakaitis, I. Golubickaite, *et al.*, "Diverse monogenic subforms of human spermatogenic failure.," *Nature communications*, vol. 13, p. 7953, dec 2022.

[237] L. Kasak and M. Laan, "Monogenic causes of non-obstructive azoospermia: challenges, established knowledge, limitations and perspectives," *Human Genetics*, vol. 140, no. 1, pp. 135–154, 2021.

[238] R. Cannarella, R. A. Condorelli, Y. Duca, S. La Vignera, and A. E. Calogero, "New insights into the genetics of spermatogenic failure: a review of the literature," *Human Genetics*, vol. 138, no. 2, pp. 125–140, 2019.

[239] M. S. Oud, R. Smits, H. Smith, F. K. Mastrorosa, G. Holt, B. Houston, *et al.*, "A de novo paradigm for male infertility," *Nature communications*, vol. 13, no. 1, p. 154, 2022.

[240] L. Nagirnaja, N. Mørup, J. E. Nielsen, R. Stakaitis, I. Golubickaite, M. S. Oud, *et al.*, "Variant pnldc1, defective pirna processing, and azoospermia," *New England Journal of Medicine*, vol. 385, no. 8, pp. 707–719, 2021.

[241] G. Quarantani, A. Sorgente, M. Alfano, G. B. Pipitone, L. Boeri, E. Pozzi, *et al.*, "Whole exome data prioritization unveils the hidden weight of Mendelian causes of male infertility. A report from the first Italian cohort.," *PloS one*, vol. 18, no. 8, p. e0288336, 2023.

[242] J. Young, C. Xu, G. E. Papadakis, J. S. Acierno, L. Maione, J. Hietamäki, *et al.*, "Clinical management of congenital hypogonadotropic hypogonadism," *Endocrine reviews*, vol. 40, no. 2, pp. 669–710, 2019.

[243] C. Kouri, G. Sommer, and C. E. Flück, "Oligogenic Causes of Human Differences of Sex Development: Facing the Challenge of Genetic Complexity," *Hormone Research in Paediatrics*, vol. 96, pp. 169–179, 05 2023.

[244] F. Del Giudice, A. M. Kasman, T. Chen, E. De Berardinis, G. M. Busetto, A. Sciarra, *et al.*, "The association between mortality and male infertility: systematic review and meta-analysis," *Urology*, vol. 154, pp. 148–157, 2021.

[245] P. N. Schlegel, M. Sigman, B. Collura, C. J. De Jonge, M. L. Eisenberg, D. J. Lamb, *et al.*, "Diagnosis and treatment of infertility in men: Aua/asrm guideline part i," *The Journal of urology*, vol. 205, no. 1, pp. 36–43, 2021.

[246] S. Karavolos, N. Panagiotopoulou, H. Alahwany, and S. Martins da Silva, "An update on the management of male infertility," *The Obstetrician & Gynaecologist*, vol. 22, no. 4, pp. 267–274, 2020.

[247] WHO, "WHO manual for the standardized investigation, diagnosis and management of the infertile male," tech. rep., World Health Organization, WHO, 2000.

[248] I. Khourdaji, H. Lee, and R. P. Smith, "Frontiers in hormone therapy for male infertility.," *Translational andrology and urology*, vol. 7, pp. S353–S366, jul 2018.

[249] N. Garrido and I. Hervás, "Personalized Medicine in Infertile Men.," *The Urologic clinics of North America*, vol. 47, pp. 245–255, may 2020.

[250] C. L. R. Barratt, C. Wang, E. Baldi, I. Toskin, J. Kiarie, and D. J. Lamb, "What advances may the future bring to the diagnosis, treatment, and care of male sexual and reproductive health?," *Fertility and sterility*, vol. 117, pp. 258–267, feb 2022.

[251] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander, "The mystery of missing heritability: Genetic interactions create phantom heritability," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 1193–1198, jan 2012.

[252] K. M. Boycott, T. Hartley, L. G. Biesecker, R. A. Gibbs, A. M. Innes, O. Riess, J. Belmont, S. L. Dunwoodie, N. Jojic, T. Lassmann, *et al.*, "A diagnosis for all rare genetic diseases: the horizon and the next frontiers," *Cell*, vol. 177, no. 1, pp. 32–37, 2019.

[253] C. Nachtegael, B. Gravel, A. Dillen, G. Smits, A. Nowé, S. Papadimitriou, *et al.*, "Scaling up oligogenic diseases research with OLIDA: the Oligogenic Diseases Database," *Database*, vol. 2022, 2022.

[254] N. Versbraegen, B. Gravel, C. Nachtegael, A. Renaux, E. Verkinderen, A. Nowé, *et al.*, "Faster and more accurate pathogenic combination predictions with VarCoPP2.0," *BMC Bioinformatics*, vol. 24, no. 1, p. 179, 2023.

[255] B. Gravel, A. Renaux, S. Papadimitriou, G. Smits, A. Nowé, and T. Lenaerts, "Prioritization of oligogenic variant combinations in whole exomes," *Bioinformatics*, vol. 40, p. btae184, apr 2024.

[256] S. Papadimitriou, B. Gravel, C. Nachtegael, E. De Baere, B. Loeys, M. Vikkula, *et al.*, "Toward reporting standards for the pathogenicity of variant combinations involved in multilocus/oligogenic diseases," *Human Genetics and Genomics Advances*, vol. 4, jan 2023.

[257] X. Zheng-Bradley and P. Flicek, "Applications of the 1000 Genomes Project resources.," *Briefings in functional genomics*, vol. 16, pp. 163–170, may 2017.

[258] T. K. Oleksyk, V. Brukhin, and S. J. O'Brien, "The Genome Russia project: closing the largest remaining omission on the world Genome map.," *GigaScience*, vol. 4, p. 53, 2015.

[259] M. Byrska-Bishop, U. S. Evani, X. Zhao, A. O. Basile, H. J. Abel, A. A. Regier, *et al.*, "High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios.," *Cell*, vol. 185, pp. 3426–3440.e19, sep 2022.

[260] M. Bouaziz, J. Mullaert, B. Bigio, Y. Seeleuthner, J.-L. Casanova, A. Alcais, *et al.*, "Controlling for human population stratification in rare variant association studies," *Scientific Reports*, vol. 11, no. 1, p. 19015, 2021.

[261] Y. Zhang, X. Shen, and W. Pan, "Adjusting for Population Stratification in a Fine Scale With Principal Components and Sequencing Data," *Genetic Epidemiology*, vol. 37, pp. 787–801, dec 2013.

[262] A. Boyd, J. Golding, J. Macleod, D. A. Lawlor, A. Fraser, J. Henderson, *et al.*, "Cohort profile: the 'children of the 90s'—the index offspring of the avon longitudinal study of parents and children," *International journal of epidemiology*, vol. 42, no. 1, pp. 111–127, 2013.

[263] A. Moayyeri, C. J. Hammond, D. J. Hart, and T. D. Spector, "The uk adult twin registry (twinsuk resource)," *Twin Research and Human Genetics*, vol. 16, no. 1, pp. 144–149, 2013.

[264] K. Walter, J. L. Min, J. Huang, L. Crooks, Y. Memari, S. McCarthy, *et al.*, "The UK10K project identifies rare variants in health and disease.," *Nature*, vol. 526, pp. 82–90, sep 2015.

[265] D. Muddyman, "The UK10K Project: 10,000 UK Genome Sequences—Accessing the Role of Rare Genetic Variants in Health and Disease," in *Assessing Rare Variation in Complex Traits: Design and Analysis of Genetic Studies* (E. Zeggini and A. Morris, eds.), pp. 87–105, Springer New York, 2015.

[266] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, *et al.*, "The human genome browser at UCSC.," *Genome research*, vol. 12, pp. 996–1006, jun 2002.

[267] L. R. Nassar, G. P. Barber, A. Benet-Pagès, J. Casper, H. Clawson, M. Diekhans, *et al.*, "The UCSC Genome Browser database: 2023 update.," *Nucleic acids research*, vol. 51, pp. D1188–D1195, jan 2023.

[268] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, *et al.*, "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, no. 7616, pp. 285–291, 2016.

[269] K. J. Karczewski, B. Weisburd, B. Thomas, M. Solomonson, D. M. Ruderfer, D. Kavanagh, *et al.*, "The ExAC browser: displaying reference data information from over 60 000 exomes," *Nucleic Acids Research*, vol. 45, pp. D840–D845, jan 2017.

[270] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, *et al.*, "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.," *PLoS medicine*, vol. 12, p. e1001779, mar 2015.

[271] S. Gudmundsson, M. Singer-Berk, N. A. Watts, W. Phu, J. K. Goodrich, M. Solomonson, *et al.*, "Variant interpretation using population databases: Lessons from gnomAD," *Human Mutation*, vol. 43, pp. 1012–1030, aug 2022.

[272] S. Chen, L. C. Francioli, J. K. Goodrich, R. L. Collins, M. Kanai, Q. Wang, *et al.*, "A genomic mutational constraint map using variation in 76,156 human genomes," *Nature*, vol. 625, no. 7993, pp. 92–100, 2024.

[273] K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, *et al.*, "The mutational constraint spectrum quantified from variation in 141,456 humans," *Nature*, vol. 581, no. 7809, pp. 434–443, 2020.

[274] S. Petrovski, A. B. Gussow, Q. Wang, M. Halvorsen, Y. Han, W. H. Weir, *et al.*, "The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity," *PLOS Genetics*, vol. 11, no. 9, pp. 1–25, 2015.

[275] L. Kasak, M. Punab, L. Nagirnaja, M. Grigorova, A. Minajeva, A. M. Lopes, *et al.*, "Bi-allelic Recessive Loss-of-Function Variants in FANCM Cause Non-obstructive Azoospermia.," *American journal of human genetics*, vol. 103, pp. 200–212, aug 2018.

[276] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, *et al.*, "The variant call format and VCFtools.," *Bioinformatics (Oxford, England)*, vol. 27, pp. 2156–2158, aug 2011.

[277] E. Garrison, Z. N. Kronenberg, E. T. Dawson, B. S. Pedersen, and P. Prins, "A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar.," *PLoS computational biology*, vol. 18, p. e1009123, may 2022.

[278] K. M. Boycott, A. Rath, J. X. Chong, T. Hartley, F. S. Alkuraya, G. Baynam, *et al.*, "International cooperation to enable the diagnosis of all rare genetic diseases," *The American Journal of Human Genetics*, vol. 100, no. 5, pp. 695–705, 2017.

[279] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160018, 2016.

[280] G. M. Dall'Olio, J. Bertranpetit, and H. Laayouni, "The annotation and the usage of scientific databases could be improved with public issue tracker software," *Database*, vol. 2010, p. baq035, jan 2010.

[281] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, "OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders," *Nucleic Acids Research*, vol. 43, pp. D789–D798, jan 2015.

[282] J. E. Harrison, S. Weber, R. Jakob, and C. G. Chute, "ICD-11: an international classification of diseases for the twenty-first century," *BMC Medical Informatics and Decision Making*, vol. 21, no. 6, p. 206, 2021.

[283] V. A. McKusick, *Mendelian inheritance in man: a catalog of human genes and genetic disorders*, vol. 1. JHU Press, 1998.

[284] T. Groza, S. Köhler, D. Moldenhauer, N. Vasilevsky, G. Baynam, T. Zemojtel, *et al.*, "The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease.," *American journal of human genetics*, vol. 97, pp. 111–124, jul 2015.

[285] C. L. Smith, C.-A. W. Goldsmith, and J. T. Eppig, "The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information.," *Genome biology*, vol. 6, no. 1, p. R7, 2005.

[286] J. Sprague, L. Bayraktaroglu, D. Clements, T. Conlin, D. Fashena, K. Frazer, *et al.*, "The Zebrafish Information Network: the zebrafish model organism database.," *Nucleic acids research*, vol. 34, pp. D581–5, jan 2006.

[287] Y. Bradford, T. Conlin, N. Dunn, D. Fashena, K. Frazer, D. G. Howe, *et al.*, "ZFIN: enhancements and updates to the Zebrafish Model Organism Database.," *Nucleic acids research*, vol. 39, pp. D822–9, jan 2011.

[288] D. Smedley, A. Oellrich, S. Köhler, B. Ruef, M. Westerfield, P. Robinson, *et al.*, "PhenoDigm: analyzing curated annotations to associate animal models with human diseases.," *Database : the journal of biological databases and curation*, vol. 2013, p. bat025, 2013.

[289] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, and M. A. Haendel, "Uberon, an integrative multi-species anatomy ontology.," *Genome biology*, vol. 13, p. R5, jan 2012.

[290] S. Köhler, S. C. Doelken, B. J. Ruef, S. Bauer, N. Washington, M. Westerfield, *et al.*, "Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research.," *F1000Research*, vol. 2, p. 30, 2013.

[291] P. N. Robinson and C. Webber, "Phenotype ontologies and cross-species analysis for translational research.," *PLoS genetics*, vol. 10, p. e1004268, apr 2014.

[292] W. P. Bone, N. L. Washington, O. J. Buske, D. R. Adams, J. Davis, D. Draper, *et al.*, "Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency.," *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 18, pp. 608–617, jun 2016.

[293] M. A. Gargano, N. Matentzoglu, B. Coleman, E. B. Addo-Lartey, A. V. Anagnostopoulos, J. Anderton, *et al.*, "The Human Phenotype Ontology in 2024: phenotypes around the world.," *Nucleic acids research*, vol. 52, pp. D1333–D1346, jan 2024.

[294] R. Steinhaus, S. Proft, E. Seelow, T. Schalau, P. N. Robinson, and D. Seelow, "Deep phenotyping: symptom annotation made simple with SAMS.," *Nucleic acids research*, vol. 50, pp. W677–W681, jul 2022.

[295] F. Dhombres and O. Bodenreider, "Interoperability between phenotypes in research and healthcare terminologies–Investigating partial mappings between HPO and SNOMED CT.," *Journal of biomedical semantics*, vol. 7, p. 3, 2016.

[296] M. Girdea, S. Dumitriu, M. Fiume, S. Bowdin, K. M. Boycott, S. Chénier, *et al.*, "PhenoTips: patient phenotyping software for clinical and research use.," *Human mutation*, vol. 34, pp. 1057–1065, aug 2013.

[297] J. A. McMurry, S. Köhler, N. L. Washington, J. P. Balhoff, C. Borromeo, M. Brush, *et al.*, "Navigating the Phenotype Frontier: The Monarch Initiative," *Genetics*, 2016.

[298] K. A. Shefchek, N. L. Harris, M. Gargano, N. Matentzoglu, D. Unni, M. Brush, *et al.*, "The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species," *Nucleic Acids Research*, vol. 48, pp. D704–D715, 1 2020.

[299] S. Choobdar, M. E. Ahsen, J. Crawford, M. Tomasoni, T. Fang, D. Lamparter, *et al.*, "Assessment of network module identification across complex diseases," *Nature Methods*, vol. 16, no. 9, pp. 843–852, 2019.

[300] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016. Community detection in networks: A user guide.

[301] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[302] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the national academy of sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[303] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães, "Gene co-expression analysis for functional classification and gene-disease predictions.," *Briefings in bioinformatics*, vol. 19, pp. 575–592, jul 2018.

[304] D. Amar, H. Safer, and R. Shamir, "Dissection of regulatory networks that are altered in disease via differential co-expression.," *PLoS computational biology*, vol. 9, no. 3, p. e1002955, 2013.

[305] GTEx Consortium, "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.," *Science*, vol. 348, pp. 648–660, may 2015.

[306] A. Battle, C. D. Brown, B. E. Engelhardt, and S. B. Montgomery, "Genetic effects on gene expression across human tissues.," *Nature*, vol. 550, pp. 204–213, oct 2017.

[307] P. Raina, R. Guinea, K. Chatsirisupachai, I. Lopes, Z. Farooq, C. Guinea, *et al.*, "GeneFriends: gene co-expression databases and tools for humans and model organisms.," *Nucleic acids research*, vol. 51, pp. D145–D158, jan 2023.

[308] T. Obayashi, S. Kodate, H. Hibara, Y. Kagaya, and K. Kinoshita, "COXPRESdb v8: an animal gene coexpression database navigating from a global view to detailed investigations.," *Nucleic acids research*, vol. 51, pp. D80–D87, jan 2023.

[309] N. Safari-Alighiarloo, M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, and A. A. Peyvandi, "Protein-protein interaction networks (PPI) and complex diseases.," *Gastroenterology and hepatology from bed to bench*, vol. 7, no. 1, pp. 17–31, 2014.

[310] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, *et al.*, "The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest.," *Nucleic acids research*, vol. 51, pp. D638–D646, jan 2023.

[311] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, *et al.*, "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic Acids Research*, vol. 42, pp. D358–D363, jan 2014.

[312] A. Calderone, L. Castagnoli, and G. Cesareni, "mentha: a resource for browsing integrated protein-interaction networks," *Nature Methods*, vol. 10, no. 8, pp. 690–691, 2013.

[313] D. V. Veres, D. M. Gyurkó, B. Thaler, K. Z. Szalay, D. Fazekas, T. Korcsmáros, *et al.*, "ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis," *Nucleic acids research*, vol. 43, pp. D485–D493, 1 2015.

[314] M. Gillespie, B. Jassal, R. Stephan, M. Milacic, K. Rothfels, A. Senff-Ribeiro, *et al.*, "The reactome pathway knowledgebase 2022," *Nucleic Acids Research*, vol. 50, pp. D687–D692, 1 2022.

[315] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, "KEGG: integrating viruses and cellular organisms," *Nucleic acids research*, vol. 49, pp. D545–D551, 1 2021.

[316] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, *et al.*, "Pathway Commons, a web resource for biological pathway data.," *Nucleic acids research*, vol. 39, pp. D685–90, jan 2011.

[317] C. Pesquita, "Semantic Similarity in the Gene Ontology.," *Methods in molecular biology (Clifton, N.J.)*, vol. 1446, pp. 161–173, 2017.

[318] G. K. Mazandu, E. R. Chimusa, and N. J. Mulder, "Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 886–901, 2017.

[319] A. B. Kamran and H. Naveed, "GOntoSim: a semantic similarity measure based on LCA and common descendants," *Scientific Reports*, vol. 12, no. 1, p. 3818, 2022.

[320] C. Zhao and Z. Wang, "GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms," *Scientific Reports*, vol. 8, no. 1, p. 15107, 2018.

[321] Y. Itan, S.-Y. Zhang, G. Vogt, A. Abhyankar, M. Herman, P. Nitschke, *et al.*, "The human gene connectome as a map of short cuts for morbid allele discovery," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, pp. 5558–5563, apr 2013.

[322] Y. Itan, M. Mazel, B. Mazel, A. Abhyankar, P. Nitschke, L. Quintana-Murci, *et al.*, "HGCS: an online tool for prioritizing disease-causing gene variants by biological distance," *BMC Genomics*, vol. 15, no. 1, p. 256, 2014.

[323] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[324] P. Jaccard, "The distribution of the flora in the Alpine zone," *New Phytologist*, vol. 11, pp. 37–50, feb 1912.

[325] G. de Anda-Jáuregui, "Guideline for comparing functional enrichment of biological network modular structures," *Applied Network Science*, vol. 4, no. 1, p. 13, 2019.

[326] I. A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, *et al.*, "Network-based prediction of protein interactions.," *Nature communications*, vol. 10, p. 1240, mar 2019.

[327] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: A universal amplifier of genetic associations," *Nature Reviews Genetics*, vol. 18, no. 9, pp. 551–562, 2017.

[328] S. Lehtinen, J. Lees, J. Bähler, J. Shawe-Taylor, and C. Orengo, "Gene Function Prediction from Functional Association Networks Using Kernel Partial Least Squares Regression," *PLOS ONE*, vol. 10, p. e0134668, aug 2015.

[329] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, *et al.*, "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.," *Nucleic acids research*, vol. 38, pp. W214–20, jul 2010.

[330] S. Erten, G. Bebek, R. M. Ewing, and M. Koyutürk, "DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization.," *BioData mining*, vol. 4, p. 19, jun 2011.

[331] O. Magger, Y. Y. Waldman, E. Ruppin, and R. Sharan, "Enhancing the Prioritization of Disease-Causing Genes through Tissue Specific Protein Interaction Networks," *PLOS Computational Biology*, vol. 8, p. e1002690, sep 2012.

[332] Y. Li and J. C. Patra, "Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network," *Bioinformatics*, vol. 26, no. 9, pp. 1219–1224, 2010.

[333] A. Valdeolivas, L. Tichit, C. Navarro, S. Perrin, G. Odelin, N. Levy, *et al.*, "Random walk with restart on multiplex and heterogeneous biological networks," *Bioinformatics*, vol. 35, pp. 497–505, 2 2019.

[334] C. Kingsford and S. L. Salzberg, "What are decision trees?," *Nature biotechnology*, vol. 26, pp. 1011–1013, sep 2008.

[335] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 281, 2019.

[336] J. Fürnkranz and E. Hüllermeier, *Preference Learning*. Berlin, Heidelberg: Springer-Verlag, 1st ed., 2010.

[337] T. Werner, *A review on instance ranking problems in statistical learning*, vol. 111. Springer US, 2022.

[338] H. Li, "A Short Introduction to Learning to Rank," *J-stage*, no. 10, pp. 1854–1862, 2011.

[339] X. Ru, X. Ye, T. Sakurai, and Q. Zou, "Application of learning to rank in bioinformatics tasks," *Briefings in Bioinformatics*, vol. 22, p. bbaa394, sep 2021.

[340] M. Taylor, J. Guiver, S. Robertson, and T. Minka, "SoftRank: optimizing non-smooth rank metrics," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, (New York, NY, USA), pp. 77–86, Association for Computing Machinery, 2008.

[341] C. J. C. Burges, "From ranknet to lambdarank to lambdamart: An overview," in *Microsoft Research Technical Report MSR-TR-2010-82*, 2010.

[342] C. J. C. Burges, R. Ragno, Q. V. Le, and C. J. C. Burges, "Learning to Rank with Non-Smooth Cost Functions," in *Advances in Neural Information Processing Systems 19*, MIT Press, Cambridge, MA, jan 2007.

[343] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, (New York, NY, USA), p. 129–136, Association for Computing Machinery, 2007.

[344] S. Agarwal and S. Sengupta, "Ranking genes by relevance to a disease," *8th International Conference on Computational Systems Bioinformatics (CSB)*, pp. 37–46, 2009.

[345] P.-F. Lee and V.-W. Soo, "An ensemble rank learning approach for gene prioritization.," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2013, pp. 3507–3510, 2013.

[346] U.-K. I. Umlai, D. K. Bangarusamy, X. Estivill, and P. V. Jithesh, "Genome sequencing data analysis for rare disease gene discovery," *Briefings in Bioinformatics*, vol. 23, p. bbab363, jan 2022.

[347] C. Kelly, A. Szabo, N. Pontikos, G. Arno, P. N. Robinson, J. O. B. Jacobsen, *et al.*, "Phenotype-aware prioritisation of rare Mendelian disease variants.," *Trends in genetics : TIG*, vol. 38, pp. 1271–1283, dec 2022.

[348] X. Yuan, J. Su, J. Wang, B. Dai, Y. Sun, K. Zhang, *et al.*, "Refined preferences of prioritizers improve intelligent diagnosis for Mendelian diseases.," *Scientific reports*, vol. 14, p. 2845, feb 2024.

[349] J. O. B. Jacobsen, C. Kelly, V. Cipriani, G. E. Research Consortium, C. J. Mungall, J. Reese, *et al.*, "Phenotype-driven approaches to enhance variant prioritization and diagnosis of rare disease.," *Human mutation*, vol. 43, pp. 1071–1081, aug 2022.

[350] D. Smedley, S. Köhler, J. C. Czeschik, J. Amberger, C. Bocchini, A. Hamosh, *et al.*, "Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases," *Bioinformatics*, vol. 30, pp. 3215–3222, nov 2014.

[351] J. E. Flores, D. M. Claborne, Z. D. Weller, B.-J. M. Webb-Robertson, K. M. Waters, and L. M. Bramer, "Missing data in multi-omics integration: Recent advances through artificial intelligence," *Frontiers in Artificial Intelligence*, vol. 6, 2023.

[352] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507–2517, aug 2007.

[353] A. M. Musolf, E. R. Holzinger, J. D. Malley, and J. E. Bailey-Wilson, "What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics," *Human Genetics*, vol. 141, no. 9, pp. 1515–1528, 2022.

[354] S. Nembrini, I. R. König, and M. N. Wright, "The revival of the Gini importance?," *Bioinformatics (Oxford, England)*, vol. 34, pp. 3711–3718, nov 2018.

[355] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, no. 1, p. 91, 2006.

[356] M. Uddin, Y. Wang, and M. Woodbury-Smith, "Artificial intelligence for precision medicine in neurodevelopmental disorders," *npj Digital Medicine*, vol. 2, no. 1, 2019.

[357] A. Holzinger and I. Jurisica, *Interactive knowledge discovery and data mining in biomedical informatics: state-of-the-art and future challenges*, vol. 8401. Springer, 2014.

[358] M. Rahman and D. N. Davis, "Addressing the Class Imbalance Problem in Medical Datasets," *International Journal of Machine Learning and Computing*, vol. 3, p. 224, 4 2013.

[359] G. Forman and M. Scholz, "Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement," *SIGKDD Explor. Newsl.*, vol. 12, p. 49–57, 11 2010.

[360] S. Daskalaki, I. Kopanas, and N. Avouris, "Evaluation of classifiers for an uneven class distribution problem," *Applied Artificial Intelligence*, vol. 20, pp. 381–417, 6 2006.

[361] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[362] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, jun 2024.

[363] P. D. Stenson, E. V. Ball, M. Mort, A. D. Phillips, J. A. Shiel, N. S. T. Thomas, *et al.*, "Human Gene Mutation Database (HGMD): 2003 update.," *Human mutation*, vol. 21, pp. 577–581, 6 2003.

[364] V. Cipriani, N. Pontikos, G. Arno, P. I. Sergouniotis, E. Lenassi, P. Thawong, *et al.*, "An Improved Phenotype-Driven Tool for Rare Mendelian Variant Prioritization: Benchmarking Exomiser on Real Patient Whole-Exome Data," *Genes*, vol. 11, no. 4, 2020.

[365] W. Chen, T.-y. Liu, Y. Lan, Z.-m. Ma, and H. Li, "Ranking Measures and Loss Functions in Learning to Rank," in *Advances in Neural Information Processing Systems* (Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, eds.), vol. 22, Curran Associates, Inc., 2009.

[366] A. Gazzo, D. Raimondi, D. Daneels, Y. Moreau, G. Smits, S. Van Dooren, *et al.*, "Understanding mutational effects in digenic diseases," *Nucleic Acids Research*, vol. 45, no. 15, pp. 1–11, 2017.

[367] R. Mkaouar, L. Abdallah, C. Naouali, S. Lahbib, Z. Turki, S. Elouej, *et al.*, "Oligogenic inheritance underlying incomplete penetrance of PROKR2 mutations in hypogonadotropic hypogonadism," *Frontiers in Genetics*, 2021.

[368] M. Laan, L. Kasak, K. Timinskas, M. Grigorova, V. Česlovas, A. Renaux, *et al.*, "NR5A1 c.991-1G > C splice-site variant causes familial 46,XY partial gonadal dysgenesis with incomplete penetrance," *Clinical Endocrinology*, 2020.

[369] S. Papdimitriou, *Towards multivariant pathogenicity predictions, Using machine learning to directly predict and explore disease-causing oligogenic variant combinations*. PhD thesis, Université Libre de Bruxelles, Vrije Universiteit Brussel, 2020.

[370] R. Bhaskaran and P. K. Ponnuswamy, "Dynamics of amino acid residues in globular proteins.," *International journal of peptide and protein research*, vol. 24, pp. 180–191, aug 1984.

[371] W. C. Wimley and S. H. White, "Experimentally determined hydrophobicity scale for proteins at membrane interfaces," *Nature structural biology*, vol. 3, pp. 842–848, oct 1996.

[372] N. Huang, I. Lee, E. M. Marcotte, and M. E. Hurles, "Characterising and Predicting Haploinsufficiency in the Human Genome," *PLOS Genetics*, vol. 6, p. e1001154, 10 2010.

[373] D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, *et al.*, "A systematic survey of loss-of-function variants in human protein-coding genes," *Science (New York, N.Y.)*, vol. 335, pp. 823–828, 2 2012.

[374] A. Renaux, C. Terwagne, M. Cochez, I. Tiddi, A. Nowé, and T. Lenaerts, "A knowledge graph approach to predict and interpret disease-causing gene interactions," *BMC Bioinformatics*, vol. 24, no. 1, p. 324, 2023.

[375] S. Lee, C. Zhang, M. Arif, Z. Liu, R. Benfeitas, G. Bidkhori, *et al.*, "TCSBN: a database of tissue and cancer specific biological networks.," *Nucleic acids research*, vol. 46, pp. D595–D600, 1 2018.

[376] D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, *et al.*, "The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets.," *Nucleic acids research*, vol. 49, pp. D605–D612, 1 2021.

[377] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.," *Nature genetics*, vol. 25, pp. 25–29, 5 2000.

[378] G. ontology consortium, "The Gene Ontology resource: enriching a GOld mine.," *Nucleic acids research*, vol. 49, pp. D325–D334, 1 2021.

[379] M. Blum, H.-Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, *et al.*, "The InterPro protein families and domains database: 20 years on," *Nucleic acids research*, vol. 49, p. D344—D354, 1 2021.

[380] M. Giurgiu, J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, *et al.*, "CORUM: the comprehensive resource of mammalian protein complexes-2019.," *Nucleic acids research*, vol. 47, pp. D559–D563, 1 2019.

[381] E. M. Scott, A. Halees, Y. Itan, E. G. Spencer, Y. He, M. A. Azab, *et al.*, "Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery," *Nature Genetics*, vol. 48, no. 9, pp. 1071–1076, 2016.

[382] C. N. Rotimi and L. B. Jorde, "Ancestry and Disease in the Age of Genomic Medicine," *New England Journal of Medicine*, vol. 363, no. 16, pp. 1551–1558, 2010.

[383] C. Kopanos, V. Tsiolkas, A. Kouris, C. E. Chapple, M. Albarca Aguilera, R. Meyer, *et al.*, "VarSome: the human genomic variant search engine," *Bioinformatics (Oxford, England)*, vol. 35, pp. 1978–1980, 6 2019.

[384] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, *et al.*, "dbSNP: the NCBI database of genetic variation," *Nucleic acids research*, vol. 29, pp. 308–311, 1 2001.

[385] B. Braschi, P. Denny, K. Gray, T. Jones, R. Seal, S. Tweedie, *et al.*, "Genenames.org: the HGNC and VGNC resources in 2019," *Nucleic acids research*, vol. 47, pp. D786–D792, 1 2019.

[386] J. M. Schwarz, D. N. Cooper, M. Schuelke, and D. Seelow, "MutationTaster2: mutation prediction for the deep-sequencing age," *Nature methods*, vol. 11, no. 4, pp. 361–362, 2014.

[387] K. Horackova, M. Janatova, P. Kleiblova, Z. Kleibl, and J. Soukupova, "Early-Onset Ovarian Cancer <30 Years: What Do We Know about Its Genetic Predisposition?," *International Journal of Molecular Sciences*, vol. 24, no. 23, 2023.

[388] C. M. Neuhofer and H. Prokisch, "Digenic Inheritance in Rare Disorders and Mitochondrial Disease—Crossing the Frontier to a More Comprehensive Understanding of Etiology," *International Journal of Molecular Sciences*, vol. 25, no. 9, 2024.

[389] J. M. Parmar, N. G. Laing, M. L. Kennerson, and G. Ravenscroft, "Genetics of inherited peripheral neuropathies and the next frontier: looking backwards to progress forwards," *Journal of Neurology, Neurosurgery & Psychiatry*, pp. jnnp–2024–333436, 2024.

[390] Q. T. Cowan, S. Gu, W. Gu, B. L. Ranzau, T. S. Simonson, and A. C. Komor, "Development of multiplexed orthogonal base editor (MOBE) systems," *Nature Biotechnology*, 2024.

[391] C. Nachtegael, J. De Stefani, A. Cnudde, and T. Lenaerts, "DUVEL: an active-learning annotated biomedical corpus for the recognition of oligogenic combinations," *Database*, vol. 2024, p. baae039, jan 2024.

[392] K. Ichioka, T. Yoshikawa, H. Kimura, and R. Saito, "Additional mutation in PROKR2 and phenotypic differences in a Kallmann syndrome/normosmic congenital hypogonadotropic hypogonadism family carrying FGFR1 missense mutation," *BMJ Case Reports CP*, vol. 17, no. 1, 2024.

[393] L. Graziani, S. Zampatti, M. L. Carriero, C. Minotti, C. Peconi, M. Bengala, *et al.*, "Co-Inheritance of Pathogenic Variants in PKD1 and PKD2 Genes Determined by Parental Segregation and De Novo Origin: A Case Report," *Genes*, vol. 14, no. 8, 2023.

[394] M. V. Nachury, A. V. Loktev, Q. Zhang, C. J. Westlake, J. Peränen, A. Merdes, *et al.*, "A core complex of bbs proteins cooperates with the gtpase rab8 to promote ciliary membrane biogenesis," *Cell*, vol. 129, no. 6, pp. 1201–1213, 2007.

[395] W. A. Haynes, A. Tomczak, and P. Khatri, "Gene annotation bias impedes biomedical research," *Scientific Reports*, vol. 8, no. 1, p. 1362, 2018.

[396] T. Stoeger, M. Gerlach, R. I. Morimoto, and L. A. Nunes Amaral, "Large-scale investigation of the reasons why potentially important genes are ignored," *PLOS Biology*, vol. 16, no. 9, pp. 1–25, 2018.

[397] A. M. Edwards, R. Isserlin, G. D. Bader, S. V. Frye, T. M. Willson, and F. H. Yu, "Too many roads not taken," *Nature*, vol. 470, no. 7333, pp. 163–165, 2011.

[398] M. Bosio, O. Drechsel, R. Rahman, F. Muyas, R. Rabionet, D. Bezdan, *et al.*, "ediva—classification and prioritization of pathogenic variants for clinical diagnostics," *Human mutation*, vol. 40, no. 7, pp. 865–878, 2019.

[399] S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, and D. B. Goldstein, "Genic intolerance to functional variation and the interpretation of personal genomes.," *PLoS genetics*, vol. 9, no. 8, p. e1003709, 2013.

[400] H. A. Shihab, M. F. Rogers, C. Campbell, and T. R. Gaunt, "HIPred: an integrative approach to predicting haploinsufficient genes," *Bioinformatics*, vol. 33, p. 1751, 6 2017.

[401] Z. Yang and J. P. Bielawski, "Statistical methods for detecting molecular adaptation," *Trends in ecology & evolution*, vol. 15, pp. 496–503, 12 2000.

[402] J. S. Hsu, J. S. Kwan, Z. Pan, M. M. Garcia-Barcelo, P. C. Sham, and M. Li, "Inheritance-mode specific pathogenicity prioritization (ISPP) for human protein coding genes," *Bioinformatics*, vol. 32, pp. 3065–3071, 10 2016.

[403] C. Pesquita, D. Faria, H. Bastos, A. E. N. Ferreira, A. O. Falcão, and F. M. Couto, "Metrics for GO based protein semantic similarity: a systematic evaluation," *BMC Bioinformatics*, vol. 9, no. 5, p. S4, 2008.

[404] A. G. Karegowda, M. A. Jayaram, and A. S. Manjunath, "Feature subset selection problem using wrapper approach in supervised learning," *International journal of Computer applications*, vol. 1, no. 7, pp. 13–17, 2010.

[405] C. Chen, A. Liaw, L. Breiman, and others, "Using random forest to learn imbalanced data," *University of California, Berkeley*, vol. 110, no. 1-12, p. 24, 2004.

[406] T. Doğan, "HPO2GO: prediction of human phenotype ontology term associations for proteins using cross ontology annotation co-occurrences," *PeerJ*, vol. 6, no. 8, 2018.

[407] A.-G. Juchnewitsch, K. Pomm, A. Dutta, E. Tamp, A. Valkna, K. Lillepea, *et al.*, "Undiagnosed RASopathies in infertile men," *Frontiers in Endocrinology*, vol. 15, no. April, pp. 1–13, 2024.

[408] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, *et al.*, "Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, p. 841, feb 2015.

[409] J. Zschocke, P. H. Byers, and A. O. M. Wilkie, "Mendelian inheritance revisited: dominance and recessiveness in medical genetics," *Nature Reviews Genetics*, vol. 24, no. 7, pp. 442–463, 2023.

[410] C. Fallerini, M. Baldassarri, E. Trevisson, V. Morbidoni, A. La Manna, R. Lazzarin, *et al.*, "Alport syndrome: impact of digenic inheritance in patients management.," *Clinical genetics*, vol. 92, pp. 34–44, jul 2017.

[411] M. H. Guo, L. Plummer, Y.-M. Chan, J. N. Hirschhorn, and M. F. Lippincott, "Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data.," *American journal of human genetics*, vol. 103, pp. 522–534, oct 2018.

[412] Y.-C. A. Feng, D. P. Howrigan, L. E. Abbott, K. Tashman, F. Cerrato, T. Singh, *et al.*, "Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals," *The American Journal of Human Genetics*, vol. 105, no. 2, pp. 267–282, 2019.

[413] L. Bundalian, Y.-Y. Su, S. Chen, A. Velluva, A. S. Kirstein, A. Garten, *et al.*, "Epilepsies of presumed genetic etiology show enrichment of rare variants that occur in the general population.," *American journal of human genetics*, vol. 110, pp. 1110–1122, jul 2023.

[414] E. Eisenberg and E. Y. Levanon, "Human housekeeping genes, revisited," *Trends in Genetics*, vol. 29, pp. 569–574, oct 2013.

[415] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation plink: rising to the challenge of larger and richer datasets," *GigaScience*, vol. 4, pp. s13742–015–0047–8, 02 2015.

[416] A. Kaprara and I. T. Huhtaniemi, "The hypothalamus-pituitary-gonad axis: Tales of mice and men," *Metabolism*, vol. 86, pp. 3–17, 2018.

[417] M. Quinodoz, B. Royer-Bertrand, K. Cisarova, S. A. Di Gioia, A. Superti-Furga, and C. Rivolta, "DOMINO: Using Machine Learning to Predict Genes Associated with Dominant Disorders," *The American Journal of Human Genetics*, vol. 101, no. 4, pp. 623–629, 2017.

[418] X. Wang, X. Zhang, S. Liu, G. Li, L. Cui, Y. Qin, *et al.*, "Novel mutations in the TP63 gene are potentially associated with Müllerian duct anomalies," *Human Reproduction*, vol. 31, pp. 2865–2871, dec 2016.

[419] A. Bashamboo, R. Brauner, J. Bignon-Topalovic, S. Lortat-Jacob, V. Karageorgou, D. Lourenco, *et al.*, "Mutations in the FOG2/ZFPM2 gene are associated with anomalies of human testis determination," *Human Molecular Genetics*, vol. 23, no. 14, pp. 3657–3665, 2014.

[420] J. J. Tremblay, N. M. Robert, and R. S. Viger, "Modulation of endogenous gata-4 activity reveals its dual contribution to mullerian inhibiting substance gene transcription in sertoli cells," *Molecular Endocrinology*, vol. 15, no. 9, pp. 1636–1650, 2001.

[421] N. L. Manuylov, Y. Fujiwara, I. I. Adameyko, F. Poulat, and S. G. Tevosian, "The regulation of sox9 gene expression by the gata4/fog2 transcriptional complex in dominant xx sex reversal mouse models," *Developmental biology*, vol. 307, no. 2, pp. 356–367, 2007.

[422] S. G. Tevosian, K. H. Albrecht, J. D. Crispino, Y. Fujiwara, E. M. Eicher, and S. H. Orkin, "Gonadal differentiation, sex determination and normal sry expression in mice require direct interaction between transcription partners gata4 and fog2," *Development*, 2002.

[423] J. A. van den Bergen, G. Robevska, S. Eggers, S. Riedl, S. R. Grover, P. B. Bergman, *et al.*, "Analysis of variants in GATA4 and FOG2/ZFPM2 demonstrates benign contribution to 46,XY disorders of sex development.," *Molecular genetics I& genomic medicine*, vol. 8, p. e1095, mar 2020.

[424] A. De Luca, A. Sarkozy, R. Ferese, F. Consoli, F. Lepri, M. L. Dentici, *et al.*, "New mutations in ZFPM2/FOG2 gene in tetralogy of Fallot and double outlet right ventricle.," *Clinical genetics*, vol. 80, pp. 184–190, aug 2011.

[425] L. C. A. D'Alessandro, S. Al Turki, A. K. Manickaraj, D. Manase, B. J. M. Mulder, L. Bergin, *et al.*, "Exome sequencing identifies rare variants in multiple genes in atrioventricular septal defect.," *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 18, pp. 189–198, feb 2016.

[426] H.-G. Kim, I. Kurth, F. Lan, I. Meliciani, W. Wenzel, S. H. Eom, *et al.*, "Mutations in CHD7, encoding a chromatin-remodeling protein, cause idiopathic hypogonadotropic hypogonadism and Kallmann syndrome.," *American journal of human genetics*, vol. 83, pp. 511–519, oct 2008.

[427] R. Balasubramanian, J.-H. Choi, L. Francescatto, J. Willer, E. R. Horton, E. P. Asimacopoulos, *et al.*, "Functionally compromised CHD7 alleles in patients with isolated GnRH deficiency.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, pp. 17953–17958, dec 2014.

[428] A. Malcher, T. Stokowy, A. Berman, M. Olszewska, P. Jedrzejczak, D. Sielski, *et al.*, "Whole-genome sequencing identifies new candidate genes for nonobstructive azoospermia.," *Andrology*, vol. 10, pp. 1605–1624, nov 2022.

[429] Y. Yasukochi, J. Sakuma, I. Takeuchi, K. Kato, M. Oguri, T. Fujimaki, H. Horibe, and Y. Yamada, "Identification of three genetic variants as novel susceptibility loci for body mass index in a japanese population," *Physiological genomics*, vol. 50, no. 3, pp. 179–189, 2018.

[430] R. Cacace, K. Sleegers, and C. Van Broeckhoven, "Molecular genetics of early-onset alzheimer's disease revisited," *Alzheimer's & dementia*, vol. 12, no. 6, pp. 733–748, 2016.

[431] M. Shoji, T. Kawarabayashi, Y. Harigaya, H. Yamaguchi, S. Hirai, T. Kamimura, *et al.*, "Alzheimer amyloid beta-protein precursor in sperm development.," *The American journal of pathology*, vol. 137, pp. 1027–1032, nov 1990.

[432] C. S. von Koch, H. Zheng, H. Chen, M. Trumbauer, G. Thinakaran, L. H. van der Ploeg, *et al.*, "Generation of APLP2 KO mice and early postnatal lethality in APLP2/APP double KO mice.," *Neurobiology of aging*, vol. 18, no. 6, pp. 661–669, 1997.

[433] J. V. Silva, S. Yoon, S. Domingues, S. Guimarães, A. V. Goltsev, E. F. da Cruz E Silva, *et al.*, "Amyloid precursor protein interaction network in human testis: sentinel proteins for male reproduction.," *BMC bioinformatics*, vol. 16, p. 12, jan 2015.

[434] N. Camats, M. Fernández-Cancio, L. Audí, A. Schaller, and C. E. Flück, "Broad phenotypes in heterozygous NR5A1 46,XY patients with a disorder of sex development: an oligogenic origin?," *European journal of human genetics : EJHG*, vol. 26, pp. 1329–1338, sep 2018.

[435] K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, *et al.*, "Predicting Splicing from Primary Sequence with Deep Learning.," *Cell*, vol. 176, pp. 535–548.e24, jan 2019.

[436] L. C. Patiño, I. Beau, A. Morel, B. Delemer, J. Young, N. Binart, *et al.*, "Functional evidence implicating NOTCH2 missense mutations in primary ovarian insufficiency etiology," *Human Mutation*, vol. 40, pp. 25–30, jan 2019.

[437] C. M. Bui, T. Pukhalskaya, B. R. Smoller, H. B. Zengin, S. Heneidi, E. Vail, *et al.*, "Two distinct pathogenic pathways of digital papillary adenocarcinoma — BRAF mutation or low-risk HPV infection," *Journal of Cutaneous Pathology*, vol. 50, pp. 568–576, jun 2023.

[438] S. M. King, "Axonemal dynein arms," *Cold Spring Harbor Perspectives in Biology*, vol. 8, no. 11, p. a028100, 2016.

[439] M. Weng, Y. Sha, Y. u. Zeng, N. Huang, W. Liu, X. Zhang, and H. Zhou, "Mutations in dnah8 contribute to multiple morphological abnormalities of sperm flagella and male infertility," *Acta Biochimica et Biophysica Sinica*, vol. 53, no. 4, pp. 472–480, 2021.

[440] E. Engelen, U. Akinci, J. C. Bryne, J. Hou, C. Gontan, M. Moen, *et al.*, "Sox2 cooperates with Chd7 to regulate genes that are mutated in human syndromes," *Nature Genetics*, vol. 43, no. 6, pp. 607–611, 2011.

[441] P. Sengupta, S. Dutta, I. R. Karkada, and S. V. Chinni, "Endocrinopathies and Male Infertility.," *Life (Basel, Switzerland)*, vol. 12, dec 2021.

[442] G. E. Moldovan, L. Miele, and A. T. Fazleabas, "Notch signaling in reproduction.," *Trends in endocrinology and metabolism: TEM*, vol. 32, pp. 1044–1057, dec 2021.

[443] D. Murta, M. Batista, E. Silva, A. Trindade, D. Henrique, A. Duarte, *et al.*, "Notch signaling in the epididymal epithelium regulates sperm motility and is transferred at a distance within epididymosomes," *Andrology*, vol. 4, no. 2, pp. 314–327, 2016.

[444] A. Mottis, L. Mouchiroud, and J. Auwerx, "Emerging roles of the corepressors NCoR1 and SMRT in homeostasis.," *Genes & development*, vol. 27, pp. 819–835, apr 2013.

[445] R. H. Costa-e Sousa, I. Astapova, F. Ye, F. E. Wondisford, and A. N. Hollenberg, "The thyroid axis is regulated by NCoR1 via its actions in the pituitary.," *Endocrinology*, vol. 153, pp. 5049–5057, oct 2012.

[446] A. Alahmar, S. Dutta, and P. Sengupta, "Thyroid hormones in male reproduction and infertility," *Asian Pacific Journal of Reproduction*, vol. 8, no. 5, 2019.

[447] P. D. Thomas, D. Ebert, A. Muruganujan, T. Mushayahama, L.-P. Albou, and H. Mi, "PANTHER: Making genome-scale phylogenetics accessible to all," *Protein Science*, vol. 31, no. 1, pp. 8–22, 2022.

[448] E. Sjöstedt, W. Zhong, L. Fagerberg, M. Karlsson, N. Mitsios, C. Adori, P. Oksvold, F. Edfors, A. Limiszewska, F. Hikmet, J. Huang, Y. Du, L. Lin, Z. Dong, L. Yang, X. Liu, H. Jiang, X. Xu, J. Wang, H. Yang, L. Bolund, A. Mardinoglu, C. Zhang, K. von Feilitzen, C. Lindskog, F. Pontén, Y. Luo, T. Hökfelt, M. Uhlén, and J. Mulder, "An atlas of the protein-coding genes in the human, pig, and mouse brain.," *Science (New York, N.Y.)*, vol. 367, mar 2020.

[449] M. Wang, Y. Xu, Y. Zhang, Y. Chen, G. Chang, G. An, *et al.*, "Deciphering the autophagy regulatory network via single-cell transcriptome analysis reveals a requirement for autophagy homeostasis in spermatogenesis.," *Theranostics*, vol. 11, no. 10, pp. 5010–5027, 2021.

[450] M. D. Griswold, "The central role of Sertoli cells in spermatogenesis.," *Seminars in cell & developmental biology*, vol. 9, pp. 411–416, aug 1998.

[451] R. Shiri-Sverdlov, A. Custers, J. V. van Vliet-Ostaptchouk, P. J. J. van Gorp, P. J. Lindsey, J. H. O. van Tilburg, *et al.*, "Identification of TUB as a novel candidate gene influencing body weight in humans.," *Diabetes*, vol. 55, pp. 385–389, feb 2006.

[452] A. D. Borman, L. R. Pearce, D. S. Mackay, K. Nagel-Wolfrum, A. E. Davidson, R. Henderson, *et al.*, "A homozygous mutation in the TUB gene associated with retinal dystrophy and obesity.," *Human mutation*, vol. 35, pp. 289–293, mar 2014.

[453] V. J. M. Nies, D. Struik, M. G. M. Wolfs, S. S. Rensen, E. Szalowska, U. A. Unmehopa, *et al.*, "TUB gene expression in hypothalamus and adipose tissue and its association with obesity in humans.," *International journal of obesity (2005)*, vol. 42, pp. 376–383, mar 2018.

[454] M. Zheng, X. Chen, Y. Cui, W. Li, H. Dai, Q. Yue, *et al.*, "TULP2, a New RNA-Binding Protein, Is Required for Mouse Spermatid Differentiation and Male Fertility," *Frontiers in Cell and Developmental Biology*, vol. 9, no. February, pp. 1–13, 2021.

[455] Y. Oyama, H. Miyata, K. Shimada, T. Larasati, Y. Fujihara, and M. Ikawa, "TULP2 deletion mice exhibit abnormal outer dense fiber structure and male infertility," *Reproductive Medicine and Biology*, vol. 21, p. e12467, jan 2022.

[456] C.-W. A. Feng, C. Spiller, D. J. Merriner, M. K. O'Bryan, J. Bowles, and P. Koopman, "SOX30 is required for male fertility in mice," *Scientific Reports*, vol. 7, no. 1, p. 17619, 2017.

[457] F. Han, X. Jiang, Z.-M. Li, X. Zhuang, X. Zhang, W.-M. Ouyang, *et al.*, "Epigenetic Inactivation of SOX30 Is Associated with Male Infertility and Offers a Therapy Target for Non-obstructive Azoospermia.," *Molecular therapy. Nucleic acids*, vol. 19, pp. 72–83, mar 2020.

[458] P. Jelinic, J. J. Mueller, N. Olvera, F. Dao, S. N. Scott, R. Shah, *et al.*, "Recurrent SMARCA4 mutations in small cell carcinoma of the ovary.," *Nature genetics*, vol. 46, pp. 424–426, may 2014.

[459] S. La Salle, F. Sun, X.-D. Zhang, M. J. Matunis, and M. A. Handel, "Developmental control of sumoylation pathway proteins in mouse male germ cells.," *Developmental biology*, vol. 321, pp. 227–237, sep 2008.

[460] J. V. Silva, M. J. Freitas, B. R. Correia, L. Korrodi-Gregório, A. Patrício, S. Pelech, *et al.*, "Profiling signaling proteins in human spermatozoa: biomarker identification for sperm quality evaluation," *Fertility and Sterility*, vol. 104, pp. 845–856.e8, oct 2015.

[461] A. R. Latchford, K. Neale, R. K. S. Phillips, and S. K. Clark, "Juvenile polyposis syndrome: a study of genotype, phenotype, and long-term outcome.," *Diseases of the colon and rectum*, vol. 55, pp. 1038–1043, oct 2012.

[462] T. Matsumoto, J. Umeno, K. Jimbo, M. Arai, I. Iwama, H. Kashida, *et al.*, "Clinical Guidelines for Diagnosis and Management of Juvenile Polyposis Syndrome in Children and Adults-Secondary Publication.," *Journal of the anus, rectum and colon*, vol. 7, no. 2, pp. 115–125, 2023.

[463] R. A. Rey and R. P. Grinspon, "Anti-Müllerian hormone, testicular descent and cryptorchidism.," *Frontiers in endocrinology*, vol. 15, p. 1361032, 2024.

[464] J. Gillis, S. Ballouz, and P. Pavlidis, "Bias tradeoffs in the creation and analysis of protein-protein interaction networks.," *Journal of proteomics*, vol. 100, pp. 44–54, apr 2014.

[465] J. Yang, L. Shu, H. Duan, and H. Li, "A robust phenotype-driven likelihood ratio analysis approach assisting interpretable clinical diagnosis of rare diseases," *Journal of Biomedical Informatics*, vol. 142, p. 104372, 2023.

[466] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering*, vol. 2, no. 10, pp. 719–731, 2018.

[467] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[468] C.-C. Huang and Z. Lu, "Community challenges in biomedical text mining over 10 years: success, failure and the future," *Briefings in bioinformatics*, vol. 17, no. 1, pp. 132–144, 2016.

[469] C.-H. Wei, A. Allot, R. Leaman, and Z. Lu, "Pubtator central: automated concept annotation for biomedical full text articles," *Nucleic acids research*, vol. 47, no. W1, pp. W587–W593, 2019.

[470] J. Saez-Rodriguez, J. C. Costello, S. H. Friend, M. R. Kellen, L. Mangravite, P. Meyer, *et al.*, "Crowdsourcing biomedical research: leveraging communities as innovation engines.," *Nature reviews. Genetics*, vol. 17, pp. 470–486, jul 2016.

[471] S. Nunes, R. T. Sousa, and C. Pesquita, "Multi-domain knowledge graph embeddings for gene-disease association prediction," *Journal of Biomedical Semantics*, vol. 14, no. 1, p. 11, 2023.

[472] I. Bosch, *Knowledge graph embeddings for the prediction of pathogenic gene pairs.* PhD thesis, Université Libre de Bruxelles, 2023.

[473] F. Shen, S. Peng, Y. Fan, A. Wen, S. Liu, Y. Wang, *et al.*, "Hpo2vec+: Leveraging heterogeneous knowledge resources to enrich node embeddings for the human phenotype ontology," *Journal of biomedical informatics*, vol. 96, p. 103246, 2019.

[474] C. Peng, S. Dieck, A. Schmid, A. Ahmad, A. Knaus, M. Wenzel, *et al.*, "CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph.," *NAR genomics and bioinformatics*, vol. 3, p. lqab078, sep 2021.

[475] F. Gualdi, B. Oliva, and J. Piñero, "Predicting gene disease associations with knowledge graph embeddings for diseases with curtailed information.," *NAR genomics and bioinformatics*, vol. 6, p. lqae049, jun 2024.

[476] M. Schubach, T. Maass, L. Nazaretyan, S. Röner, and M. Kircher, "CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions.," *Nucleic acids research*, vol. 52, pp. D1143–D1154, jan 2024.

[477] P. W. Harrison, M. R. Amode, O. Austine-Orimoloye, A. Azov, M. Barba, I. Barnes, *et al.*, "Ensembl 2024," *Nucleic Acids Research*, vol. 52, pp. D891–D899, jan 2024.

[478] B. Langmead and A. Nellore, "Cloud computing for genomic data analysis and collaboration," *Nature Reviews Genetics*, vol. 19, no. 4, pp. 208–219, 2018.

[479] J. Davis-Turak, S. M. Courtney, E. S. Hazard, W. B. J. Glen, W. A. da Silveira, T. Wesselman, *et al.*, "Genomics pipelines and data integration: challenges and opportunities in the research setting.," *Expert review of molecular diagnostics*, vol. 17, pp. 225–237, mar 2017.

[480] A. K. Manrai, B. H. Funke, H. L. Rehm, M. S. Olesen, B. A. Maron, P. Szolovits, *et al.*, "Genetic misdiagnoses and the potential for health disparities," *New England Journal of Medicine*, vol. 375, no. 7, pp. 655–665, 2016.

[481] D. C. Hoskinson, A. M. Dubuc, and H. Mason-Suares, "The current state of clinical interpretation of sequence variants.," *Current opinion in genetics & development*, vol. 42, pp. 33–39, feb 2017.

[482] M. A. Field, "Bioinformatic Challenges Detecting Genetic Variation in Precision Medicine Programs.," *Frontiers in medicine*, vol. 9, p. 806696, 2022.

[483] E. Masson, W.-B. Zou, E. Génin, D. N. Cooper, G. Le Gac, Y. Fichou, *et al.*, "Expanding ACMG variant classification guidelines into a general framework.," *Human genomics*, vol. 16, p. 31, aug 2022.

[484] A. Takata, K. Hamanaka, and N. Matsumoto, "Refinement of the clinical variant interpretation framework by statistical evidence and machine learning.," *Med (New York, N.Y.)*, vol. 2, pp. 611–632.e9, may 2021.

[485] A. M. Roberts, M. T. DiStefano, E. R. Riggs, K. S. Josephs, F. S. Alkuraya, J. Amberger, *et al.*, "Toward robust clinical genome interpretation: Developing a consistent terminology to characterize Mendelian disease-gene relationships-allelic requirement, inheritance modes, and disease mechanisms.," *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 26, p. 101029, feb 2024.

[486] P. Kountouris, C. Stephanou, C. W. Lederer, J. Traeger-Synodinos, C. Bento, C. L. Harteveld, *et al.*, "Adapting the ACMG/AMP variant classification framework: A perspective from the ClinGen Hemoglobinopathy Variant Curation Expert Panel.," *Human mutation*, vol. 43, pp. 1089–1096, aug 2022.

[487] B. J. Livesey and J. A. Marsh, "Interpreting protein variant effects with computational predictors and deep mutational scanning," *Disease Models & Mechanisms*, vol. 15, p. dmm049510, jun 2022.

[488] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, *et al.*, "Big data: astronomical or genomical?," *PLoS biology*, vol. 13, no. 7, p. e1002195, 2015.

[489] P. S. Samarakoon, H. S. Sorte, A. Stray-Pedersen, O. K. Rødningen, T. Rognes, and R. Lyle, "cnvscan: a cnv screening and annotation tool to improve the clinical utility of computational cnv prediction from exome sequencing data," *BMC genomics*, vol. 17, pp. 1–11, 2016.

[490] F. Requena, H. H. Abdallah, A. García, P. Nitschké, S. Romana, V. Malan, *et al.*, "CNVxplorer: a web tool to assist clinical interpretation of CNVs in rare disease patients," *Nucleic Acids Research*, vol. 49, pp. W93–W103, jul 2021.

[491] F. J. Massey, "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[492] R. A. Armstrong, "When to use the Bonferroni correction.," *Ophthalmic & physiological optics : the journal of the British College of Ophthalmic Opticians (Optometrists)*, vol. 34, pp. 502–508, 9 2014.