Statistical strategies leveraging population data to help with the diagnosis of rare diseases

Marie-Sophie Ogloblinsky

COLLEGE DOCTORAL BRETAGNE SCIENCES DE LA VIE ET DE LA SANTE

Université de Bretagne Occidentale

THESE DE DOCTORAT DE

L'UNIVERSITE
DE BRETAGNE OCCIDENTALE

ECOLE DOCTORALE N° 637
Sciences de la Vie et de la Santé
Spécialité : *Génétique, Génomique et Bioinformatique*

Par
Marie-Sophie OGLOBLINSKY

Statistical strategies leveraging population data to help with the diagnosis of rare diseases

Thèse présentée et soutenue à Brest, le 28 juin 2024
Unité de recherche : INSERM UMR1078 – Génétique, Génomique fonctionnelle et Biotechnologies

Rapporteurs avant soutenance :

Aurélie COBAT	CR – Institut Imagine UMR1163
Antonio RAUSELL	MCU-PH – Institut Imagine UMR1163

Composition du Jury :

Président :	Gérald LE GAC	PU-PH – UMR1078, INSERM, UBO, EFS, CHRU
Rapporteurs :	Aurélie COBAT	CR – Institut Imagine UMR1163
	Antonio RAUSELL	MCU-PH – Institut Imagine UMR1163
Examinateurs :	Donald CONRAD	Associate Professor – Oregon National Primate Research Center
	Marie DE TAYRAC	PU-PH – IGDR UMR6290, CNRS, UR
	Gérald LE GAC	PU-PH – UMR1078, INSERM, UBO, EFS, CHRU
	Gaël NICOLAS	PU-PH – CBG UMR1245, INSERM, UNIROUEN
Dir. de thèse :	Emmanuelle GENIN	DR – UMR1078, INSERM, UBO, EFS, CHRU
Encdr. de thèse :	Gaëlle MARENNE	IR – UMR1078, INSERM, UBO, EFS, CHRU

TABLE OF CONTENTS

Acknowledgements

Embarking on a PhD journey was a very deliberate decision on my part, enough for me to put on hold my medical studies which are already quite lengthy. In the last 3 years, I have grown as a researcher and as person in ways I could not have imagined. This PhD has been a scientific and human adventure, and I am extremely grateful for the support I have been given along this hard but very rewarding journey. I have been supported by many great people, and I hope that I will be able to pay a well-deserved tribute to all of them here. I have done my best to stay concise in the rest of this manuscript, so I may be a bit lengthier here.

First of all, I would like to thank the members of my jury for giving me the honor of examining my research work. I thank Antonio Rausell and Aurélie Cobat, who have accepted to read and report my thesis manuscript; Gaël Nicolas and Gérald Le Gac for agreeing to be part of my jury. I thank Marie de Tayrac as well for being part of my jury and my CSI. I thank Matthieu Bouaziz as well, who was also member of my CSI. Both of you have offered me very valuable insights and comments on the advancement of my PhD.

I would like to give a very heartfelt thank you to my last jury member, Don Conrad. I could not be more grateful for your mentorship during the course of my PhD. Thank you for following my work consistently, giving me very valuable feedback and advice, and welcoming me for 3 months in your lab in Portland. I will miss our early morning runs in the beautiful sceneries of Oregon. This visit to Portland and the people I met there have most definitely transformed my life for the better. I thank all of the people from the Conrad Lab who became more than colleagues in such a short time and made me feel like home away from home. I thank Katinka especially, who helped me immensely with all of the paperwork and made my visit to the Conrad Lab possible. Thank you to the frenchies Clément and Aude, who welcomed me from day one, Caitlin, who was the most brilliant and thoughtful friend, Antoni, who was such an enjoyable presence every day in the lab, Helena and little Auggie, whom I miss already, Ana, Eisa and Marianna, who brought me along for fun paddleboarding sessions, and Helen, Jasvinder and Arvand, who were always there to help me or chat.

All of my gratefulness goes to Emmanuelle Génin and Gaëlle Marenne, who have been great thesis advisors and mentors way before I even began this PhD. Emmanuelle, thank you for trusting me and offering me an internship in your team while I was still in my second year of medical school and did not have any notion of programming. Without you, I may not have chosen the fields of statistical genetics and bioinformatics, which have since then become my passion. Thank you for your advice, which has always been invaluable, and for challenging me to grow as a researcher. Gaëlle, thank you for your continuous presence and support. Since my first internship, you have taught me so much and reassured me during difficult times. I could not have imagined doing this PhD under a better supervision and I could not thank the both of you enough for that. I hope I have been up to the trust you have put into me.

I could not be more grateful for the lab members and colleagues I have shared time with during these last few years. I want to start by thanking the people I have been lucky to share an office with during these last 3 years, and who have supported or witnessed my addiction to coffee. Cloé and Salomé, thank you for being my daily dose of happiness for almost 2 years. We have shared so much during this time, in the ups and downs of our respective journeys. Preparing our office for the Christmas Competition of 2022 will forever be one of my favorite memories from the lab. I was very sad to see both of you go, but I wish you the best in your respective futures.

In the last few months of my PhD, I had the great pleasure of sharing the office with two other colleagues. Maël, it was a great pleasure to get to know you more during the end of my PhD. I wish you the best for your own PhD, I have no doubts you will do great. Marc, thank you for being a kindred spirit as well as a fellow MD-PhD student. I cannot believe I only got to spend a few months in your company, but they have been worth it. You made me feel extremely understood and I am sure that all of our laughs have added years to my life expectancy. I believe whole-heartedly that you will succeed in any path you choose, and best of luck for the PhD.

I would like to thank deeply the other members of the Bioinfo-Genstat team for being such great colleagues on a day-to-day basis. Thank you Anthony, for your insightful comments especially in statistics; Thomas, for being an invaluable help when I was struggling on the servers; Gaëlle LF, for being so helpful despite administrative burdens; Aude, for allowing me to have a great first teaching experience; Kevin, for taking time to talk with me about clinical genetics; to Lourdes, for always brightening my day with your good mood; and Véronique, for bringing even more joyful vibes to the team. Thank you also to the colleagues who left the lab during my PhD, especially Ozvan and Rim. I will keep a very fond memory of our time together. A special thanks to my running buddies as well, especially Tanguy and Sophie. I will miss our midday excursions and chats very much.

I want to extend a general thank you to all of the UMR 1078 members for welcoming me so well within the lab. I will miss this cheerful and supportive work environment very much.

I would also like to thank the people who collaborated with me during this PhD without whom this work would not have been possible, including the GENETWORK4DIAG project members Elisabeth Tournier-Lasserve, Chaker Aloui, Anne-Louise Leutenegger, Anaïs Baudot and Ozan Ozisik. Your expertise and our exchanges have brought a depth to this work that would not have been achieved otherwise.

On a more personal note, I would like to thank my friends and family for their continuous support, both during happy and challenging times. This has truly been a team effort. Tim, thank you for putting up with me every day and always supporting my unconventional studies. I have no doubts that you will succeed in your own PhD. I also want to thank Tamara, for always being up for cooking experiments and overall being a wonderful friend; Laos, for having the best although debatable jokes to lift my spirits; Alice, and all of the Hellfest team, for sharing my love for metal music; and Aurélien, for teaching me C++ and chess when I needed to get my mind off writing this manuscript.

A very warm thank you to my parents for always following closely my studies and being my unconditional supporters not matter the path I chose. I would not be where I am today without your education and unwavering faith in me.

Those who know me very well will not be surprised by this last personal thank to Tethys, my lovely cat, who has been a great mental health support during this PhD despite eating my plants regularly and waking me up during the night.

Finally, I am extremely grateful for the Ecole de l'INSERM Liliane Bettencourt (EdILB), without which I would have not done this PhD at this time in my life, and their directors at the time I joined, Boris Barbour and Eric Clauser. Being part of the EdILB family has quite literally changed my life and has also allowed me to meet wonderful people I am lucky to call friends. Among them, a special thanks to Clémence, Imran, Florian and Henri, who have been there for me at times when I needed it the most. All of my gratitude also goes to Barro Sow, Christine Tanga and Pierre-Yves Holtzmann, who have managed the administrative gestion of the EdILB. I am also thankful for the Bettencourt Schueller Foundation and their continuous support to the EdILB.

Scientific Productions

PUBLICATIONS

M-S. Ogloblinsky, The FrEx Consortium, D. Conrad, E. Génin, G. Marenne

Computational methods to detect digenism in sequencing data: a comprehensive review and benchmark

In preparation

M-S. Ogloblinsky, D. Lewinsohn, M. Nguyen, L. Velo-Suarez, A. Herzig, T. Ludwig, H. Castillo-Madeen, D. Conrad, E. Génin, G. Marenne

Easy-PSAP: an integrated workflow to prioritize pathogenic variants in sequence data from a single individual

In preparation

M-S. Ogloblinsky, O. Bocher, C. Aloui, A-L. Leutenegger, O. Ozisik, A. Baudot, E. Tournier-Lasserve, H. Castillo-Madeen, D. Lewinsohn, D. Conrad, E. Génin, G. Marenne

PSAP-genomic-regions: a method leveraging population data to prioritize coding and non-coding variants in whole genome sequencing for rare disease diagnosis

Submitted to Genetic Epidemiology

BioRxiv, 2024. doi: 10.1101/2024.02.13.580050

O. Bocher, Thomas E. Ludwig, **M-S. Ogloblinsky**, G. Marenne, J-F. Deleuze, S. Suryakant, J. Odeberg, P-E. Morange, D-A. Trégouët, H. Perdry, E. Génin

Testing for association with rare variants in the coding and non-coding genome: RAVAFIRST, a new approach based on CADD deleteriousness score

PLOS Genetics, 2023. doi: 10.1371/journal.pgen.1009923

ORAL PRESENTATIONS (SPEAKER UNDERLINED)

M-S. Ogloblinsky, O. Bocher, H. Castillo-Madeen, D. Lewinsohn, D. Conrad, E. Génin, G. Marenne

« Easy-PSAP: an integrated workflow to prioritize pathogenic variants in sequence data from a single individual »

July 7 2023 : AMPS congress (Paris – France)

M-S. Ogloblinsky, O. Bocher, C. Aloui, E. Tournier-Lasserve, D. Conrad, E. Génin, G. Marenne

« Utiliser les données de la population générale pour évaluer la pathogénicité de variants rares dans le séquençage pan-génomique : extension de la méthode PSAP au génome non codant »

November 8 2022 : Colloque IBSAM (Brest – France)

POSTER PRESENTATIONS (SPEAKER UNDERLINED)

M-S. Ogloblinsky, O. Bocher, C. Aloui, A-L. Leutenegger, O. Ozisik, A. Baudot, E. Tournier-Lasserve, H. Castillo-Madeen, D. Lewinsohn, D. Conrad, E. Génin, G. Marenne

« Easy-PSAP : un workflow complet de priorisation des variants pour aider au diagnostic des maladies monogéniques hétérogènes »

January 9-12 2024 : Assises de Génétique Humaine et Médicale (Paris – France)

M-S. Ogloblinsky, O. Bocher, C. Aloui, A-L. Leutenegger, O. Ozisik, A. Baudot, E. Tournier-Lasserve, H. Castillo-Madeen, D. Lewinsohn, D. Conrad, E. Génin, G. Marenne

« Using population data to assess significance of rare variants in WGS of n=1 disease cases »

November 1-5 2023 : ASHG (Washington – USA)

M-S. Ogloblinsky, O. Bocher, C. Aloui, A-L. Leutenegger, O. Ozisik, A. Baudot, E. Tournier-Lasserve, H. Castillo-Madeen, D. Lewinsohn, D. Conrad, E. Génin, G. Marenne

« Leveraging healthy population data to assess the pathogenicity of rare variants in WGS using an extension of the PSAP method »

June 10-13 2023 : ESHG (Glasgow – UK)

M-S. Ogloblinsky, O. Bocher, C. Aloui, E. Tournier-Lasserve, D. Conrad, E. Génin, G. Marenne

« Leveraging healthy population data to assess the pathogenicity of rare variants in WGS: extension of PSAP method to the non-coding genome »

September 7-9 2022 : IGES (Paris – France)

PhD Achievements

MOBILITY

August - November 2023 : **Visit to the Conrad Lab** (Division of Genetics, Oregon Health and Science University ; Oregon National Primate Research Center, Portland OR, USA)

Supported by a grant from the ED SVS and UBO

- Implementation of the Easy-PSAP variant prioritization workflow on OHSU cluster
- Application of Easy-PSAP to patient data available in the Conrad Lab
- Research of genetic variants implicated in male infertility in whole genome sequencing data from Pakistani consanguineous families
- Reinforcement of the collaboration between research teams

STUDENT SUPERVISION

July 2023 : Lucie-Garance Barot (1st year student at AgroParisTech)

Determination and evaluation of functional regions in the new build 38 of the genome from population data (continuation of Gabriel Clerempuy's work)

May – July 2023 : Gabriel Clerempuy (2nd year student at AgroParisTech)

Determination and evaluation of functional regions in the new build 38 of the genome from population data

June 2022 : Mathilde Nguyen (1st year student at AgroParisTech)

Evaluation of the PSAP variant prioritization method through artificially-simulated disease exomes

TEACHING

September – December 2022 : Graduate Teaching Assistant in Statistics for undergraduate students in Psychology

List of Abbreviations

1kGP	1000 Genomes Project
ACMG	American College of Medical Genetics and Genomics
ACS	functionally-Adjusted CADD Score
AD	Autosomal Dominant
AI	Artificial Intelligence
AR	Autosomal Recessive
bp	Base Pair
CADD	Combined Annotation–Dependent Depletion
CCR	Constrained Coding Regions
CGH	Comparative Genomic Hybridization
CSVD	Cerebral Small Vessel Disease
DI	Digenic Inheritance
DIDA	DIgenic diseases DAtabase
DM	Digenic Method
DNA	DeoxyriboNucleic Acid
ExAc	Exome Aggregation Consortium
FREX	FRench EXome project
GATK	Genome Analysis Toolkit
GEMINI	GEnetics of Male INfertility Initiative
gnomAD	Genome Aggregation Database
GRCh	Genome Research Consortium human build
GWAS	Genome-Wide Association Study
HBD	Homozygosity by Descent
HGMD	Human Gene Mutation Database
HGP	Human Genome project
HISTA	Human Infertility Single-cell Testis Atlas
HMM	Hidden Markov Model
HPO	Human Phenotype Ontology
IBD	Identical by Descent
InDel	Insertion-Deletion
LD	Linkage Disequilibrium
LOD	Logarithm of the odds
LOF	Loss-Of-Function
MAF	Minor Allele Frequency
ML	Machine Learning
NGS	Next-Generation Sequencing
OLIDA	OLIgogenic diseases Database
OMIM	Online Mendelian Inheritance in Man
OS	Open Science
PSAP	Population SAmpling Probability
QC	Quality Control
RD	Rare Disease
RNA	RiboNucleic Acid

RVAT	Rare Variant Association Test
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SVM	Support Vector Machine
TAD	Topologically Associated Domain
UTR	UnTranslated Regions
VCF	Variant Call Format
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Table of Figures

Table of Tables

Part I

General Introduction

In the world, one-third of newborns admitted to intensive care suffer from a rare disease (RD). RDs are a heterogeneous group of conditions classically characterized by a very low prevalence (less than 1 in 2,000 individuals in Europe). RDs encompass 5,000 to more than 9,000 diseases depending on the source (Ferreira 2019; Haendel *et al.* 2020; Sequeira *et al.* 2021), and this number continues to grow with new conditions being described every year. Although they may not touch every family like a common condition would – heart disease, cancer, diabetes and dementia, just to name a few – RDs often have extremely severe consequences and impose a dramatic health burden on societies around the world. While RDs are by definition rare when looked at separately, they are frequent as a whole and affect around 350 million people worldwide including 3 million people in France, which is nearly one in twenty French citizens. Many RDs affect children, 30% of whom do not survive past their 5th birthday (Wright, FitzPatrick and Firth 2018), but they can also develop later in life during adulthood. RDs are often chronic, severe and alter significantly the quality of life of patients.

Added to this public health issue is the fact that nearly 50% of RDs go undiagnosed, and when they are diagnosed, it is most often after months, or even years, of diagnostic "wandering". Diagnosis is a crucial step for patients, as it allows a better understanding of their disease and improve the medical care or therapeutics they can receive (Uhlenbusch, Löwe and Depping 2019). This elusive diagnosis is often a genetic or molecular one, as around 80% of RDs are believed to be of genetic origin (Wright, FitzPatrick and Firth 2018), with immunological, oncological and toxicological causes explaining the remainder. This commonly cited percentage of genetic etiology in RDs is sometimes challenged, with a recent review article (Ferreira 2019) describing that only 39% of RDs have a confirmed genetic origin, but the genetic basis of many RDs is undeniable.

Overall, the main challenge that I have address during my PhD project is: how can we improve the methods of analysis of a RD patient's genetic information to ultimately offer them a genetic diagnosis? My objective was to provide statistical and bioinformatic methods that could easily be applied in a clinical or research context and that would offer valuable information for RDs diagnosis.

Nowadays, we have a growing amount of genetic data at our disposal due to the progress of DNA sequencing technologies. The issue now becomes, how to analyze this data for currently undiagnosed RDs? To understand why this question is so challenging, there are a few key facts that I need to mention about RDs. The genetic architecture of RDs is very complex and can be different from one RD to another. My work has thus been focused on addressing three important factors that can explain RDs genetic diagnosis shortcomings: genetic heterogeneity, non-coding variants and complex modes of inheritance.

Indeed, many conventional paradigms of genetics, notably the one-gene-one-disease model, inadequately capture the genetic diversity observed in RDs. Most RDs are characterized by a strong genetic heterogeneity, wherein genetic variants in different genes lead to phenotypically indistinguishable diseases. The existence of multiple causative genes contributing to a singular clinical phenotype makes the grouping of patients even more challenging, often resulting in only one individual harboring a specific causal genetic variant. Moreover, the increasing availability of whole-genome sequencing data has unveiled the prominence of non-coding variants in RDs etiology, challenging traditional coding-centric approaches to genetic diagnosis. Non-coding variants can have profound effects on gene expression regulation and protein function, unraveling novel mechanisms underlying RDs. Compounding the already challenging nature of RDs genetics are complex patterns of inheritance. The main example of complex genetic inheritance explored in this work is digenism, according to which the combined effects of variants within two distinct genes are necessary to develop a disease.

In the Part II of this thesis, I present some key concepts of human genetics that are necessary to understand my work. Afterwards, in Part III, I discuss the strategies I have implemented to address genetic heterogeneity in RDs and analyze variants in the non-coding genome. I introduce the method that I have developed to prioritize variants in the coding and non-coding genome in RD patient genomes, which is named PSAP-genomic-regions. I also present Easy-PSAP, the bioinformatic workflow that I have developed and that allows the application PSAP-genomic-regions to patient data. The performance of PSAP-genomic-regions is highlighted in this part by an application to real-life cases of RDs, in consanguineous families affected by male infertility. Part IV is focused on elucidating a complex mode of inheritance of RDs, the digenic model. I review the different methods developed to detect digenic inheritance in sequencing data and benchmark some of the methods in realistic scenarios of RD diagnosis. I finish in Part V by discussing all of my results, and put them back in the context of understanding the genetic basis of RDs and using new technologies to bridge the diagnostic gap for patients affected by a RD.

Part II

Main concepts in human genetics

This part will provide the necessary background and information to discuss my results. I start by giving an overview of the genome's organization and its variations. Then, I delve deeper into the interpretation of DNA variations and their link to disease. I highlight different strategies that can be used to associate genes to diseases, and contrast them with strategies to prioritize potentially causal variant in a diagnostic setting. Finally, I focus on the specific genetic characteristics of RDs genetics and their diagnosis and mention some key facts on the two RDs that I will use as case studies in this manuscript, which are Cerebral Small Vessel Disease and male infertility.

Chapter 1 ORGANIZATION OF THE GENOME

Genetics, derived from the Greek work *genesis* meaning "origin", is the study of genes, DNA variation, and heredity in organisms. This section of the introduction aims at describing the main concepts in human genetics.

1.1 DNA STRUCTURE AND FUNCTION

1.1.1 Discovery of genetic inheritance and DNA

Modern genetics can be traced back to the discovery of **trait inheritance** by Gregor Mendel. In his study entitled "Experiments on Plant Hybridization", published in 1865, Mendel describes trait transmission patterns in pea plants. Some traits followed what he called a "**dominant**" transmission, unchanged by the hybridization. Other traits were characterized as "**recessive**", becoming latent in the hybridization process. Mendel also described an heritable substance, the "elements", which determined the expression of each trait he studied (Mendel, 1865). Although the importance of Mendel's work did not gain recognition until after his death, the mathematical equations he used to describe dominant and recessive trait transmission still hold true to this day.

Only a few years later, in 1869, chemist Friedrich Miescher made a major leap in the understanding of human genetics by describing the "nuclein", that we know as DNA (DeoxyriboNucleic Acid) today, inside the nuclei of human white blood cells. Miescher, much like Mendel, was not recognized by the scientific community of his time for his ground-breaking findings. In the decades that followed, biochemists Phoebus Levene and Erwin Chargaff expanded upon Miescher's discovery by describing the primary chemical components of the DNA molecule and the bonds that linked these components. Building upon Levene and Chargaff's work, as well as pivotal X-ray crystallography work by often-forgotten English researchers Rosalind Franklin and Maurice Wilkins, James Watson and Francis Crick described in 1953 the **double-helical structure of DNA**. This discovery would earn Watson, Crick and Wilkins the 1962 Nobel Prize in Medicine.

The genetic information is now known to be carried by the DNA molecule. A DNA molecule is a polymer shaped in a double-stranded helix (Figure 1.1). Each of the two strands of DNA is a polynucleotide, as they are composed of monomeric units call **nucleotides**. A single nucleotide is made up of three components: a nitrogenous base, a sugar (deoxyribose in the case of DNA) and a phosphate group. The nitrogen-containing base, or **nucleobase**, can either be a purine (Adenine [A] and Guanine [G]) or a pyrimidine (Cytosine [C] and Thymine [T]). Bases are paired by hydrogen bonds, thus connecting the two strands of DNA, in the following way: [T] with [A], and [C] with [G]. Nucleotides themselves are linked by covalent phosphodiester bonds between the sugar of one nucleotide and the phosphate group of the next nucleotide. The ends of DNA strands are said asymmetric, and have a **directionality** of five prime (5') end and three (3') prime end, with the 5' end having a terminal phosphate group and the 3' end a terminal hydroxyl group.

In the nuclei of the human cell, DNA is structured in 23 pairs of **chromosomes** (22 autosomal pairs and 1 pair of sexual chromosomes, also known as gonosomes). The complementary nature of base pairing ensures a redundancy of the genetic information encoded by each strand of DNA. The term **base pair** (bp) is used when referring to the nucleotides present on each chromosome at a certain position in the DNA molecule. A specific position on a chromosome is also known as a **locus**. The total length of the human genome is of 6.37 billion bp for females and 6.27 billion bp for males. The **genotype** of an individual in a particular genetic location refers to the nucleotides found on each chromosome of the pair. This genotype can be **homozygote** if the two nucleotides are the same, or **heterozygote** if not. The two possible nucleotides at a position are called alleles. A sequence of consecutive alleles on a particular chromosome is known as a **haplotype**.

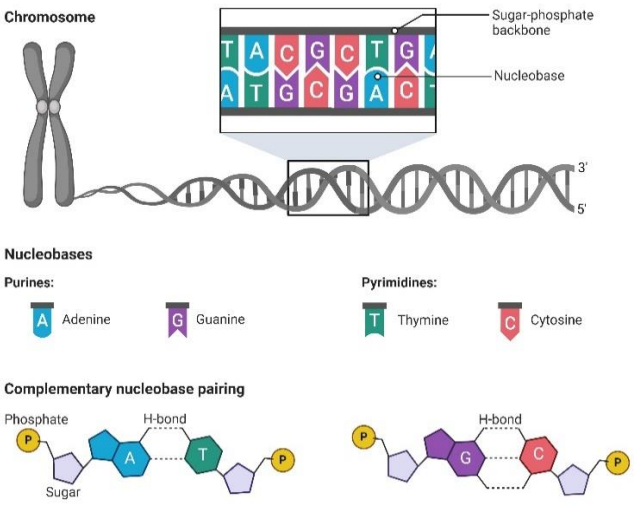


Figure 1.1 : Chemical structure of DNA

Adapted from BioRender

1.1.2 Physical structure of DNA

Knowing the sequence of DNA is key to understanding its function, but DNA must be looked at in the context of its physical organization in the cell. To fit in the nuclei, DNA needs to be highly compacted. The compact complex formed by DNA and proteins is called **chromatin**. The **euchromatin** is active and organized in nucleosomes, while the **heterochromatin** is extremely condensed and not very accessible. The localization and level of compaction of DNA thus determines its activity. **Nucleosomes** are octamers of proteins called histones, around which 150 to 200 bp of DNA is wrapped (Figure 1.2). Nucleosomes form 10 nm "beads" on a string: this is the form of euchromatin. The H1 histone allows a supplementary compaction of nucleosomes in 30 nm **fibers**: this is the structure of heterochromatin. The fibers themselves coil to create chromosomes during the metaphase step of meiosis and mitosis.

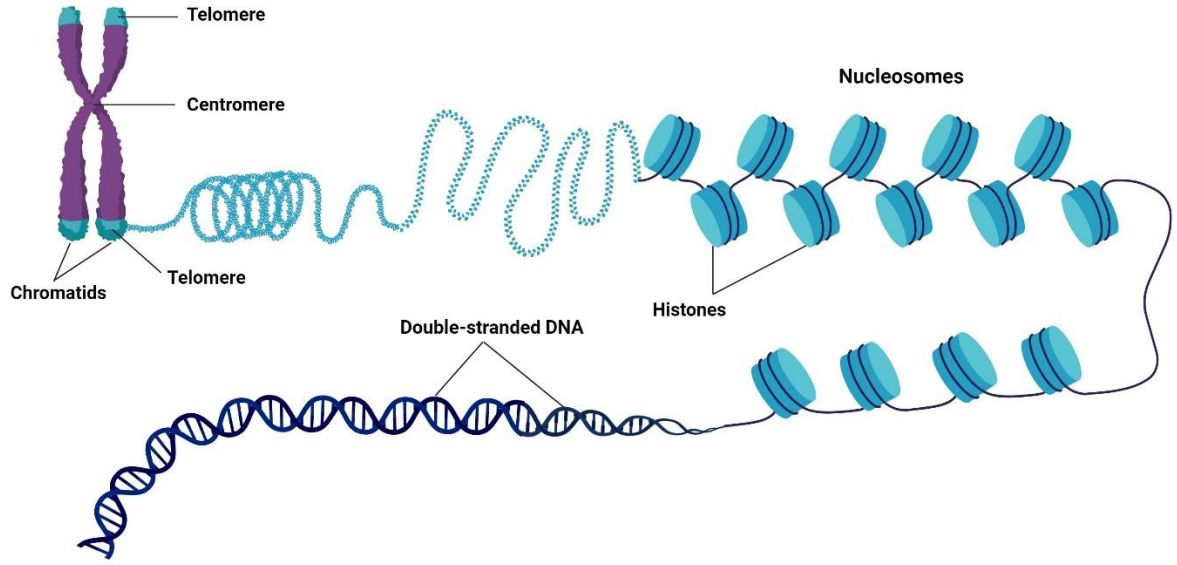


Figure 1.2 : DNA organization

Adapted from BioRender

The state of chromatin is dynamic: it depends on the stage of the cell cycle and the action of enzymes responsible for the remodeling of chromatin. These enzymes can methylate or acetylate different structures, including histones, that will alter the condensation of chromatin. This process of chromatin structure alteration is called **epigenetic**, as it alters cell function without changing the DNA sequence. When chromatin is highly compact, it is "**inactive**" or "turned off". In contrary, a less compact and open chromatin is "**active**" or "turned on". The local structure of chromatin depends on the specific genes that are found in the region.

It is also important to consider the genome in terms of 3D organization and interactions. DNA is now known to be organized in **Topologically Associated Domains** (TADs, Figure 1.3). These domains are characterized by more chromatin contacts, and can encompass multiple genes and their regulatory elements that will be described in the next paragraphs. Within a TAD, these elements can interact much more easily despite being physically distant, whereas interactions are much more limited between different TADs (Dixon *et al.* 2012). TADs are separated by regions of 300 to 2000 bp, known as **insulators**, which limit contacts between TADs (Ong and Corces 2014). This 3D organization of the genome heavily influences the impact of any variation in the DNA sequence on the function of the cell.

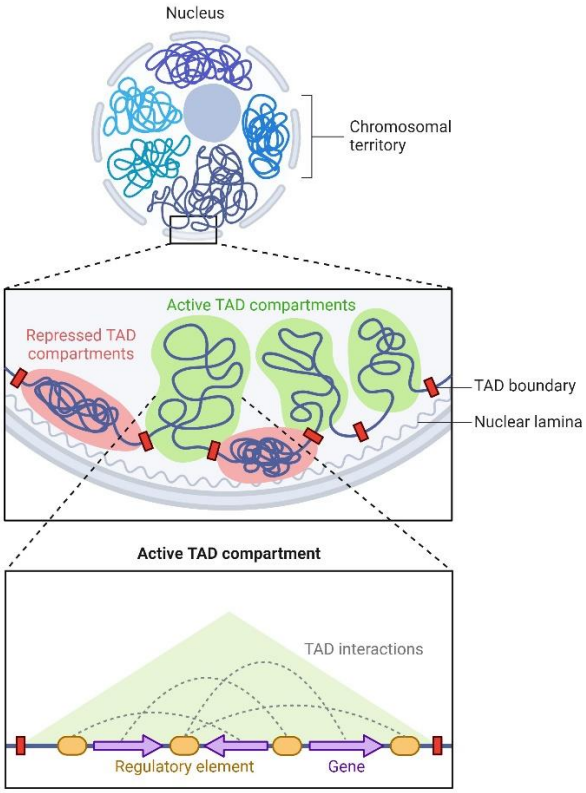


Figure 1.3 : Topologically Associated Domains

Adapted from BioRender

1.1.3 Coding genome: from the DNA sequence to proteins

A gene is the basic unit of heredity, and usually encodes information that allows the synthesis of a functional product in the cell. A gene is composed of several elements, including exons and introns (Figure 1.4). **Exons** are the coding parts of genes, meaning they encode the information that will ultimately create proteins, unlike **introns** that will not be translated into a protein. The genetic information contained in coding parts of genes or exons constitutes the **exome** of an individual, while

the entire genetic information of an individual constitutes their **genome**. Around 26,000 genes can be found in nucleic DNA, within the previously described euchromatin.

In the process of creating proteins necessary for the life of the cell, DNA is first transcribed into **pre-messenger ribonucleic acid** (pre-mRNA) by an enzyme called a RNA polymerase. RNA is also a kind of nucleic acid. Unlike DNA, RNA is a single-stranded molecule, contains ribose in place of deoxyribose and Uracile in place of the Thymine base. The pre-mRNA is then spliced and processed to make a **mature mRNA**. During the **splicing** process, introns are removed and exons are joined together. From the same gene, multiple mRNA with a varying composition of exons can be created by a process called alternative splicing. The mRNA is characterized by a 5' cap and a 3' poly-A tail. Finally, mRNAs are exported from the nucleus to the cytoplasm to be translated into protein by ribosomes. The genetic sequence of mRNAs consists of ribonucleotides, which are arranged by three to form a **codon**. In this **translation** step, each codon is associated with an amino acid, except stop codons which terminate protein synthesis. The genetic code is **redundant** but not ambiguous: one codon codes for only one amino acid, the building blocks of proteins, but multiple codons can code for the same amino acid. During translation, the **5' UTR** and **3' UTR** (UnTranslated Regions) parts of the mRNA are not translated, as their name indicates.

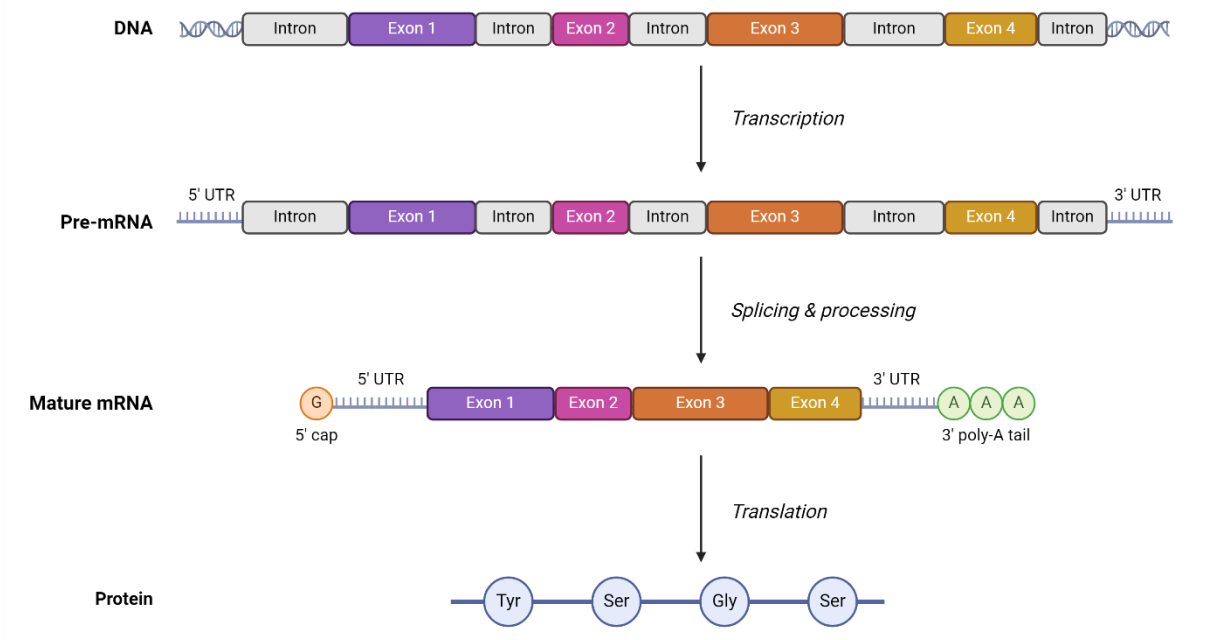


Figure 1.4 : From DNA to protein

Adapted from BioRender

1.1.4 Non-coding genome and gene regulation

The coding genome only makes up about 2% of the total genome, meaning the rest and biggest part of the genome is referred to as "**non-coding**". The totality of the heterochromatin is non-coding, as well as the majority of euchromatin. In each cell, only around 20% of genes are expressed. This expression depends on multiple factors, like the type of cell, metabolic signals or the state of differentiation of the cell. The **regulation of gene expression** can intervene at different steps: transcription, post-transcription or post-translation. Here, we will focus on the genomic elements modulating gene transcription (Figure 1.5). This regulation involves a combination of proteins present at a specific moment in the cell (**trans factors**) and binding sites on the DNA sequence (**cis factors**). These regulatory elements have been described and annotated by several big projects like FANTOM5 (Forrest *et al.* 2014) or ENCODE (Dunham *et al.* 2012). This knowledge has informed our understanding of the impact of genetic variations in such regulatory regions and their link with human disease, a key problematic that will be explored in the rest of this manuscript.

Cis regulatory elements can be separated in two main categories: proximal and distal control elements. The main **proximal** regulatory element is the promoter of the gene and is typically located in the 5' UTR part of the gene. The **promoter** controls the initiation a gene's transcription. The proximal regulation of gene expression also involves introns, as variations in key regions of introns can lead to modulations of the splicing process (Chong *et al.* 2019). More **distal** regulatory elements include enhancers and silencers. These elements are mostly located in 5' of the transcription initiation site, but can also be found in 3' of the gene or within introns. **Enhancers** are activators of the transcription when linked to a trans factor, while **silencers** have an inhibitory effect with the action of trans factors (Kolovos *et al.* 2012). The previously described TADs can encompass multiple genes, their promoters and enhancers, and help to explain the interactions between different and sometimes distant regions of the DNA. The transcriptional activity of a gene can also depend on epigenetic factors, mentioned earlier, that will make cis elements more or less accessible to trans elements.

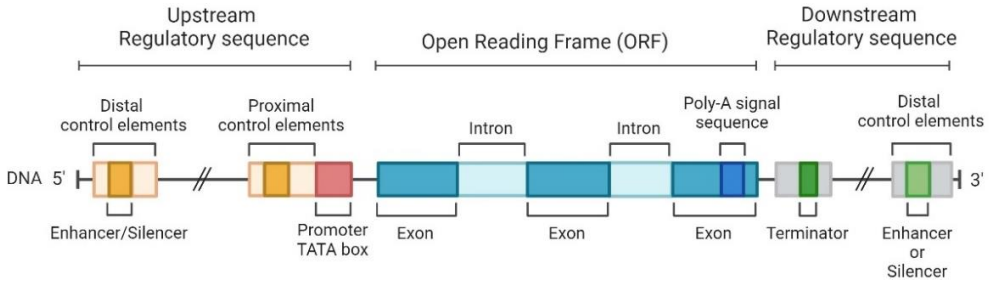


Figure 1.5 : Regulatory and genic elements

Adapted from BioRender

1.2 DNA VARIATIONS

1.2.1 Variant types

Despite having a mostly consensus human reference genome, it is evident that every individual is unique. This polymorphism in the human population is caused by stable modifications of the DNA sequence, called mutations or **variants**. A variant can be passed down to cell descendants (**somatic variant**) or human descendants (**germline variant**). Variations in the human genome can be separated in two categories: sequence variants and structural variants. **Structural variants** involve big chromosomal rearrangements, while sequence variants alter the DNA sequence at a smaller scale. In this work, we will focus on germline sequence variants when the term "variant" is used.

Sequence variants are modification of one or a few bases of the DNA compared to the reference sequence. When the reference nucleotide is substituted with another nucleotide, the term **Single Nucleotide Variant** (SNV) is used, or point mutation. For instance, a [T] nucleotide can be replaced by a [G] in some individuals, the **reference allele** being [T] in that case and the **alternative allele** being [G]. There can be multiple alternative alleles or variants at a single position. **InDels** are insertions or deletions of a few nucleotides in the DNA sequence. InDels can be distinguished from structural insertions and deletions by their smaller length, usually inferior to 50 bases or 1 kilobase depending on the definition.

Variants are characterized by their frequency in the general population. The **MAF** (Minor Allele Frequency) of a variant refers to the frequency at which the less common allele occurs in a given population. If a SNV is sufficiently frequent, present in more than 5% of individuals in a population, it can be referred to as a **Single-Nucleotide Polymorphism** (SNP). SNVs with a frequency of 1-5% are usually known as low-frequency variants. Variants that are found in less than 1% of the general population are called **rare variants**.

1.2.2 Variant consequence

The potential consequences of variants depending on their location is represented in Figure 1.6. When looking at SNVs in the coding sequence of a gene, a variant can be **synonymous** if the change in nucleotide results in a codon that leads to the same amino acid as in the reference, due to the redundancy of the genetic code. A variant is called **non-synonymous** if it leads to an alteration of the amino acid sequence of the protein. **Missense** variants change the amino acid in the protein sequence. This change in amino acid can be **conservative** if the new amino acid has the same chemical properties as the initial amino acid or **non-conservative** if not. **Non-sense** variants have a more important impact on the protein structure. Non-sense variants either create a premature stop codon (**stop-gain**) leading

to a truncated protein, or remove a transcription start codon (**start-loss**). There can also be **start-gain** and **stop-loss** variants, leading to the gain of a transcription start codon or loss of a stop codon, respectively.

The consequence of InDel variants will depend on several factors. If the length of the InDel is not divisible by 3, it is called a **frameshift variant** as it changes the reading frame (grouping of codons). The earlier (i.e. close to the beginning) in the coding sequence the frameshift variant happens, the more it affects the protein by changing all subsequent amino acids in the protein sequence. Frameshift variants can also modify the stop codon, creating an altered and shortened or lengthened protein. InDels with a length divisible by three are called "**in-frame**" and lead to the insertion or deletion of one or more amino acids but not the change of the entire protein sequence afterwards.

A protein also has to be understood in terms of domains, that allow the protein to fulfill its biological functions. This domain structure of proteins is crucial to keep in mind while evaluating the functional consequences of a variant, as it depends on the domain it affects. A **protein domain** is a distinct functional or structural unit within a protein molecule. Proteins can contain one or multiple domains, each of which may have a specific function. For example, a protein involved in cell signaling might contain a domain responsible for binding to a specific molecule, another domain for catalyzing a chemical reaction, and yet another domain for interacting with other proteins. Protein domains are documented in several databases, like Pfam (Mistry *et al.* 2021).

Variants outside of the coding sequence of a gene are more difficult to interpret in terms of consequence, as they do not as directly impact the structure of a protein as variants in the coding sequence. A variant in the promoter of a gene may affect its expression. Variants affecting the splicing process, also known as **splicing variants**, can occur in both introns and exons. They can modify existing splice sites or create new ones, leading to a disruption of the splicing process (Anna and Monika 2018). Splicing variants can lead to improper intron removal or retention, thus altering the open reading frame and ultimately the amino acid sequence. Other types of **non-coding variants** with a more difficult interpretation include intronic, 5' and 3' UTR, upstream gene, downstream gene, regulatory and intergenic variants.

Several tools, like VEP (McLaren *et al.* 2016) or Annovar (Wang, Li and Hakonarson 2010), have been developed to assign functional information to DNA variants, as well as other information about variants like their frequency in general population databases. This step is called **variant annotation** and is usually performed using as input a VCF file, which will be described in section 1.3.2. We can also note that a gene can have several **transcripts**, due to alternative splicing and other mechanisms. A variant can thus be annotated to more than one transcript and have different consequences depending

on the transcript. The full scope of a variant's consequence can be underestimated if it is not associated to the correct transcript, which heavily depends on the set of transcripts and annotation tool chosen (McCarthy *et al.* 2014).

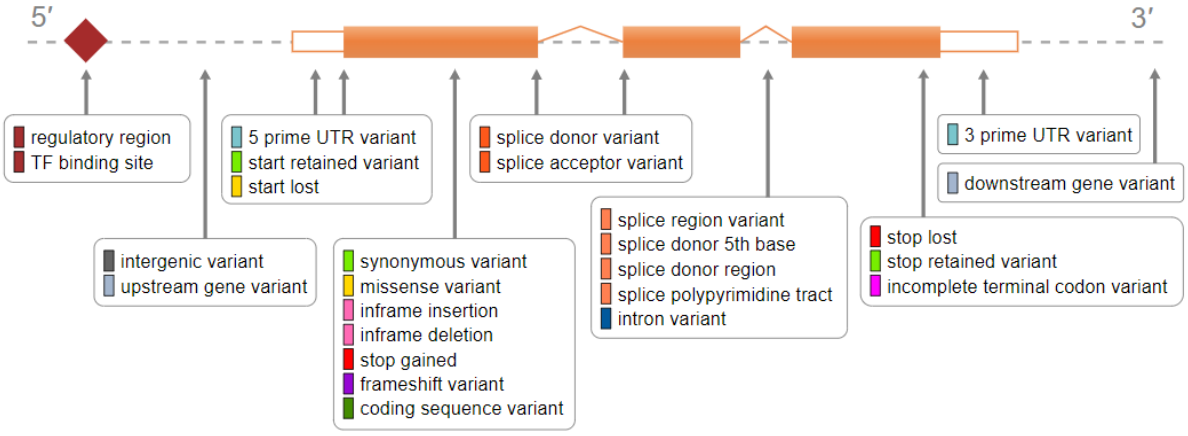


Figure 1.6 : Variant consequences

Adapted from ensembl.org

1.3 DNA SEQUENCING

1.3.1 Sequencing technologies

Being able to decipher the genetic information of an individual has been a challenge since the discovery of DNA. The development of new technologies in the field of DNA sequencing has allowed, in less than 50 years, to go from the sequencing of a few thousand bp of DNA to sequencing thousands of whole genomes a year with the latest sequencing machines. In 1977, English biochemist Frederick Sanger was the first to sequence a full DNA genome (Giani *et al.* 2020) of bacteriophage ϕX174. The "**Sanger method**" of sequencing relies on four separate polymerization reactions using normal deoxynucleotide triphosphates and a small quantity of modified dideoxynucleotide triphosphates, which terminate the DNA strand elongation process. To each reaction, only one of the four dideoxynucleotides is added while the other three are normal deoxynucleotides. The final DNA sequence is deduced by comparing the length of the DNA fragments produced by each reaction. The Sanger method can produce DNA sequence reads of less than 500 nucleotides with a very low error rate, and is still used today for the validation of sequencing results. However, it is not suitable for large-scale sequencing projects.

Over the next 40 years, the Sanger method stayed the gold standard for DNA sequencing. Improvements like the idea of "**shotgun sequencing**" allowed the sequencing of longer genomes in

shorter amounts of time. The **Human Genome project** (HGP), officially launched in 1990, aimed at producing the complete sequence of the human genome. The completion of the HGP took 14 years and the joined efforts of thousands of researchers, with a final cost of $2.7 billion. The HGP lead to the publication of the first **human reference genome** (International Human Genome Sequencing Consortium 2004), which has been continuously improved upon since then by the Genome Reference Consortium. One of the latest version or "build" of the human reference genome, is the **GRCh38** (Schneider *et al.* 2017) (for Genome Research Consortium human build 38) or hg38 (for Human genome build 38). Compared to the previous **GRCh37** (Church *et al.* 2011) build (also known as hg19), the hg38 build is characterized by 178 regions containing 261 alternative locus sequences, in highly variable regions of the genome, collectively representing 3.6 million bp of novel sequence and over 150 genes not represented in the primary assembly (Jäger *et al.* 2016).

With the drive initiated by the HGP, came the advent of **Next-Generation Sequencing** (NGS). NGS allowed the next breakthroughs in DNA sequencing and has shaped the field of genetics as we know it now. Unlike Sanger sequencing, NGS, also known as massively parallel sequencing, sequences multiple fragments of DNA simultaneously. Among the major developments that allowed the NGS revolution, there is the introduction in 1996 of **pyrosequencing** by Mostafa Ronaghi, a "sequencing-by-synthesis" method that was latter on refined by the use of fluorescent dyes.

NGS includes the following steps (Figure 1.7):

- **Library preparation**: the individual's DNA is fragmented and associated with adapters, which are known DNA fragments. These adapters vary among library preparation kits.
- **Library amplification**: the library is hybridized on the sequencing support, called "flow cell", using complementary fragments to the adapters. These adapters also contain binding sites for the sequencing primers, allowing DNA amplification. Multiple bridge amplifications through polymerase chain reaction create numerous copies of each fragment in the same area of the flow cell. These groups of the same fragment are called clusters.
- **Sequencing**: each fragment on the flow cell is "read" base by base by a polymerase that adds fluorescent nucleotides ([A], [T], [G] or [C]). Each addition of a nucleotide sends a fluorescent signal that is detected by the sequencer. This step is known as "calling", or base identification.

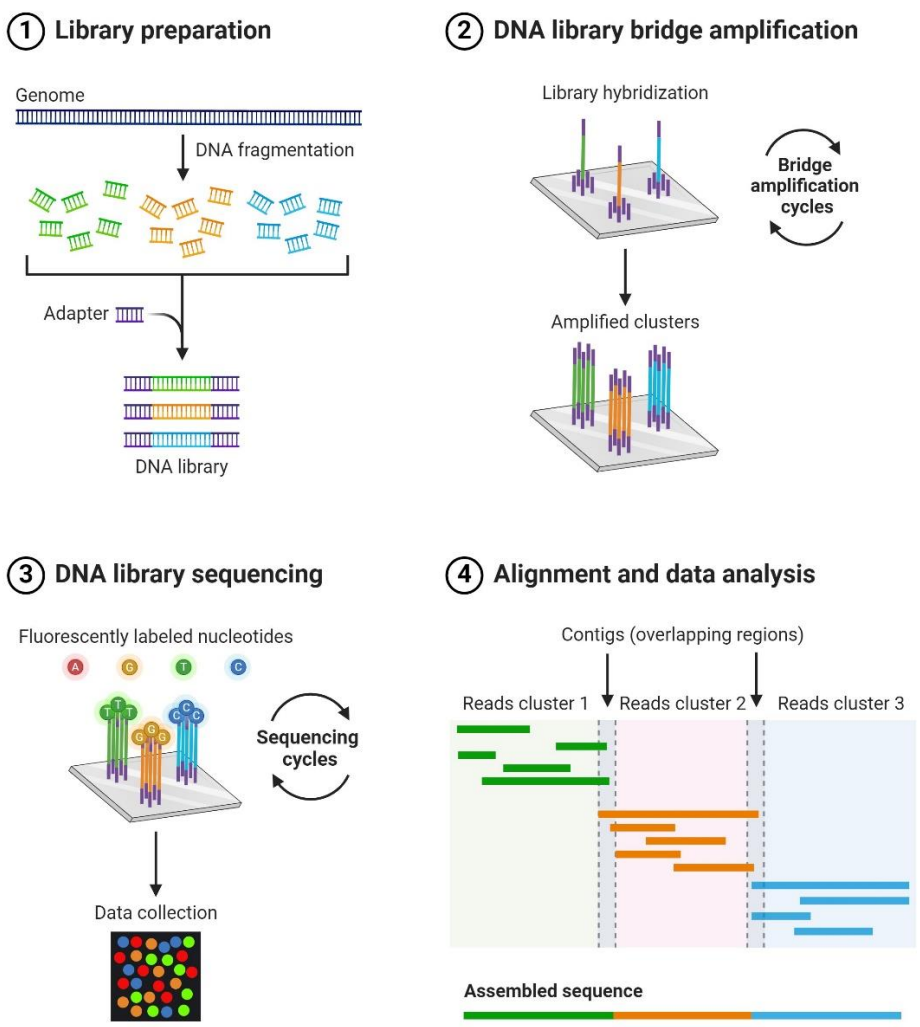


Figure 1.7 : NGS workflow

Adapted from BioRender

1.3.2 Processing sequencing data

A bioinformatic analysis is necessary to process the data coming from sequencing and allow the reconstitution of the genome. A **fastq file** is created with all of the sequenced fragments, also called "**reads**", from the base calling step. In most cases, reads are then "**aligned**" to the human reference genome using different algorithms that match each read to the most similar sequence of the reference genome. This step is also known as "**mapping**" and gives from which precise location in the genome each base pair in each read comes from. Another alternative to mapping is to use a de novo assembly, which is not covered here. The number of reads matching each position gives the "**depth**" at which the position is covered, due to the amplification step performed during the sequencing process. This depth of coverage depends on a number of factors including how complex or repeated the region is. The alignment phase produces **bam** files.

Once the sequences are aligned to the reference genome, the **variant calling** step uses those mapped reads to identify genetic variations and the genotypes of individuals at different locations in the genome. This step involves identifying positions where the sample mapped genome differs from the reference genome. Variant calling algorithms use statistical models to distinguish true genetic variations from sequencing errors and alignment artifacts, considering factors such as read depth, base quality scores, mapping quality, and strand bias. The variant calling algorithm calculates the **likelihoods of different genotypes** for the individual at each position, depending on the number of reads for each allele, and then assigns the genotype with the highest likelihood. Once variants are identified, they are typically assigned quality scores and filtered based on preset criteria to remove low-confidence calls and potential false positives.

At the individual or cohort scale, variants can be reported using the **Variant Call Format** (VCF) to facilitate their annotation and further analysis. The VCF is a standardized file format used in genetics to represent genetic variations identified through sequencing technologies. This format is widely used in the field of genomics for storing and sharing information about genetic variants, including SNVs, InDels, and structural variants. A VCF file is comprised of two main components. The **header** provides metadata about the VCF file, including information about the reference genome used, sample IDs, file format version, and other relevant details. The **Variant Call Records** constitute the main body of the VCF file. The first eight columns of the VCF records (CHROM, POS, ID, REF, ALT, QUAL, FILTER and INFO) represent the properties observed at the level of the variant site, whilst the remainder of the columns contain sample-specific information. Among the information in the VCF file, the **QUAL** field corresponds to the quality score representing the confidence in the variant call, the **FILTER** field indicates whether the variant passed certain quality control filters and the **genotype fields** give the genotype of each sample at the variant locus, and can also include information about allele depth and genotype quality.

The VCF format offers a certain flexibility by allowing the user to add annotation data within the "**INFO**" column of the file, which is itself composed of several fields separated by ";" and described in the header. The VCF file also contains key information used for **Quality Control** (QC), which is an essential step to mitigate potential artifacts or errors introduced during sequencing, alignment, and variant calling processes before any downstream analysis. In this manuscript, we have used the R package RAVAQ (Marenne *et al.* 2022), developed by Dr Gaëlle Marenne from our team at UMR1078 in Brest, to perform the QC of all of the VCF files used in subsequent analyses. All of the QC parameters are extensively described in the corresponding article.

The main function of the RAVAQ package allows a QC with 3 main steps, with the following default parameters:

- **Variant QC**: genotypes with allele depth < 10 or genotype quality < 20 are set to missing. Standard Genome Analysis Toolkit (GATK) (McKenna *et al.* 2010) hard filtering criteria are used for filtering. GATK is an industry standard genomic analysis toolkit used for variant discovery from NGS DNA sequencers. In addition, there is filtering on the mean allele balance computed across heterozygous genotypes (ABhet range 0.25–0.75) and call rates (CR > 90% in each group). Call rates homogeneity across groups is also tested by a Fisher's exact test (p-value > 0.001).
- **Sample QC**: four criteria are looked at during the sample QC to account for potential sample contamination. A sample can be excluded if it has a missing rate > 10%, an inbreeding coefficient > 4%, outlier for the heterozygous rate and outlier for the median relatedness measure.
- **Allele QC**: in this step, multiallelic variants are split so that each allele makes one line of a new VCF file. The allele QC process is then similar to the variant QC.

Chapter 2 INTERPRETING GENETIC VARIATIONS

As described in the previous section, DNA variations can arise through different processes and be passed on from parent to offspring. These variations in our genetic code are essential for our evolution and long-term survival. However, a very small percentage of genetic variants can also lead to disease. This section will outline several ways to interpret DNA variations, especially in relation to disease.

2.1 GENETIC FACTORS IN HUMAN DISEASE

2.1.1 Key principles in genetic epidemiology

The role of genetic factors in determining health and disease within populations, in conjunction with environmental factors, is studied by the field of **genetic epidemiology** (Panoutsopoulou and Wheeler 2018). Different types of diseases can be defined depending on the role of the genetic factors in the manifestation of the disease. **Association** refers to the statistical relationship between a genotype and a particular disease of interest. **Penetrance** of a genotype corresponds to the probability of expressing a certain disease given the genotype. Additionally, a **phenocopy** refers to instances where individuals exhibit a phenotype without harboring the corresponding risk-inducing genotype. Finally, **heritability** measures how much of the phenotypic variance is attributable to genotypic variance (Robette, Génin and Clerget-Darpoux 2022). Finding the genetic factors contributing to this genetic component aids in understanding the extent to which genetics influence trait variation or disease susceptibility within a population.

2.1.2 Genetic basis of monogenic versus multifactorial diseases

As mentioned previously, a significant proportion of RDs have a genetic origin. To date, most of RDs for which the genetic cause is known are described as having a **monogenic** mode of inheritance, according to which the alteration of a single gene is responsible for the disease (Ziegler 1999). Monogenic diseases exhibit **high or complete penetrance**, meaning that a large proportion of individuals with the disease-causing variant, or even all of them, develop the associated clinical phenotype.

In contrast to monogenic diseases, more common diseases like type 2 diabetes or asthma are associated with multiple small-effect and low penetrance common genetic variants. These common diseases, also known as **multifactorial diseases**, involve the interplay of multiple genetic factors (polygenic inheritance) as well as environmental factors in determining the phenotype. In the case of this **polygenic inheritance** (Crouch and Bodmer 2020), the combination of multiple variants determines a trait or susceptibility to a disease. The inheritance pattern is complex, and the phenotype

is often influenced by the combined effects of multiple genetic variants. Multifactorial diseases often involve **variable penetrance** influenced by complex genetic and environmental factors.

While in the context of monogenic diseases, the search aims to identify a gene presenting variants with high penetrance, in the case of multifactorial diseases, the goal is to identify variants conferring **susceptibility** to the disease, without being necessary or sufficient. The study of monogenic diseases has thus revealed highly penetrant pathogenic variants, in contrast to the low to medium penetrance variants that can be uncovered by the investigation into multifactorial diseases (Antonarakis *et al.* 2010).

2.1.3 Monogenic patterns of inheritance

In this work, we will focus more on monogenic rather than multifactorial diseases. Mendel's laws are crucial to understand possible monogenic inheritance patterns and Mendel himself described two of these modes of inheritance during his experiments: the autosomal dominant and recessive transmissions.

The monogenic modes of transmission can be categorized as followed (Figure 2.1):

- **Autosomal dominant**: a single copy of the variant inherited from one parent is enough to cause the disease. Affected individuals almost always have an affected parent. Each child of an affected individual has a 50% chance of inheriting the variant and therefore the disease. Autosomal dominance can be recognized in a pedigree because affected individuals are male or female, and they occur in each generation of a disease family. This is the mode of inheritance of conditions like Huntington's disease.

- **Autosomal recessive**: a single copy of a variant is insufficient to cause the disease, but the presence of two alterations, one on each copy of the gene leads to the disease phenotype. Both parents of an affected individual are usually carriers of a heterozygote variant, meaning they are not affected by the disease themselves. Each child of two carrier parents has a 25% chance of inheriting two altered copies of the gene and therefore having the disease, a 50% chance of being a carrier like the parents, and a 25% chance of inheriting two functional copies of the gene. Cystic fibrosis and sickle-cell anemia are well-known for having an autosomal recessive mode of inheritance.

    - **Homozygote recessive**: the affected individual inherits a copy of the same variant from each of his parents, and this homozygote variant leads to the disease.

- **Compound heterozygote**: each parent is a carrier of a different heterozygote variant in the same gene. The individual inherits one variant from each parent, resulting in disease expression. Neither parent exhibits the disease phenotype because they only carry one altered copy of the gene each.

- **X-linked**: the disease-associated gene is located on the X chromosome, in contrary to the previously cited modes of inheritance which involve autosomes only. This type of transmission is known as **hemizygote** for men who only have on copy of the X chromosome. Hemophilia A is an example of X-linked disease. It can be noted that X and Y chromosomes have short regions of homology named **pseudoautosomal regions** (PAR1 and PAR2) and that genes in this region are inherited in an autosomal rather than a sex-linked pattern.

  - **X-linked dominant**: in females, who have two X chromosomes, inheriting one copy of the variant is enough to cause the disease. In males, who only have one X chromosome, inheriting one copy of the variant on their single X chromosome cause the disease. Affected males pass the altered copy of the gene to all of their daughters but not their sons. Affected females can pass the altered copy of the gene to both sons and daughters.

  - **X-linked recessive**: males, having only one X chromosome, are more commonly affected. They inherit the variant from their carrier mother. Females need to inherit two copies of the variant (one from each parent) to manifest the disease, making them less commonly affected than males. Carrier females can pass the altered copy of the gene to both sons and daughters.

- **Y-linked**: Y-linked inheritance involves genes located on the Y chromosome and will not be further explored in this manuscript. These diseases are passed from fathers to all their sons.

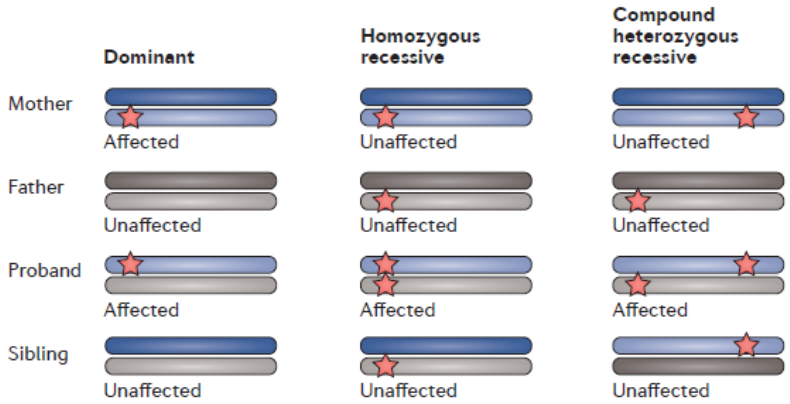


Figure 2.1 : Main monogenic modes of inheritance

From Eilbeck et al. 2017

2.2 VARIANT PATHOGENICITY

2.2.1 Damaging is not equal to pathogenic

Each healthy individual's genome contains thousands of variants, meaning that most variants have no detectable biological consequence for the cell or the organism. The real **impact** of a variant is complex to predict (Zschocke, Byers and Wilkie 2023), and a distinction has to be made between the deleteriousness and pathogenicity of a variant. **Deleteriousness** refers to the potential harmful or damaging effect of a variant. A variant is considered deleterious if it disrupts the normal functioning of a gene or protein. **Pathogenicity** specifically refers to the ability of a variant to cause or contribute to a disease phenotype, as discussed in the previous sections. So while deleteriousness assesses the potential functional impact of a variant, pathogenicity evaluates its role in causing or predisposing individuals to a disease. Variants can be deleterious without being pathogenic. For instance, for a recessive disease like cystic fibrosis, an individual will have to harbor two copies of the deleterious variant to develop the disease and heterozygote carriers are unaffected. However, in many cases, deleterious variants are also pathogenic as they disrupt critical biological processes or pathways.

As early as 2003, the SIFT (Ng and Henikoff 2003) method was proposed to evaluate the impact of amino acid substitutions on protein function. A plethora of **bioinformatic prediction tools** and scores have been developed since then to evaluate the deleteriousness and potential pathogenicity of genetic variations. Currently, one tool widely used both in research and clinics is the CADD (Kircher *et al.* 2014) score, that will be described thereafter. As we will discuss in the following sections of this manuscript, these scores are based on several annotations of the genome like the phylogenic **conservation** compared to homologous protein sequences from other organisms, or the **allele frequency** of a variant in a general population database. A caveat to these prediction of pathogenicity tools can be found in a study from 2010 (Dorfman *et al.* 2010) which assessed common variants implicated in cystic fibrosis using three popular variant pathogenicity prediction tools. They found that although the CFTR variant p.Arg75Gln is predicted damaging as it alters a very conserved position in the protein, it has a mild phenotypic effect. In contrast, the p.Val520Phe variant is known to be pathogenic but affects a non-conserved position in the CFTR protein. Pathogenicity prediction scores thus cannot be taken at face value, and have to be combined and confronted with other lines of evidence to determine the actual impact of a variant.

2.2.2 CADD pathogenicity score

Among variant pathogenicity scores, the **CADD** (Combined Annotation–Dependent Depletion) **score** (Kircher *et al.* 2014) stands out as a particularly useful score to predict the deleteriousness of variants. In contrary to most other scores, defined only for a specific type of variants like missense

variants, the CADD score provides a single pre-computed score for all possible SNVs and a large set of InDels. A recent version also allows the scoring of structural variants (Kleinert and Kircher 2022a). CADD also has the advantage of evaluating the **deleteriousness** of variants, which strongly correlates with both molecular functionality and pathogenicity, but is less likely to suffer from a major ascertainment bias by relying on a small set of genetically or experimentally well-characterized pathogenic variants.

The CADD score is a **machine learning** (ML) method that combines multiple annotations of the genome into a single score to assess the deleteriousness and thus potential pathogenicity of a variant. Initially, CADD used linear **SVMs** (Support Vector Machines) as ML model, but switched to **logistic regression** in later versions. To train the model, CADD relies on a unique framework that does not rely on a specific set of known pathogenic or benign variants, but on the contrast between "**proxy-deleterious**" and "**proxy-neutral**" sets of millions of variants and the resulting differences in their annotation features. The "proxy-neutral" variants are variants that have persisted in the human genome since the last human-ape ancestor, whilst the "proxy-deleterious" variants are simulated de novo variants that are free of selective pressure.

The initial output of the CADD machine-learning model are "**raw scores**", which summarize the extent to which the variant is likely to be from the proxy-neutral (negative values) or proxy-deleterious (positive values) class. To make the CADD score comparable between versions and assemblies of the human genome, this raw score is then converted into a "**PHRED score**" ranging from 1 to 99, based on the rank of each variant relative to all possible 8.6 billion SNVs in the human reference genome on a PHRED-scale. In the rest of this manuscript, we will be referring to the PHRED-scaled score when talking about the CADD score. A main limitation of the CADD score is the training set label given to a variant, which provides an imperfect although useful approximation of whether the variant is really benign or pathogenic. An unknown proportion of the proxy-deleterious variants are certainly neutral and could lead to a flawed prediction of deleteriousness for some variants.

Initially, **CADD v1.0** included 63 distinct annotations that spanned a variety of properties of the genome (see Table 2.1 for a full comparison of all the main version of CADD). Interestingly, the best-performing individual annotations were protein-level scores, but these evaluated only missense variants (0.63% of all training variants). In contrast, conservation metrics were the strongest individual genome-wide annotations. The next publication on the CADD score described **CADD v1.4** (Rentzsch *et al.* 2019), which is the first to start supporting the genome build GRCh38 and includes a splice score as well as measures of genome-wide variant density. **CADD v1.6** (Rentzsch *et al.* 2021), also called "CADD-splice", is the following main update of CADD and integrates two deep learning splicing models which

improve significantly the scoring of splicing variants, while not compromising performance on other types of variants. The most recent version of CADD, **CADD v1.7** (Schubach *et al.* 2024), includes new annotations like state-of-the-art protein language model scores, regulatory variant effect predictions and sequence conservation scores, that continue to further improve the performance of CADD especially in the non-coding genome.

When analyzing an individual's variants, clinicians and researchers alike often seek a **single universal cut-off value** above which a variant is considered "pathogenic". However, in all of CADD publications, the authors strongly advise against such binarization of the CADD score and recommend ranking all variants by CADD score, further investigating top-ranked variants. The CADD score should be understood as a piece of information among other lines of evidence to determine the impact of a variant. Also, it is not currently possible to precisely calibrate the relationship between the deleteriousness estimated by CADD and the likelihood that a variant is truly pathogenic.

Version	**Machine Learning model**	**Features description** (or additional features compared to v1.0)	**Reference**
CADD v1.0	Linear SVM	63 annotations • **Conservation metrics**: GERP, phastCons, phyloP • **Regulatory information**: genomic regions of DNase I hypersensitivity, transcription factor binding • **Transcript information**: distance to exon-intron boundaries, expression levels in commonly studied cell lines • **Protein-level scores**: Grantham, SIFT, PolyPhen	Kircher et al. (2014)
CADD v1.4	Logistic regression	>60 annotations • **ChromHMM** (from v1.1) • **DNA shape factor** (from v1.1) • **mirSVR** and **targetScan** (from v1.1): miRNA binding site prediction • **Mutation Index** (from v1.1) • **VEP domain annotation** (from v1.1) • **Mutation density** (from v1.4) • **dbscSNV** (from v1.1): splice prediction	Rentzsch et al. (2019)

CADD v1.6	Logistic regression	102 annotations • **SpliceAI**: splice junction prediction • **MMSplice**: exon skipping, splice site choice, splicing efficiency predictions	Rentzsch et al. (2021)
CADD v1.7	Logistic regression	>100 annotations • **ESM-1v**: Meta AI Evolutionary Scale Model for variant effects in protein coding sequences • **RegSeq**: CNN trained on open chromatin sequences of multiple tissues • **APARENT2**: Human polyadenylation • **Zoonomia**: Conservation score • **Roulette**: Mutability score	Schubach et al. (2024)

Table 2.1 : Published versions of the CADD score and their characteristics

SVM, Support vector machine ; CNN, convolutional neural network ; AI, Artificial Intelligence

2.2.3 Other main variant pathogenicity scores

While the CADD score was the main variant pathogenicity prediction score used in my work, a number of other pathogenicity scores have been developed over the years. They are sometimes called **"*in silico*" pathogenicity prediction tools**, in contrast to "in vivo" or "in vitro" experiments, as they derive knowledge from computer simulations and model analysis. The main pathogenicity prediction scores were described and compared in recent reviews (Eilbeck, Quinlan and Yandell 2017; Garcia, Andrade and Palmero 2022). Broadly, these *in silico* pathogenicity prediction tools can be separated in 5 main categories depending on the way they infer pathogenicity, although these categories are not mutually exclusive. These categories include: analyzing **sequence conservation** in both evolutionary and interspecific contexts (Ng and Henikoff 2003; Siepel *et al.* 2005; Stone and Sidow 2005; Tavtigian *et al.* 2008; Davydov *et al.* 2010; Pollard *et al.* 2010; Reva, Antipin and Sander 2011; Choi *et al.* 2012; Shihab *et al.* 2013, 2015; Gulko *et al.* 2015; Mi *et al.* 2019), evaluating **structural or physicochemical parameters** (Li *et al.* 2009; De Baets *et al.* 2012; Adzhubei, Jordan and Sunyaev 2013), employing **supervised ML** (Schwarz *et al.* 2010; Carter *et al.* 2013; Kircher *et al.* 2014; Ioannidis *et al.* 2016; Jagadeesh *et al.* 2016; Feng 2017; Steinhaus *et al.* 2021), employing **unsupervised ML** (Lu *et al.* 2015; Ionita-Laza *et al.* 2016), and analyzing modifications of **splicing** (Reese *et al.* 1997; Yeo and Burge 2004; Desmet *et al.* 2009; Jian, Boerwinkle and Liu 2014; Jaganathan *et al.* 2019). A non-exhaustive list of methods from each category, as well as methods targeted towards non-coding variants pathogenicity prediction addressed in the next section, can be found in Table 2.2.

There is not one universal best pathogenicity score for all types of variants and diseases, as evidenced by the multiple benchmarks on the subject and their conflicting conclusions depending on the variant testing set (Li *et al.* 2018; Anderson and Lassmann 2022). By the number of citations, PolyPhen, CADD and SIFT have been the most used tools in the literature (as they are also among the oldest tools), whereas the benchmarks often cite VEST, REVEL, FATHMM and BayesDel as outperforming tools (Garcia, Andrade and Palmero 2022). Another limitation of these tools is their restriction of analysis, on missense variants only for instance (including SIFT, PolyPhen, VEST and REVEL). The advantage of a tool like CADD is that it is a **meta-predictor** integrating other scores like SIFT, PolyPhen-2, phyloP and GERP and that it enables the user to score any SNV or InDel of the genome. CADD is also trained on a much larger number of variants compared to other machine-learning methods, while having a relatively modest number of features, which is why the method is not too affected by the **curse of dimensionality** according to which a large amount of training data is necessary to train a machine-learning model on large feature spaces (Ruscheinski *et al.* 2021).

Category	Software	Pros	Cons	Reference
Interspecific and evolutionary sequential conservation	SIFT	Largely used and included in several meta-predictors	Not maintained anymore, analyses missense only	Ng et al. (2003)
	Align-GVGD	Tool developed for some genes	Species limitation, analyses missense only	Tavtigian et al. (2006)
	MAPP	Compares the conservation of several physicochemical parameters	No website for analysis, for missense only, complex input format	Stone et al. (2005)
	PhastCons	Largely used and included in several meta-predictors	No website for analysis, for missense only	Siepel et al. (2005)
	PhyloP	Included in several meta-predictors	No website for analysis, for missense only	Pollard et al. (2010)
	GERP	Included in several meta-predictors	No website for analysis, for SNVs only	Davydov et al. (2010)
	Mutation Assessor	Included in several meta-predictors, uses entropy	For missense only	Reva et al. (2011)
	FATHMM FATHMM-MKL	Outperforming tool, included in several meta-predictors	-	Shihab et al. (2012, 2015)
	PROVEAN	From the same institute of SIFT, showed good performance when tested	No website for analysis	Choi et al. (2012)

		for somatic and experimentally validated variants(Li *et al.* 2018)		
	FitCons	Multi-data integration	No website for analysis	Gulko et al. (2015)
	Panther	Recently updated	Requires FASTA input, for missense only	Mi et al. (2019)
Sequence/ Structure tools	MutPred	Display possible altered molecular mechanisms, recently updated	Requires FASTA input, different algorithms for each type of variant	Li et al. (2009)
	SNPeffect	Four algorithms for specific biochemical features	Requires FASTA (or similar data) input, for missense only, complex output format	DeBeats et al. (2012)
	PolyPhen-2	Most cited tool	For missense only	Adzhubei et al. (2013)
Supervised Machine Learning Analysis	VEST	Outperforming tool	For missense only	Carter et al. (2013)
	Mutation Taster	Display several of the variant altering mechanisms	Upgraded version available	Schwarz et al. (2014)
	Mutation Taster 2021	Upgraded version	-	Steinhaus et al. (2021)
	CADD	Largely used meta-predictor	-	Kircher et al. (2014)
	M-CAP	Meta-predictor	For missense only	Jagadeesh et al. (2016)
	REVEL	Outperforming tool, meta-predictor	No website for analysis, for missense only	Ioannidis et al. (2016)
	BayesDel	Outperforming tool, meta-predictor	Requires software download to work with	Feng et al. (2017)
Unsupervised Machine	GenoCanyon	One of the few unsupervised models	-	Lu et al. (2015)

Learning Analysis	Eigen Eigen-PC	One of the few unsupervised models, good performance(Li *et al.* 2018)	-	Ionita-Laza et al. (2016)
Splicing analysis	Nnsplice	One of the first tools available	Not maintained anymore, require FASTA input	Reese et al. (1997)
	MaxEntScan	Uses entropy	Requires FASTA input	Yeo et al. (2004)
	HSF	Group several splicing algorithm	Paid tool (free credits for academic purposes)	Desmet et al. (2009)
	dbscSNV	Ensemble splicing tool	No website for analysis, SNVs only	Jian et al. (2014)
	SpliceAI	Newest tool described here based on deep learning	Doesn't support all types of InDels	Jaganathan et al. (2019)
Non-coding specific analysis	GWAVA	Meta-predictor	No website for analysis, requires software download to work with	Ritchie et al. (2014)
	FunSeq2	Weighted scoring scheme according to several annotations	Targeted towards somatic variants	Fu et al. (2014)
	DeepSea	Deep-learning based meta-predictor	Upgraded version available	Zhou et al. (2015)
	ReMM	Meta-predictor	-	Smedley et al. (2016)
	Linsight	Based on the FitCons framework, semi-supervised model	No website for analysis	Huang et al. (2017)
	ncER	Meta-predictor	No website for analysis	Wells et al. (2019)
	NCBoost	Meta-predictor	No website for analysis	Caron et al. (2019)
	FINSURF	Meta-predictor	-	Moyon et al. (2022)

Table 2.2 : Description of the main in silico predictors of variant pathogenicity

Adapted from Garcia et al. 2022 ; FASTA is a text-based format representing nucleotide sequences

2.2.4 Pathogenicity scores in the non-coding genome

The pathogenicity prediction for non-coding variants remains more challenging than in protein-coding regions, as evidenced by the limitation of most of the common pathogenicity prediction scores to missense variants. However, the potential role of non-coding parts of the genome in the etiology of disease is undeniable and currently underrated (Hindorff *et al.* 2009; Makrythanasis and Antonarakis 2013). Some of the previously cited tools are able to also evaluate the potential pathogenicity of non-coding variants, including **conservation** scores like PhastCons, phyloP, GERP, FitCons and FATHMM and **machine-learning-based** methods like CADD, Eigen or Eigen-PC. However, they often have a poorer evaluation of the pathogenicity of non-coding variants compared to coding variants due to the lack of knowledge around regulatory machinery encrypted in non-coding DNA (Eilbeck, Quinlan and Yandell 2017).

Other more recently developed methods are targeted specifically towards **non-coding variants pathogenicity prediction** through ML methods (GWAVA (Ritchie *et al.* 2014), DeepSea (Zhou and Troyanskaya 2015), ReMM (Smedley *et al.* 2016), ncER (Wells *et al.* 2019), NCBoos t(Caron, Luo and Rausell 2019), FINSURF (Moyon *et al.* 2022)), a weighted scoring scheme (FunSeq2 (Fu *et al.* 2014)) or conservation approaches (LINSIGHT (Huang, Gulko and Siepel 2017)). A recent review (Wang *et al.* 2023), although not encompassing all of the aforementioned methods, found that ncER and LINSIGH performed the best to predict the pathogenicity of non-coding variants in four benchmark datasets. Other frameworks quantify the intolerance to variation of regions of the genome, also known as **constraint**, like JARVIS (Vitsios *et al.* 2021) or ORION (Gussow *et al.* 2017). These methods follow a similar approach to other methods that looked at constraint at the gene level for coding variants, including the pLI (Lek *et al.* 2016) (probability of being loss-of-function intolerant, relative to all other genes in the human genome) or the Gene Damage Index (Itan *et al.* 2015) (mutational damage accumulated by each protein-coding human gene in the general population). However, most of these prediction methods either do not provide a variant-specific score, or are not defined in both coding and non-coding parts of the genome.

2.3 VARIANT FREQUENCY

2.3.1 The HapMap project

Allele frequency can vary widely depending on the population, due to mechanisms likes natural selection, genetic drift and gene flow. **Population databases** thus play a crucial role by providing reference data on allele frequencies in diverse populations, which allows the evaluation of how

common a variant is in this population. By identifying and cataloging these variations, researchers aimed to facilitate the characterization of human genetic diversity and its role in health and disease.

The **HapMap project** (Gibbs *et al.* 2003), short for Haplotype Map, was an international research effort aimed at creating a comprehensive catalog of common genetic variations in humans. The project was a collaboration between researchers from various institutions worldwide, including the United States, Japan, Canada, China, Nigeria, and the United Kingdom. The HapMap project aimed to identify haplotype blocks and infer patterns of **linkage disequilibrium** (LD) within several populations to avoid having to sequence all patient genomes. LD refers to the nonrandom association of alleles at different loci (Slatkin 2008). The identification of haplotype blocks revealed that LD often spans extensive chromosomal regions, implying that testing a single SNP within each block for significant association with a disease could potentially indicate association with all SNPs in that block thus decreasing the number of SNPs requiring testing in case-control studies of disease association. The project bolstered the development of genotyping platforms and methodologies for efficient SNP genotyping. HapMap was initiated in 2002 and completed in 2009. The release of HapMap datasets (Altshuler, Donnelly, and The International HapMap Consortium 2005; Frazer *et al.* 2007; Altshuler *et al.* 2010) with millions of genotyped SNPs provided a foundational resource for subsequent genetic studies.

2.3.2 Other general population variant panels

Following the HapMap project, the NGS revolution and the diminishing cost of sequencing individuals has made possible the implementation of large sequencing projects, for both individuals affected by specific pathologies and individuals from the general population. Among these projects, the **1000 Genomes Project** (1kGP) was launched in 2007 and aimed at providing a comprehensive resource on human genetic variation. The idea was to catalogue variants and their frequencies genome-wide in the studied populations, and make it accessible for the scientific community. The phase 3 analysis of the 1kGP, published in 2015 (Auton *et al.* 2015), contained low-coverage whole genome and exome sequencing data for 2,504 individuals from 26 populations in Africa, East Asia, Europe, South Asia, and the Americas. These individuals were self-reported as healthy, and were unrelated. Recently, an expansion of the 1kGP (Byrska-Bishop *et al.* 2022) made available high-coverage whole-genome sequencing data for 3,202 individuals including 602 trios parents-child (Figure 2.2). This new instalment of the 1kGP allowed the identification of more rare SNVs, as well as new InDels and structural variants. More than a resource on variant frequency, the 1kGP dataset is of particular interest for our work as it provides the whole individual genome data for all of its general population samples, which we have used to simulate disease genomes.

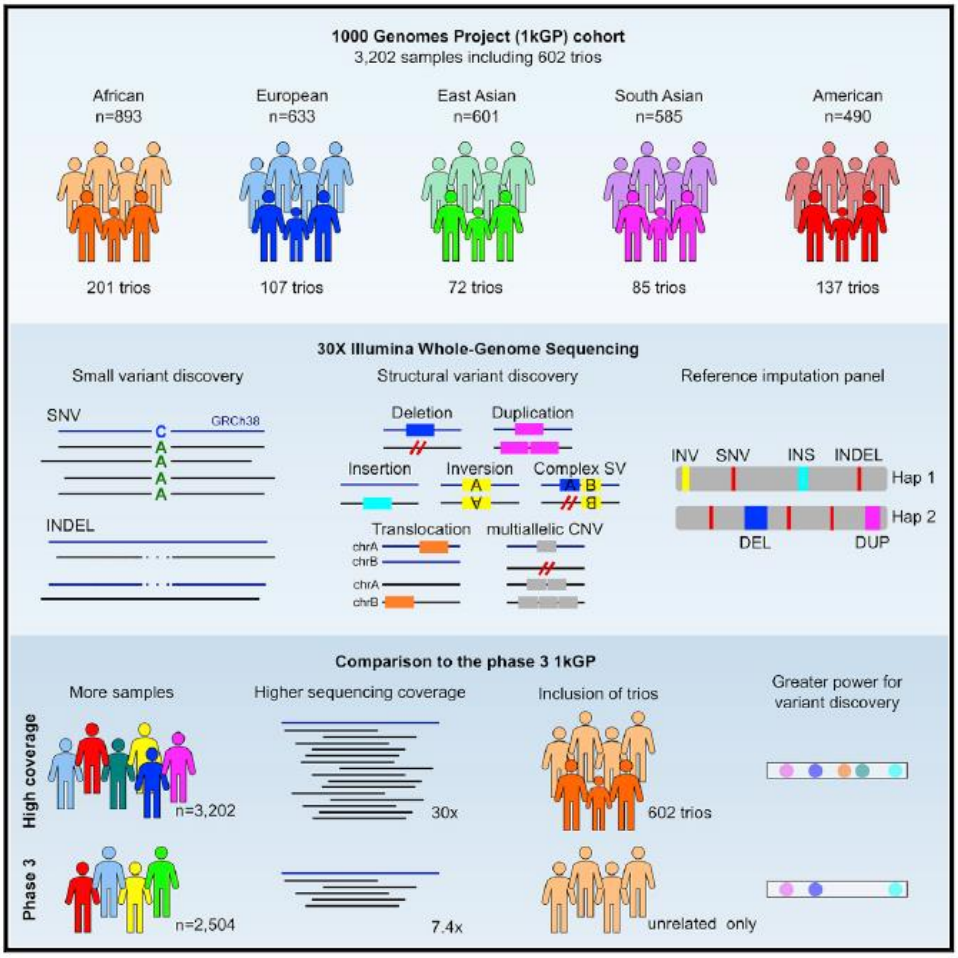


Figure 2.2 : Description of the phase 3 and high coverage 1kGP datasets

From Byrska-Bishop et al. 2022

Another of the main genomic data repositories initiatives was the **Exome Aggregation Consortium** (ExAC) (Lek *et al.* 2016), which published in 2016 a deep catalogue of protein-coding variation from high-quality exome sequence data of 60,706 individuals from 6 broad populations. In contrast to the 1kGP, ExAC included data from unrelated adults without severe pediatric disease, but that could be affected by other types of diseases. The successor to ExAC is the **Genome Aggregation Database** (gnomAD). gnomAD v2 made available in 2018 the aggregated data of 125,748 exomes and 15,708 genomes from human sequencing studies (Karczewski *et al.* 2020), for a total of 141,456 individuals (Table 2.3). In 2020, gnomAD v3 provided a dataset comprising of 76,156 genome samples mapped to the GRCh38 reference sequence (Chen *et al.* 2024). Finally, in late 2023, gnomAD released its v4 which encompasses 734,947 exomes and all of the genomes from gnomAD v3 for a total of 807,162 individuals. The gnomAD v4 exome dataset includes 416,555 individuals from the UK Biobank cohort (Bycroft *et al.* 2018), mostly from European ancestry, as well as 138,000 individuals of non-European genetic ancestry (Table 2.3). In contrast to the 1kGP project, gnomAD does not give access to individual data but aggregated variant frequencies, over the whole cohort or by population.

This information on allele frequencies over a large sample size can be key for interpreting the potential pathogenicity of a variant, and is also used in more complex variant prioritization methods discussed in section 3.2.3.

	ExAC	gnomAD v2	gnomAD v3	gnomAD v4*		
	#	#	#	#	%	Fold increase from v2
Admixed American	5,789	17,720	7,647	30,019	3.72%	1.7x
African	5,203	12,487	20,744	37,545	4.65%	3x
Ashkenazi Jewish	-	5,185	1,736	14,804	1.83%	2.9x
East Asian	4,327	9,977	2,604	22,448	2.78%	2.3x
European^	36,667	77,165	39,345	622,057	77.07%	8.1x
Middle Eastern	-	-	158	3,031	0.38%	New
Remaining Individuals^	454	3,614	1,503	31,172	3.93%	8.8x
South Asian	8,256	15,308	2,419	45,546	5.64%	3x
Total	**60,706**	**141,456**	**76,156**	**-**	**807,162**	**-**

*v4 includes all v3 samples
^ Due to small sample sizes Finnish was included in European and Amish was included in Remaining Individuals

Table 2.3 : Genetic ancestry group breakdown of ExAC and gnomAD's three main versions

From gnomAD's website https://gnomad.broadinstitute.org/news/2023-11-gnomad-v4-0/

Finally, another resource that has been used in this work is the FRench EXome project (Génin *et al.* 2017) (FREX). FREX is the first French sequencing project and aimed to create a database of genetic variations observed in the French population. This project was funded by the France Génomique infrastructure, and included the exomes of 574 individuals sampled in 6 regions of France (around the cities of Brest, Rouen, Bordeaux, Dijon, Lille and Nantes). Individuals were recruited from the general population and not on the basis of any pathology. They were selected to be representative of their geographical region: only individuals with their four grandparents born within a 30 km radius were included in the study for the Brest and Nantes centers. For the other centers, individuals had to be born in the region to be selected for sequencing. The FREX project led to the identification of almost a million variants, over 30% of which were absent from the 1kGP and ExAC databases. In line with the use of 1kGP genomes, the FREX exomes were used in this work to simulate disease exomes.

Chapter 3 METHODS TO FIND GENETIC FACTORS IMPLICATED IN HUMAN DISEASES

As presented in the two previous sections, the understanding of human diseases is closely linked to genetic variations of the human genome. In this section of the introduction, we will focus on the methods that have been developed to decipher the **genetic causes of diseases**, either by association or in a diagnostic setting.

3.1 POPULATION-BASED APPROACHES FOR GENE-DISEASE ASSOCIATION DETECTION

3.1.1 Linkage methods

Associating a portion of DNA or gene with a disease had begun way before the advent of DNA sequencing, using methods of **genetic mapping** to map the chromosomal location of a disease gene by studying patterns of inheritance within families. **Linkage analysis** (Pulst 1999) is a method that is used for genetic mapping. The idea of genetic linkage is that DNA sequences that are in close proximity on a chromosome tend to be inherited together during the meiosis phase of sexual reproduction. This can be explained by the phenomenon of **recombination** which happens during the meiosis, and during which genetic material is exchanged between homologous chromosomes. Thus, the frequency of recombination events between two loci is proportional to the physical distance between them on the chromosome. **Centimorgans** (cM) are commonly used in genetic mapping studies to estimate the distance between genetic markers or genes on a chromosome. A cM is defined as the distance between genetic loci for which one genetic crossover event (or recombination event) occurs in 1% of the gametes produced by meiosis.

By measuring the recombination frequency between markers in experimental crosses, researchers can construct **genetic maps** that provide insights into the relative positions of genes along the chromosome. The **genetic markers** used as reference points along the chromosomes were restriction fragment length polymorphisms, microsatellites (short tandem repeats) or even in later years SNPs. Linkage analysis calculates a **LOD** (logarithm of the odds) **score** for each genetic marker, which measures the likelihood of observing genetic linkage between markers and traits. LOD scores are compared to predefined threshold values to determine the **statistical significance** of linkage. Despite the power of linkage analysis to detect disease gene associations, it had several limitation including a heavy reliance on the availability of **large families** with multiple affected individuals which can be a challenge in the case of RDs. Linkage analysis is also sensitive to **locus heterogeneity** and to the predicted genetic model.

3.1.2 Genome-Wide Association Studies

With the growing availability of genotyping technologies, **Genome-Wide Association Studies** (GWAS) became increasingly popular, as they combined the genomic coverage of linkage analysis with the power of association in thousands of cases and controls (Uffelmann *et al.* 2021). They were based on SNP array data, which genotype 500,000 to 1 million of **common SNPs** covering the whole genome. GWAS allowed the discovery of genetic variants with small effects, mostly associated with common multifactorial diseases. For each SNP, a statistical test is applied to evaluate if the allele or genotype frequencies are different between cases and controls. If a genetic association is found between a phenotype and a SNP, it may mean that the SNP in question is causal for the disease (**direct association**), or that the SNP is in LD with the causal variant (**indirect association**). This process allows GWAS to capture genetic associations even without genotyping all variants of the genome. Association can also be detected and wrongly interpreted as effect of the studied SNP on the phenotype of interest if the SNP has an effect on another phenotype correlated with the phenotype of interest (for example, a SNP involved in BMI variability may be associated with diabetes) or if there is **population stratification** within the sampled data. Population stratification refers to the presence of systematic differences in allele frequencies between subpopulations within a larger population and is one of the confounding factors in GWAS, that can be corrected by statistical methods like **principal component analysis**.

The association between the phenotype and SNPs is tested by **single-variant analysis**, which means every SNP is analyzed separately. Usually, the method tests for association between the phenotype and the genotype using an **additive model,** according to which the risk associated with carrying two copies of the minor allele at a SNP is two times the risk associated with carrying one copy of the minor allele. The phenotype in question can be **discrete** (e.g. disease status), **continuous** (e.g. blood glucose, Body Mass Index) or even time to disease onset. The nature of the phenotype will determine the type of model or statistical test used. Standards methods of association testing include **regression models** (linear for continuous phenotypes, logistic for binary phenotypes), which have the advantage of including more than one predictor in the regression equation if necessary. For instance, a continuous trait can be modelled by the following type of linear regression:

$$Y = \alpha + \beta W + \beta_s X_s$$

where Y is a vector of phenotype values, α is the intercept, W is a matrix of covariates (e.g. sex, age, genetic principal components), β is a corresponding vector of effect sizes, X_s is a vector of genotype values for all individuals at SNP s and β_s is the corresponding SNP effect size. Conducting millions of association tests between SNPs and a phenotype can lead to a **multiple testing issue**. Some ways to address this issue are detailed at the end of this section (Johnson *et al.* 2010).

Some main limitations of GWAS are a hindrance to their use especially for discovering variants associated with RDs. In GWAS, the less frequent the risk allele, the larger the sample size must be to keep the statistical power to detect an association. The same goes for the effect of the variant: the smaller the effect, the larger the sample size must be. Thus, GWAS are usually **underpowered** to detect associations due to the small number of individuals carrying the rare variant involved in a RD.

Despite the substantial number of GWAS carried out in the last fifteen years and their success in identifying thousands of genetic variants associated with various diseases and traits, GWAS significant hits were unable to explain the heritability predicted from traditional genetic epidemiology studies. Many causes for this "**missing heritability**" problem have been put forward (Manolio *et al.* 2009; Génin 2020), including the involvement of rare variants with larger effects and high penetrance in multifactorial diseases which is sometimes referred to as the "rare variant hypothesis" (Bodmer and Bonilla 2008). Under this paradigm of identifying rare variants involved in disease development, the study of monogenic diseases through other methods than GWAS is even more crucial, as they exhibit a strong link between these rare pathogenic variants and the expressed phenotype and can serve to elucidate disease mechanisms at the genetic and biological level (Chong *et al.* 2015). Rare variants identified in individuals with monogenic diseases may disrupt specific biological pathways or processes that can also be relevant to the pathogenesis of common multifactorial traits.

Multiple testing corrections in GWAS

The multitude of comparisons made in a GWAS results in Type 1 errors (false positives) as the probability of observing a significant association by chance alone increases with the number of independent tests. The **Family-Wise Error Rate** (FWER) corresponds to the proportion of times we falsely reject any null hypotheses and find a significant association.

To address the multiple testing issue, researchers typically employ methods to adjust for the number of comparisons being made and control this FWER. One common approach is the **Bonferroni correction**, which divides the desired significance threshold for the Type 1 error rate (e.g., 0.05) by the number of independent tests being conducted. Projects like HapMap[114] have estimated that there were on average 1 million of independent common SNPs (or blocks of LD) across the genome in European ancestry populations, which gave the classical Bonferroni testing threshold of $P_{SNP} < 5 \times 10^{-8}$ used in GWAS.

Controlling the FWER makes it unlikely to report a false positive, but also to report a true positive. Another way to address the multiple testing issue in GWAS whilst being more tolerant to a fraction of false positives is to control the **False Discovery Rate** (FDR), which corresponds to the proportion of significant tests among true null hypotheses. The most well-known method of controlling the FDR is the Benjamini-Hochberg procedure.

3.1.3 Rare variant association tests

As mentioned previously, GWAS have allowed the identification of thousands of novel loci associated with hundreds of complex traits, but mostly with small genetic effect sizes. However, GWAS are not suited for the analysis of rare variants, which are in addition in small LD with neighboring SNPs. **Rare Variant Association Tests** (RVAT) were proposed to tackle the issue of rare variant analysis by aggregating rare variants in **genetic units** (like genes) and testing for association between each genetic unit and the phenotype (Lee *et al.* 2014). The variants are also pre-filtered using different criteria like their predicted impact to retain only the most likely causal ones, which are called **qualifying variants**. The power of aggregated tests depends on the cumulated frequency of genetic variants in the genetic unit. There is also a smaller number of tests carried out compared to GWAS if the genetic unit considered is the gene, which makes the multiple testing correction much smaller (20,000 independent tests carried out instead of 1 million). The **power** of rare variant association tests will depend on the sample size, the cumulated frequency and size of cumulated effects of variants, the proportion of causal variants among qualifying variants and the underlying biological model.

There are four main categories of RVATs: burden tests, adaptive burden tests, variance-component tests and combined tests. In the case of **burden tests**, candidate rare variants are collapsed into genetic scores and the association between the score and the phenotype is tested. Burden tests make the assumption that the effect of all variants are in the same direction (protective or pathogenic), so the presence of neutral or opposite effect variants decreases their power. This limit of burden tests is taken into account by **adaptive burden tests**, which use data-adaptive weights or thresholds. Adaptive burden tests are more robust but they are often very computationally intensive. In contrast to burden tests, **variance-component tests** evaluate the collective effect of all qualifying variants by testing the distribution of effects. They are powerful in the presence of opposite effect variants or a small fraction of causal variants, but are less powerful in the optimal conditions of the burden test. Finally, **combined tests** combine burden and variance-component tests and are more robust than each method separately. They can be slightly less powerful than burden or variance-component tests if their assumptions are largely held and some methods can be very computationally intensive.

Classically, RVATs use genes as a testing unit, which restricts the analysis to coding parts of the genome only. Assessing the consequence of variants to select qualifying variants is also much more straightforward in the coding genome, as mentioned previously. In 2022, our team answered these two challenges by designing the strategy **RAVA-FIRST** (RAre Variant Association using Functionally-InfoRmed STeps) (Bocher *et al.* 2022), allowing the use of RVATs at the scale of the genome. We defined new testing units over the whole genome called **CADD regions**, using functionally-adjusted CADD scores (ACS; “Adjusted” PHRED scaled CADD Scores by “coding”, “regulatory” and “intergenic”

regions) of variants observed in the gnomAD database. Then, a region-dependent filtering step is applied for each CADD region based on the median ACS from the gnomAD database. Finally, a functionally-informed burden test is performed. RAVA-FIRST was found to outperform other whole genome RVATs that mostly use sliding windows strategies. This work was conducted by a PhD student from the team Ozvan Bocher and I contributed to the work by implementing the analysis of InDels into RAVA-FIRST.

Initially, burden tests were proposed to identify rare variant involved in common rather than rare diseases but burden tests have also been applied successfully for RD gene discovery (Madsen and Browning 2009). The most commonly used tools to apply burden tests include KBAC (Liu and Leal 2010), SKAT-O (Lee *et al.* 2012), VT (Price *et al.* 2010) and VAAST (Kennedy *et al.* 2014). However, in the case of very rare diseases or in the presence of extreme genetic heterogeneity, these methods can still be underpowered to find the causal variant, which is why an individual-based approach is often sought out for diagnosis.

3.2 Individual-based approaches for disease diagnosis

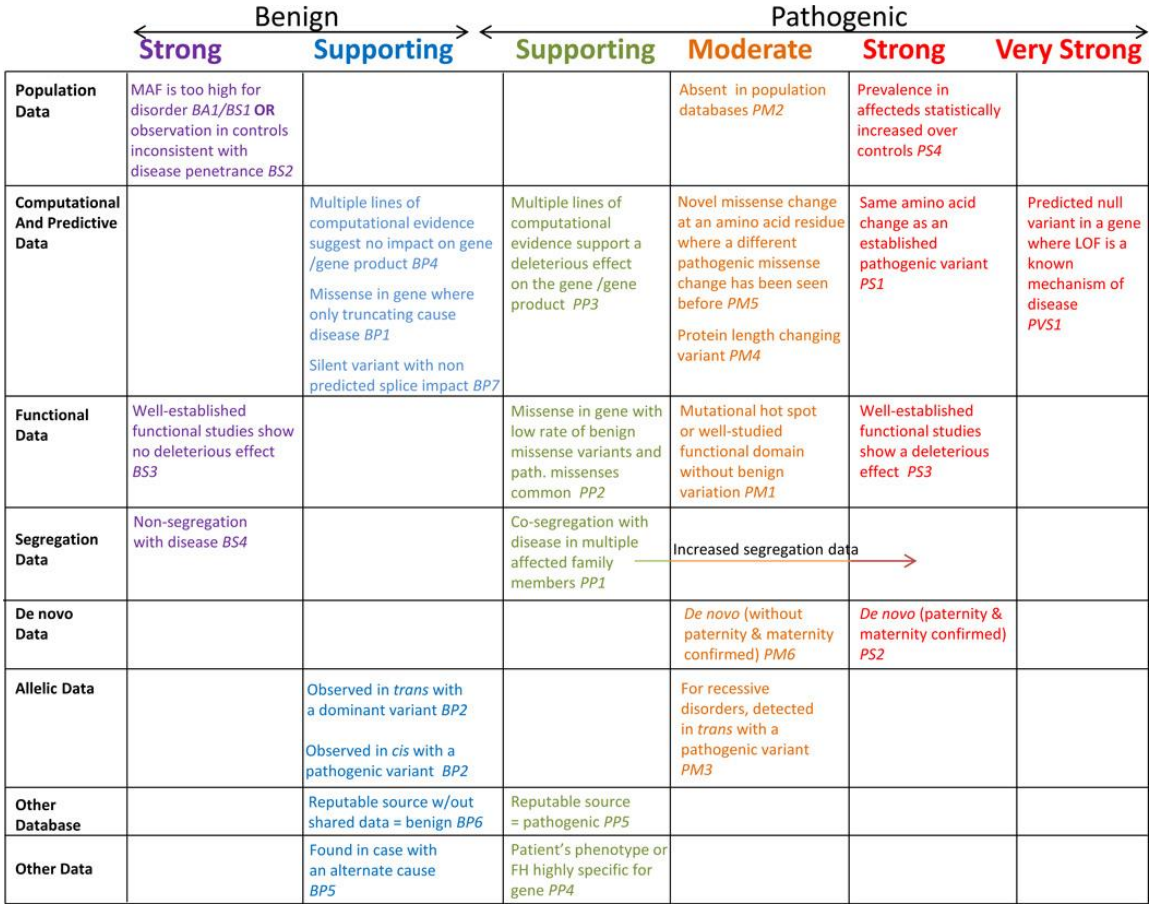
3.2.1 Variant filtering

The process of identifying disease-related variants among the background of thousands or millions of non-pathogenic polymorphisms and sequencing errors yielded by **whole exome sequencing** (WES) and **whole genome sequencing** (WGS) is a real challenge. This issue can be compared to the "**needle in the haystack**" metaphor. Indeed, the average individual carries around 20 rare loss-of-function variants (Lek *et al.* 2016), among which four are splice disrupting. Although the various databases and resources cited previously provide critical resources for the clinical interpretation of variants observed in patients suffering from RDs (Lek *et al.* 2016), a process of variant filtering often needs to be undertaken to find said "needle". Studies usually follow first a **discrete filtering** approach, according to which a set of careful filtering steps are applied to keep only variants that could be causal to the pathology, also known as **candidate variants**. This discrete filtering step is used to eliminate candidate genes by assuming that any variant found in the filter set cannot be causative, depending on the presence in controls for instance.

In 2015, the **American College of Medical Genetics and Genomics** (ACMG) provided **guidelines** (Richards *et al.* 2015) to interpret the pathogenicity of germline variants in a diagnostic context (Figure 3.1). The guidelines can allow a further filtering or **classification** of the candidate variants that passed the discrete filtering steps. Among the parameters taken into account by the ACMG guidelines, the predicted impact of the variants and *in silico* pathogenicity prediction tools play an important role.

Other crucial lines of evidence are the population frequency of the variant (available in databases like gnomAD) which can be tuned depending on the expected gene or disease at hand, segregation data, biological and clinical data. However, these guidelines have evolved a lot since their initial publication in 2015, with the development of less categorical and more nuanced Bayesian (Qian *et al.* 2018) and machine-learning (Nicora *et al.* 2022) frameworks.

In **clinical practice**, the exact filtering steps followed change drastically from case to case, and also depending on the phenotype of the individual. For RDs, if all other known monogenic causes are ruled out, a typical filtering approach involves keeping variants with a MAF in gnomAD < 10^{-4}, a high IMPACT on protein function according to the VEP software and predicted as pathogenic (according to SIFT, POLYPHEN-2 or CADD score > 20). For other diseases that are overall less rare, like intellectual disability deficiency, the MAF threshold can be higher. If familial data is available (most frequently parents/child trios), variants present in unaffected family members can be filtered out as well which facilitates the prioritization process. Once the pool of candidate variants is reduced, they are reviewed by biologists. To help further prioritize variants, the clinician can exclude variants inconsistent with the expected inheritance pattern, apply ACMG guidelines and look more closely at genes known to be associated with the phenotype or related pathways. This process is complex and can result in substantial loss of information if too stringent filters are applied.



	Benign		Pathogenic			
	Strong	**Supporting**	**Supporting**	**Moderate**	**Strong**	**Very Strong**
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*			Absent in population databases *PM2*	Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computational evidence suggest no impact on gene /gene product *BP4* Missense in gene where only truncating cause disease *BP1* Silent variant with non predicted splice impact *BP7*	Multiple lines of computational evidence support a deleterious effect on the gene /gene product *PP3*	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before *PM5* Protein length changing variant *PM4*	Same amino acid change as an established pathogenic variant *PS1*	Predicted null variant in a gene where LOF is a known mechanism of disease *PVS1*
Functional Data	Well-established functional studies show no deleterious effect *BS3*		Missense in gene with low rate of benign missense variants and path. missenses common *PP2*	Mutational hot spot or well-studied functional domain without benign variation *PM1*	Well-established functional studies show a deleterious effect *PS3*	
Segregation Data	Non-segregation with disease *BS4*		Co-segregation with disease in multiple affected family members *PP1*	Increased segregation data →		
De novo Data				*De novo* (without paternity & maternity confirmed) *PM6*	*De novo* (paternity & maternity confirmed) *PS2*	
Allelic Data		Observed in *trans* with a dominant variant *BP2* Observed in *cis* with a pathogenic variant *BP2*		For recessive disorders, detected in *trans* with a pathogenic variant *PM3*		
Other Database		Reputable source w/out shared data = benign *BP6*	Reputable source = pathogenic *PP5*			
Other Data		Found in case with an alternate cause *BP5*	Patient's phenotype or FH highly specific for gene *PP4*			

Figure 3.1 : ACMG guidelines

From Richard et al. 2015 ; BS: benign strong ; BP: benign supporting ; FH: family history ; LOF: loss-of-function ; MAF: minor allele frequency ; path.: pathogenic ; PM: pathogenic moderate ; PP: pathogenic supporting ; PS: pathogenic strong ; PVS: pathogenic very strong

3.2.2 Phenotype-informed variant prioritization integrated methods

Another approach that can be combined with variant filtering or used separately for RD variant discovery is variant prioritization, through methods integrating different lines of evidence. The output of such prioritization methods is a ranking of variants by order or predicted pathogenicity, rather than a score with a threshold of pathogenicity. Among these methods, some are based on **phenotype** information (Sifrim *et al.* 2013; Javed, Agrawal and Ng 2014; Robinson *et al.* 2014, 2020; Singleton *et al.* 2014; Zemojtel *et al.* 2014; Smedley *et al.* 2015, 2016; Yang, Robinson and Wang 2015; James *et al.* 2016; Boudellioua *et al.* 2019; Li *et al.* 2019). Phenotypic characteristics of a patient are usually provided in a standardized way though **Human Phenotype Ontology** (Robinson *et al.* 2008; Köhler *et al.* 2017) (HPO) terms. The HPO provides a standardized vocabulary and hierarchy of clinical features and disease names, as well as association between these symptoms and known disease genes. For instance, Exomiser (Smedley *et al.* 2015) uses a combination of variant frequency, *in silico* pathogenicity scoring, segregation, protein interaction networks, clinical relevance and cross-species

phenotype comparison to prioritize coding variants. The extension to the whole genome of Exomiser, called Genomiser (Smedley *et al.* 2016), uses in addition information about regulatory sequences and chromosomal topological domains to prioritize coding as well as non-coding pathogenic variants. A recent review found that among phenotype-base prioritization pipelines, Xrare (Li *et al.* 2019), Exomiser (Smedley *et al.* 2015), LIRICAL(Robinson *et al.* 2020) and PhenIX (Zemojtel *et al.* 2014) gave the best results for variant prioritization in real WES patient data (Tosco-Herrera *et al.* 2022).

3.2.3 The PSAP method

The **Population Sampling Probability** (PSAP) (Wilfert *et al.* 2016) method, which was the main focus of my work, takes another approach by integrating both **allele frequency** data and a **pathogenicity score** to prioritize variants whilst being agnostic to the phenotypic characteristics of the patient. PSAP operates under the assumption that each gene is more or less tolerant to variants with a high predicted pathogenicity, and that observing variants with the same pathogenicity scores in different genes will not have the same significance. In other words, PSAP puts variants and their pathogenicity scores back in the context of the gene they affect. The method is devoted to the analysis of case-level data and does not take into account information brought by multiple patients and altered gene recurrence as RVATs do.

The PSAP method provides for each individual a **p-value** per gene (Figure 3.2), taken from a null distribution of pathogenicity scores in this gene. This p-value per gene is the probability of observing in a healthy population a variant in the studied gene with a predicted pathogenicity score at least as high as the maximum one observed in this gene for the individual. The pathogenicity score initially used in PSAP was the CADD score v1.0, and the allele frequencies used to calibrate PSAP null distributions came from the population database ExAc. The simple yet powerful approach under which PSAP operates aims at bridging the gap between association methods and pathogenicity meta-predictors, by providing p-values that can also serve to rank and prioritize variants. In this thesis, we have taken PSAP as a conceptual idea that could be adapted and extended upon, notably by integrating new units of analysis, pathogenicity scores and allele frequencies to construct PSAP null distributions.

PSAP presented several characteristics that made it particularly useful in the context of RDs diagnosis compared to the other methods described previously: it can be applied on individual data and does not imply filtering steps or necessitate HPO terms that are not always available or easy to determine for patients as input. However, PSAP had not been updated since its initial version in 2016 and was restricted to the analysis of coding variants. There is also no commonly admitted threshold for PSAP p-value which can make them difficult to interpret and limit the usefulness of the method in practice for RD causal variant discovery.

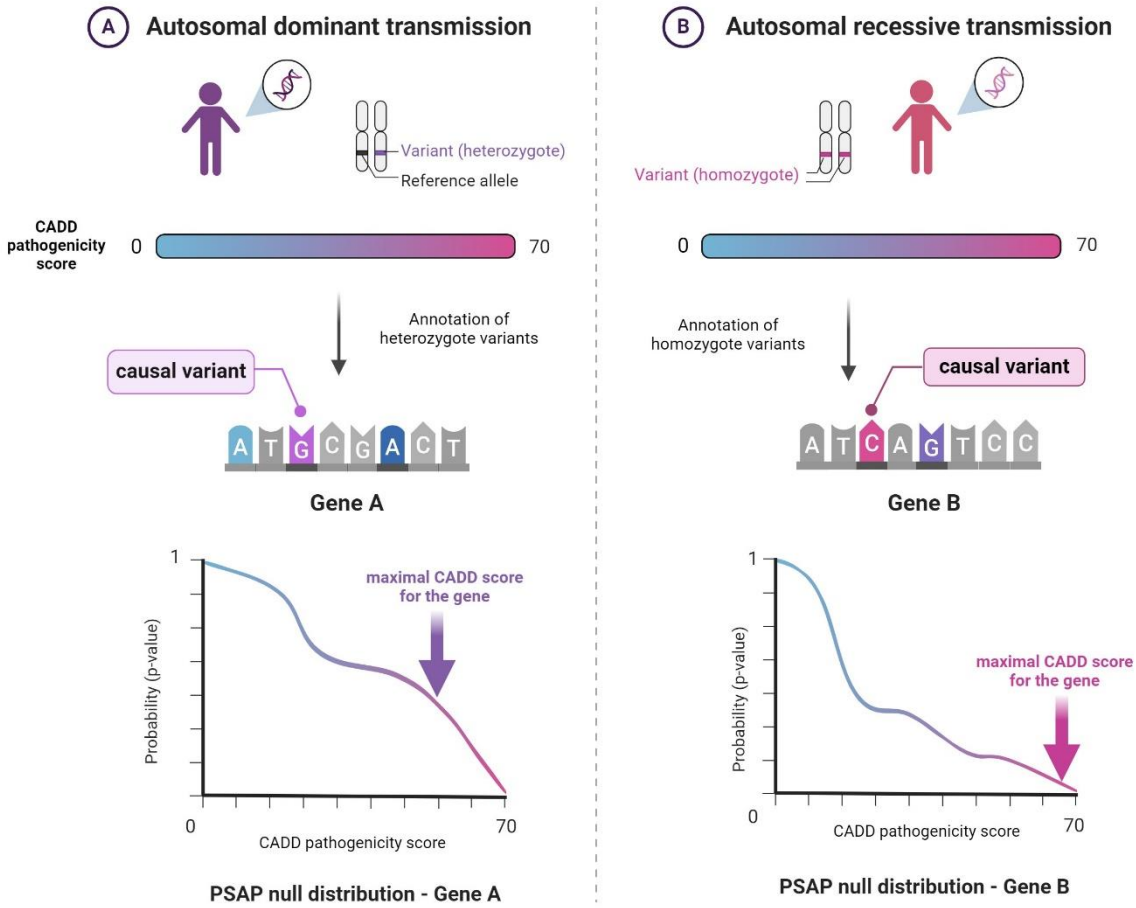


Figure 3.2 : Principle of the PSAP method

For the AD (left) and AR (right) model, PSAP scores the variant with the highest CADD pathogenicity score in the gene by looking at where it falls in the null distributions of CADD scores for this particular gene, which gives the PSAP p-value for this gene for the corresponding model of inheritance.

Created with BioRender

Chapter 4 HUMAN RARE DISEASES AND THEIR DIAGNOSIS

While most genetic variants only create polymorphisms in the human population, some germline variants in the DNA sequence can lead to the development of **inherited genetic diseases**. A significant proportion of RDs are of genetic origin. This section outlines the main characteristics of RDs and the challenge of diagnosing them.

4.1 CLINICAL AND MOLECULAR DIAGNOSIS

4.1.1 Impact of a genetic diagnosis for patients

Genetic diagnosis is crucial for RD patients in several regards. First of all, the patient can know the **risk of transmission** of the disease to any potential offspring and other genetically related individuals. Then, especially for children affected by RDs, an accurate genetic diagnosis can allow a better understanding of their disease and prognosis (Wright, FitzPatrick and Firth 2018). Knowing genetic causes of the disease can also make a **personalized treatment** and care possible for the patient. Finally, this diagnostic can have a profound impact on the patient's well-being by allowing a **recognition** of their disease by people close to them and society in general, as well as giving them the opportunity of getting in contact with patient associations.

The complex nature of RDs makes their genetic diagnosis even more difficult, with patients experiencing **diagnostic delays** and waiting on average 4 years before getting a proper diagnostic. About a quarter of patients with a RD can even have a diagnostic delay between 5 and 30 years (Uhlenbusch, Löwe and Depping 2019), that can be referred to as a "diagnostic odyssey". More alarmingly, it is believed that 40% of RD patients are given an **inaccurate diagnosis** (Chong *et al.* 2015). When a diagnostic is not currently possible for technical or knowledge reasons, RD patients are faced with a "**diagnostic deadlock**". The current diagnostic rate of RDs is still approximately of 50% (Boycott *et al.* 2017). The many reasons for the lack of diagnosis for these RD patients will be explored in section 4.2.

Although diagnosis is a crucial step that is the focus of this manuscript, 95% of RDs currently do not have a **curative treatment**. The field of **gene therapy** is giving promising results in that regard, with 50% of new genic therapies targeted towards RDs. Even when a treatment exists, drugs aiming at treating RDs, also known as **orphan drugs**, are among the most expensive drugs on the market despite some regulations by legislation. The annual worldwide expenditure on orphan drugs amounts to a staggering $125 billion (Ferreira 2019).

4.1.2 Use of Next Generation Sequencing technologies

The diagnostic process for a RD usually starts with conventional clinical practices, such as physical examination, personal and familial history, laboratory tests and image studies. Phenotypic features can be evocative of a specific syndrome or disease for the clinician and yield a first **clinical diagnosis**, or not. However, finding the true genetic causes or **molecular diagnosis** of a RD is a much more complex process, which was improved tremendously by advances discoveries around DNA.

Historically, diagnosis for RDs had been sought out by gene mapping strategies and linkage analysis. However, the advent of NGS has also allowed a shift in the techniques used to find genes involved in these RDs. A few years ago, the cost of a genome sequencing fell under $1,000 and is now dropping at $200 per genome. Thus, **WES** and more recently **WGS** have become more prevalent even in a clinical setting, as they often provide patients with a faster genetic diagnosis even where more conventional approaches are eventually expected to succeed, and avoid lengthy, expensive and invasive investigations. More importantly, they are **agnostic** to both known biology and mapping data, which makes them even more powerful for discovering genes underlying RDs. The contribution of NGS technologies to the diagnostic yield of RDs has been increasing since 2010 and has allowed the continuous discovery of new genes associated with RDs (Bamshad *et al.* 2011; Boycott *et al.* 2017; Ehrhart *et al.* 2021) (Figure 4.1). For instance, in the case of intellectual disability, the introduction of WES in 2010 and onwards added a diagnostic yield of 24–33% and a first pilot study using WGS added a further 26% in 2014 (Vissers, Gilissen and Veltman 2016). Depending on the pathology at hand, some less extensive options can be explored before attempting WES or WGS. Among them, **comparative genomic hybridization** (CGH) **array** analysis, which is part of the chromosomal microarray analysis methods, can be used to detect clinically significant major chromosomal abnormalities and sub microscopic copy number variations throughout the genome. A **gene panel** testing can also be carried out and involves analyzing a set of genes known to be associated with specific genetic conditions related to the patient's symptoms.

However, exome sequencing has been estimated to lead to a diagnosis for only around 25% of RDs (Frésard and Montgomery 2018). Thus, whole genome sequencing is becoming even more prevalent, as it allows the potential discovery of **non-coding pathogenic variants** and a more reliable detection of structural variants even in protein-coding regions (Posey 2019; Marwaha, Knowles and Ashley 2022; Wojcik *et al.* 2023). However, there is growing evidence that WGS could be used **as first-line testing** in some specific cases, without attempting CGH array of gene panel analyses. A recent review regrouping 71 studies noted that diagnostic yield was notably higher in cases where WGS was used as the initial diagnostic tool (45%), and similar in cohorts that had undergone previous genetic testing (33%) or were found to be negative on exome sequencing (33%) (Wigby *et al.* 2024). This trend

will probably be confirmed in the coming years as WGS becomes more integrated in the clinical routine and exploited through novel analysis methods.

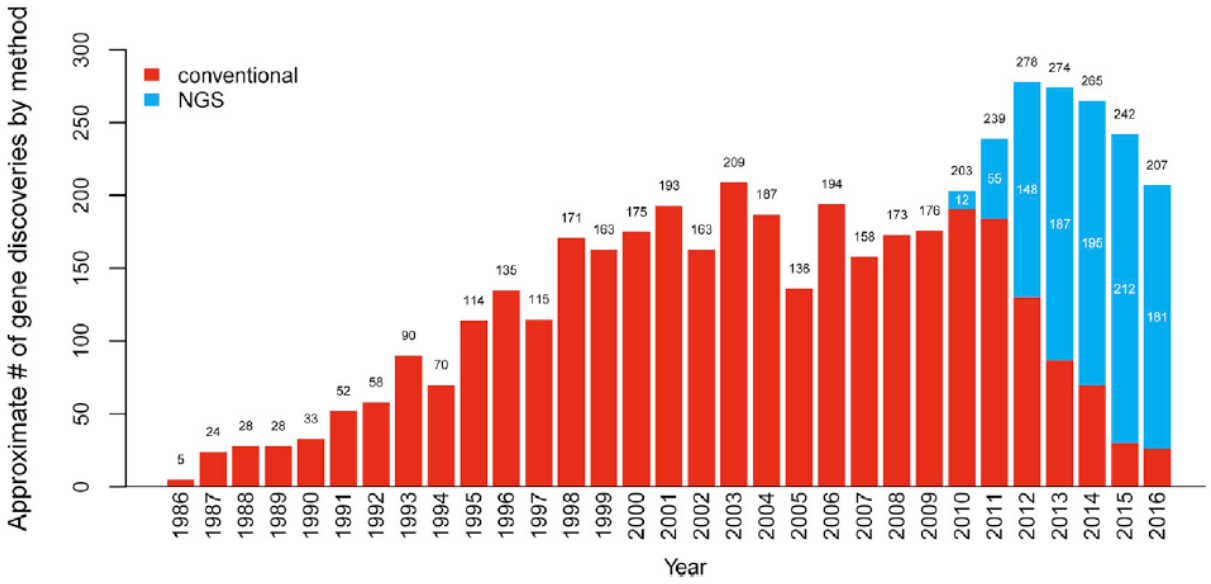


Figure 4.1 : Approximate Number of Gene Discoveries Made by WES and WGS versus Conventional Approaches since 2010 according to OMIM Data

From Boycott at al. 2017

4.1.3 Gene-disease and variant-disease association databases

As more and more genes were described and linked to RDs, it became apparent that resources aggregating this information would play a crucial role for both researchers and clinicians. Among them, **Orphanet** is the European reference portal for information on RDs and orphan drugs, created by the INSERM in 1997. Orphanet provides data on RDs including their prevalence (Nguengang Wakap *et al.* 2020), symptoms, treatments, and genetic associations. Orphanet's database is available through its platform **Orphadata**, which provides comprehensive and high-quality datasets. **OMIM** (Online Mendelian Inheritance in Man) (Amberger *et al.* 2015) is also a comprehensive collection of human genes and genetic phenotypes, based on published peer-reviewed literature. It catalogues information on monogenic diseases including RDs and their associated genes, including detailed descriptions, inheritance patterns, and molecular mechanisms. OMIM genes that are linked to a genetic disease are called **OMIM morbid**. As of February 6th 2024, there were a total of 17,207 genes described in OMIM, 6,789 phenotype descriptions with a known molecular basis, 1,504 phenotype descriptions or loci with an unknown molecular basis and 1,742 other entries, mainly phenotypes with a suspected monogenic basis.

When information from the major databases on RDs including Orphanet and OMIM is combined and compared, a **total of 10,393 RDs** can be identified (Haendel *et al.* 2020). The majority (6,370 RDs) are present in three or more resources, whereas 4,023 are unique to one source, which highlights the heterogeneity of definition and collection of RDs. This number continues to grow, as approximately **300 new monogenic diseases** or associated phenotypes are added to OMIM each year, the vast majority of which being novel gene-disease associations (Chong *et al.* 2015; Ferreira 2019).

Other databases are more focused on gathering information at the level of the genetic variant, albeit linked to phenotypes and diseases as well. Among them, **ClinVa r**(Landrum *et al.* 2018) is a freely accessible archive of reports on genetic variants, including those associated with RDs, from various sources like clinical testing laboratories and research studies. A major strength of ClinVar is the readily-available information on the **clinical significance** (benign, likely benign, uncertain significance, likely pathogenic, pathogenic) and the **review status** (no assertion criteria provided, criteria provided single submitter, criteria provided conflicting interpretations, criteria provided, multiple submitters no conflicts, reviewed by expert panel, practice guideline) of the variant, which allows an evaluation of the supporting evidence regarding the status of this variant. ClinVar was an especially useful resource in this thesis for simulating disease genomes and exomes by inserting some known pathogenic ClinVar variants in sequence data from general population individuals, for instance from the 1kGP individuals described in section 2.3.2. Another resource called **HGMD** (Human Gene Mutation Database) (Stenson *et al.* 2020) provides expert-curated information on disease-causing mutations and their associated phenotypes, including RDs, but is not freely available like ClinVar which is why it was not used in this work. ClinVar encompasses around 2,800,000 unique variation records as of March 2024, a majority of which being provided with both their clinical significance and review status. In comparison, HGMD recapitulates more than 450,000 detailed mutation reports after curation.

4.2 Specificities of rare diseases genetics

4.2.1 Complex modes of inheritance: digenism and modifier genes

For a long time, genetic diseases were believed to be caused only by alterations in a single gene, which is what we called previously a monogenic mode of inheritance. In the pre-genomic era, there was a strict divide between strictly monogenic RDs and common polygenic diseases (Messaoud *et al.* 2021). However, numerous exceptions to this rule have been described in the literature, leading to the understanding that more **genetically complex modes of inheritance** exist even for RDs (Lupski 2012). Exploring the genetically complex modes of inheritance of RDs appeared as a way to inform the etiology of both rare and common diseases (Antonarakis *et al.* 2010). In this section, we will mostly focus on digenism, the simplest form of genetically complex inheritance.

A disease is defined as having a **digenic** mode of inheritance if variants in two distinct genes are necessary to develop the disease, according to Schäffer's definition (Schäffer 2013) (Figure 1.10). These genes may act together in a pathway or have related functions that, when disrupted, lead to the observed phenotype. Individuals harboring only one of the variants are not affected. If alterations in more than two distinct genes are necessary to express the disease, it is called **oligogenism**. The first case of digenism was described in the literature in 1994 for retinis pigmentosa (Kajiwara, Berson and Dryja 1994), a degenerative disorder causing progressive loss of vision, for which heterozygote variants in both *ROM1* and *PRPH2* are needed to develop the disease. Bardet-Biedl syndrome is another disease very well-known for being linked to multiple cases of digenic and oligogenic inheritance, involving several different genes (Katsanis *et al.* 2001).

A distinction has to be made between "true" digenism and a monogenic mode of inheritance with **modifier gene** (Génin, Feingold and Clerget-Darpoux 2008; Kousi and Katsanis 2015; Rahit and Tarailo-Graovac 2020) (Figure 1.10). In the latter case, having the first monogenic variant is sufficient to exhibit the disease, but the modifier gene controls the intensity or severity of the phenotype. Whereas looking for digenic inheritance means differentiating between affected and unaffected individuals, searching for a modifier gene requires a measure of the clinical variability within affected individuals, which can be difficult to get. A last situation that has to be distinguished from digenism is the possibility of having a **dual molecular diagnosis**, which means two distinct monogenic diseases coexist in the same individual (Figure 4.2) and can lead to a blended phenotype.

To make information on digenic diseases more accessible, the database **DIDA** (DIgenic diseases DAtabase) (Gazzo *et al.* 2016) was created in 2015. DIDA provided a manually curated collection of genes and associated variants involved in digenic diseases, including SNVs and InDels. More recently, DIDA was updated and extended to become **OLIDA** (OLIgogenic diseases Database) (Nachtegael *et al.* 2022). OLIDA incorporates all oligogenic variant combinations published in the scientific literature, even those implicating structural variants. A more in-depth description and benchmark of the methods used to detect digenism, especially in sequencing data, will be the focus of Part IV - Chapter 8.

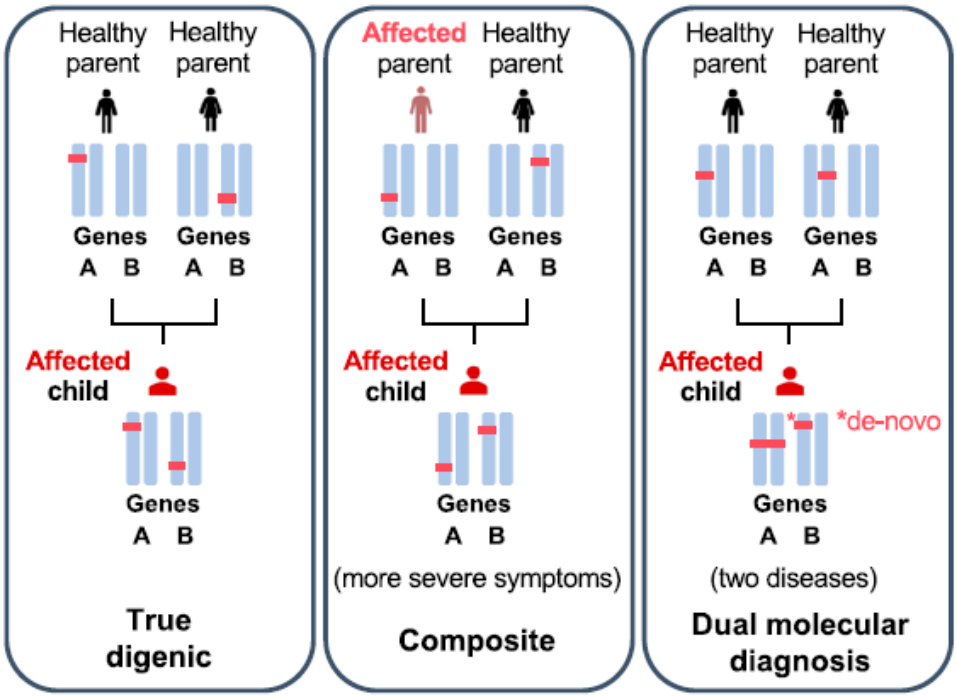


Figure 4.2 : Examples of bi-locus genetically complex modes of inheritance

From Papadimitriou et al. 2019

Other types of genetically complex modes of inheritance that are less relevant to our study of RDs and that will not be addressed in this manuscript also exist. For instance, mitochondrial DNA and thus **mitochondrial diseases** (Wei and Chinnery 2020) are inherited from the mother only. Both daughters and sons of an affected mother are affected, whereas an affected father never transmits a mitochondrial disease to his children; **imprinting disorders** (Bajrami and Spiroski 2016; Butler 2020) are caused by disturbances of the imprinting process**,** according to which certain genes are expressed or silenced based on their parental origin due to epigenetic modifications such as DNA methylation.

4.2.2 Genetic heterogeneity and the challenge of molecular diagnosis for rare diseases

Despite the great advances allowed by the advent of sequencing technologies and understanding of RD inheritance patterns, RD diagnosis in clinical genetics remains a challenge, as evidenced by the current molecular diagnosis rates of only around 50% of RD patients. Solving the molecular determinant of RD cases is made difficult by the very genetic architecture of RDs, the detection and interpretation of variants and the availability of samples.

Strong **genetic heterogeneity** (Chong *et al.* 2015) of RDs plays a huge part in this low diagnosis rate, meaning that variants in different genes (**locus heterogeneity**) or different variants in the same gene (**allelic heterogeneity**) may lead to the same disorder (McClellan and King 2010). This locus heterogeneity challenges in itself the monogenic nature of RDs, as it is one gene but not always the

same gene that leads to the disease. In addition, the monogenic model of inheritance hypothesized by several statistical methods has proven incorrect to understand the molecular causes of several RDs, as evidenced by the digenic or even oligogenic nature of RDs. There is a lack of approaches to interpret these increasingly complex scenarios in multiple variants interact together to cause or modify a RD (Frésard and Montgomery 2018).

Genetic testing relies on the accurate detection of genetic variants, including SNVs, InDels and structural variants. However, the detection of rare and novel variants can be challenging, especially in regions of the genome that are difficult to sequence or analyze. Even when genetic variants are identified, interpreting their clinical significance can be challenging especially for **non-coding variants**. Variants of Uncertain Significance are common, and determining whether a variant is causative or benign requires careful consideration of multiple factors, including population frequency, evolutionary conservation, and functional impact. In addition, high incidence of novel and ultra-rare benign variants, false assignment of variant pathogenicity, even false association of genes with disease in literature can confound results.

Finally, obtaining samples from individuals with RDs, as well as their family members for segregation analysis, can be challenging due to the rarity of the condition and the limited availability of affected individuals. Combined with genetic heterogeneity, it means that in some extreme cases, only one patient harbors a specific causal variant. This is known as the **"n-of-one" problem**, in the context of which classical association tests are powerless to detect the causal variant (Frésard and Montgomery 2018). This thesis was thus heavily focused on a method aimed at addressing this "n-of-one" problem, the PSAP method, which is the subject of section 3.2.3 of this PART II.

4.3 TWO CASE STUDIES OF RARE DISEASES

4.3.1 Cerebral Small Vessel Disease

The first pathology used as a case study in this manuscript is non-amyloid **Cerebral Small Vessel Disease** (CSVD). CSVD is a heterogeneous group of disorders that affect the structure and function of small blood vessels in the brain (Joutel *et al.* 2016), including small arteries and arterioles but also capillaries and small veins. The main phenotypic manifestations associated with CSVD are lacunar infarcts (small, localized strokes) and white matter lesions in the brain, although it is also associated with impairments of other organs (Thompson and Hakim 2009; Rannikmäe *et al.* 2020). CSVD accounts for around 20% of **ischemic strokes** and the majority of **hemorrhagic strokes** (Pantoni 2010; Marini, Anderson and Rosand 2020) and is a leading cause of vascular cognitive impairment and disability in adults. CSVD is commonly associated with aging and conditions such as hypertension,

diabetes, and smoking. While most cases of CSVD are **sporadic** and related to environmental factors and aging, **rare monogenic forms** of CSVD exist and their genetic basis has been explored over the years. Monogenic forms of CSVD are characterized by their **early onset** (before 50) and familial history. A list of genes currently known to be involved in monogenic CSVD with strong evidence and their associated phenotypes can be found in Table 4.1.

The first gene described in 1996 as causal for monogenic forms of CSVD is ***NOTCH3*** (Joutel *et al.* 1996), which is involved in the most common familial stroke disorder called CADASIL (Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy). Since then, a few additional genes have been identified including, by order of publication, ***COL4A1***, ***TREX1***, ***HTRA1***, ***COL4A2*** (Rannikmäe *et al.* 2020). For other diseases, like the Deficiency of Adenosine Deaminase 2 (associated with alterations of ***ADA2***) and Fabry disease (associated with alterations of ***GLA***), CSVD can be observed but is not the primary associated phenotype. Finally, other genes described in the literature have a weaker causal association with CSVD, like ***FOXC1*** (Marini, Anderson and Rosand 2020). More recently, ***CTSA*** has been found as a causal gene of a monogenic form of CSVD (Whittaker *et al.* 2022) and rare truncating variants in ***LAMB1*** have also been associated with CSVD (Aloui *et al.* 2021), continuing to expand the genetic etiology of the disease.

Investigations in the monogenic forms of CSVD has shed light on several pathogenic processes involved in the disease, like the impaired function of the **extracellular matrix** (Joutel *et al.* 2016; Mustapha *et al.* 2019) and its proteins, the core matrisome. Hereditary CSVD is genetically heterogeneous and represents different disease entities (Mustapha *et al.* 2019), which makes finding the molecular causes of CSVD very difficult. A substantial portion of CSVD cases believed to be of genetic origin remain unresolved, and elucidating these heritable forms of CSVD can inform the understanding and ultimately the treatment of sporadic forms as well.

Name (OMIM listing)	Inheritance	Gene	CSVD phenotypes
CADASIL (OMIM: 600276)	AD	*NOTCH3*	Lacunar infarcts, WMH, dCMB
CARASIL (OMIM: 602194)	AR	*HTRA1*	Lacunar infarcts, WMH
COL4A1-related disorders (OMIM: 120103 and 120090)	AD	*COL4A1* and *COL4A2*	Lacunar strokes, WMH, deep ICH
Retinal vasculopathy with cerebral leukodystrophy (OMIM: 606609)	AD	*TREX1*	WMH
Fabry disease (OMIM: 300644)	X-linked	*GLA*	Lacunar strokes, WMH
Deficiency of ADA2 (OMIM: 182410)	AR	*ADA2*	Lacunar strokes, WMH, dCMBs, deep ICH

Table 4.1 : Monogenic disorders exhibiting CSVD

Adapted from Marini et al. 2015
CADASIL, cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy;
CARASIL, cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy;
WMH = White matter hyperintensities, ICH = Intracerebral hemorrhage, dCMB = Deep cerebral microbleed

4.3.2 Male infertility

Male infertility is the second pathology explored in this manuscript. Male infertility is a complex multifactorial condition that affects around 7% of the male population (Krausz and Riera-Escamilla 2018), which is not in itself rare. However, while environmental factors, lifestyle choices, and other medical conditions can play significant roles in male infertility, there is a **genetic cause** for approximately 4% of diagnosed cases of infertility. The biological etiologies of male infertility are very broad. It can be caused by anomalies of the sperm, including **azoospermia** (absence of sperm in the semen), **oligozoospermia** (lower-than-normal concentration of sperm in the semen), **athenozoospermia** (poor sperm motility) and **teratospermia** (abnormal sperm morphology). Other causes of male infertility include hormonal imbalances, obstructions of the reproductive tract, erectile dysfunction and other medical conditions or treatments. The great heterogeneity characterizing male infertility makes some of its causes very rare and difficult to diagnose from an etiological standpoint. In 60 to 70% of cases of male infertility, the causes are still unknown (**idiopathic infertility**) and genetic factors are believed to play an important role in these unexplained cases.

Azoospermia is the condition with the most known genetic factors associated (**25%** of men with azoospermia carry known genetic anomalies), but new genetic factors involved in other types of male infertility are discovered regularly. In a recent review article (Houston *et al.* 2022) on the validated monogenic causes of human male infertility, **120 genes** were described as moderately,

strongly or definitively linked to 104 infertility phenotypes at different levels (Figure 4.3). This number marks a 33% increase compared to the number of genes described in 2019, highlighting great advances in the field of male infertility gene discovery mainly due to NGS studies. Large-scale worldwide initiatives like the **GEnetics of Male INfertility Initiative** (GEMINI) consortium aim at improving even more the understanding of male infertility genetics. This initiative has proven fruitful, with a study using the GEMINI data describing a **plausible recessive mongenic cause** in 20% of 1,011 cases of non-obstructive azoospermia (Nagirnaja *et al.* 2022).

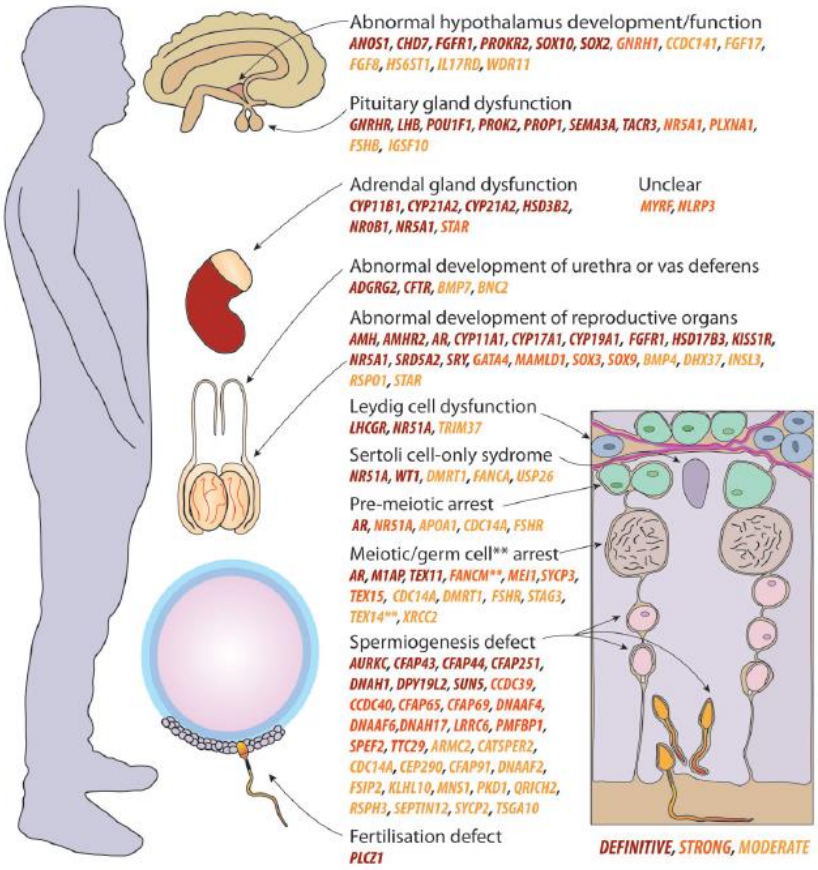


Figure 4.3 : Genes associated with male infertility

From Houston et al. 2021

Part III

Tackling genetic heterogeneity in the coding and non-coding genome

Chapter 5 PSAP-GENOMIC-REGIONS: PRIORITIZING VARIANTS IN WHOLE GENOME SEQUENCING DATA

The starting point of my PhD project was the PSAP method, which presented key advantages in the case of RDs as it was suited for individual-by-individual analysis. To benefit from the most recent developments in the analysis of sequence data, I updated the PSAP pipeline to account for larger population reference panels and the latest version of the CADD score. In addition, I took advantage of methodological developments from our team to extend the PSAP method to analyze non-coding regions of the genome. In our extension of PSAP, called PSAP-genomic-regions, I used predefined functional regions (CADD regions defined in RAVA-FIRST (Bocher *et al.* 2022)) as testing units for the construction of PSAP null distributions. The goal of PSAP-genomic-regions was to broaden the spectrum of variants detectable by PSAP, especially in introns, regulatory and splicing regions, but also to improve the performance of PSAP for patients for which the causal variant is localized in a more constrained sub-regions of a gene's coding region. PSAP-genomic-regions is described in the following preprint (Ogloblinsky *et al.* 2024), which is in review in Plos Genetics. The supplementary materials for the article can be found in Appendix I.

5.1 BACKGROUND AND SUMMARY

High genetic heterogeneity in RDs poses the challenge of identifying an n-of-one patient's causal variant using sequencing data and standard analysis methods. Initially, the PSAP method (Wilfert *et al.* 2016) used gene-specific null distributions of CADD pathogenicity scores to assess the probability of observing a given genotype in a healthy population. We propose PSAP-genomic-regions, an extension of the PSAP method to the non-coding genome using as testing units predefined regions reflecting functional constraint at the scale of the whole genome, the CADD regions. We propose two alternative PSAP-genomic-regions strategies by constructing PSAP null distributions on CADD regions with two pathogenicity scores: the initial CADD score (PSAP-genomic-regions-CADD strategy) or the ACS (PSAP-genomic-regions-ACS strategy) built to mitigate the higher CADD scores of coding variants (see section 3.1.3).

We evaluated the proposed prioritization strategies using artificially-generated disease exomes and genome. We generated these disease exomes and genomes by inserting coding and non-coding pathogenic ClinVar SNVs in 574 healthy exomes from the FrEnch Exome (FREX) Project and in 533 whole genomes from the 1kGP respectively, under both the autosomal dominant and autosomal recessive (AD and AR) models. Inserted variants were ranked based on their PSAP p-values on one hand, and on their pathogenicity score alone on the other hand. This evaluation protocol allowed us

to compare our two PSAP-genomic-regions-CADD and PSAP-genomic-regions-ACS strategies against the initial PSAP-genes (also called PSAP-genes-CADD) strategy, and against a prioritization using only the maximal CADD or ACS score by CADD region.

On the artificially-generated disease data, the two PSAP-genomic-regions strategies perform systematically better at prioritizing all types of pathogenic variants in a genome background, compared to the strategies of using maximal pathogenicity score only (CADD or ACS depending on the strategy). For coding pathogenic variant prioritization, PSAP-genomic-regions-CADD gives the best performance, and manages to rank 45.5% and 96% of variants in the top 10 of the genome for the AD and AR models, respectively. PSAP-genomic-regions-ACS prioritizes better non-coding pathogenic variants, especially splicing variants, with 56.5% and 83.3% reaching the top 10 of the genome for the AD and AR models, respectively.

We also tested our method on exome data from 6 patients with known variants causing a monogenic form of Cerebral Small Vessel Disease and genome data from 9 patients with familial forms of male infertility. Overall, the PSAP strategies always performed better than the CADD only strategies. PSAP-genomic-regions prioritized the causal variants within the top 100 variants for every individual. PSAP-genomic-regions improved notably the ranking of the causal variants in 4 out of the 6 CSVD individuals compared to PSAP-genes, and maintained a similar ranking for the 2 remaining individuals. On genome data, PSAP-genomic-regions ranks candidate genes at higher ranks than PSAP-genes. This can be explained by the fact that PSAP-gene only ranks around 4,000 variants by individual, as it analyzes only variants falling in genes, against around 70,000 variants for PSAP-genomic-regions, which analyzes the whole genome.

PSAP-genomic-regions is an efficient agnostic prioritization tool, which offers promising results for the diagnosis of unresolved n-of-one cases of RDs. To prioritize non-coding variants, the PSAP-genomic-regions-ACS gives the best results both in WES and WGS data. In the specific case of WGS coding variant prioritization, using coding parts of CADD regions as units of analysis (PSAP-coding-genomic-regions) still yields better results than PSAP-genes. Thus if the expected causal variant is coding, we advise the use of PSAP-genomic-regions-CADD in WES, and PSAP-coding-genomic-regions-CADD in WGS.

5.2 RESULTS

PSAP-genomic-regions: a method leveraging population data to prioritize coding and non-coding variants in whole genome sequencing for rare disease diagnosis

Marie-Sophie C. Ogloblinsky[1,*], Ozvan Bocher[1,2], Chaker Aloui[3], Anne-Louise Leutenegger[3], Ozan Ozisik[4], Anaïs Baudot[4], Elisabeth Tournier-Lasserve[3,5], Helen Castillo-Madeen[6], Daniel Lewinsohn[6], Donald F. Conrad[6], Emmanuelle Génin[1,7,¶], Gaëlle Marenne[1,*,¶]

[1]Univ Brest, Inserm, EFS, UMR 1078, GGB, Brest, France

[2]Institute of Translational Genomics, Helmholtz Zentrum München, Munich, Germany

[3]Université Paris Cité, Inserm, NeuroDiderot, Unité Mixte de Recherche 1141, Paris, France

[4]Aix Marseille Univ, INSERM, Marseille Medical Genetics (MMG), Marseille, France

[5]Assistance publique-Hôpitaux de Paris, Service de Génétique Moléculaire Neurovasculaire, Hôpital Saint-Louis, Paris, France

[6]Division of Genetics, Oregon National Primate Research Center, Oregon Health & Science University, Portland, Oregon, United States of America

[7]Centre Hospitalier Régional Universitaire de Brest, Brest, France

*Corresponding authors:

Email: marie-sophie.ogloblinsky@inserm.fr (M-S.O.); gaelle.marenne@inserm.fr (G.M.)

[¶]These authors contributed equally to this work

Abstract

The introduction of next generation sequencing technologies in the clinics has improved rare disease diagnosis. Nonetheless, for very heterogeneous or very rare diseases, more than half of cases still lack molecular diagnosis. Novel strategies are needed to prioritize variants within a single individual. The PSAP (Population Sampling Probability) method was developed to meet this aim but only for coding variants in exome data. To address the challenge of the analysis of non-coding variants in whole genome sequencing data, we propose an extension of the PSAP method to the non-coding genome called PSAP-genomic-regions. In this extension, instead of considering genes as testing units (PSAP-genes strategy), we use genomic regions defined over the whole genome that pinpoint potential functional constraints.

We conceived an evaluation protocol for our method using artificially-generated disease exomes and genomes, by inserting coding and non-coding pathogenic ClinVar variants in large datasets of exomes and genomes from the general population.

We found that PSAP-genomic-regions significantly improves the ranking of these variants compared to using a pathogenicity score alone. Using PSAP-genomic-regions, more than fifty percent of non-coding ClinVar variants, especially those involved in splicing, were among the top 10 variants of the genome. In addition, our approach gave similar results compared to PSAP-genes regarding the scoring of coding variants. On real sequencing data from 6 patients with Cerebral Small Vessel Disease and 9 patients with male infertility, all causal variants were ranked in the top 100 variants with PSAP-genomic-regions.

By revisiting the testing units used in the PSAP method to include non-coding variants, we have developed PSAP-genomic-regions, an efficient whole-genome prioritization tool which offers promising results for the diagnosis of unresolved rare diseases. PSAP-genomic-regions is implemented as a user-friendly Snakemake workflow, accessible to both researchers and clinicians which can easily integrate up-to-date annotation from large databases.

Author summary

In recent years, improvement in DNA sequencing technologies has allowed the identification of many genes involved in rare diseases. Nonetheless, the molecular diagnosis is still unknown for more than half of rare diseases cases. This is in part due to the large heterogeneity of molecular causes in rare diseases. This also highlights the need for the development of new methods to prioritize pathogenic variants from DNA sequencing data at the scale of the whole genome and not only coding regions. With PSAP-genomic-regions, we offer a strategy to prioritize coding and non-coding variants in whole-genome data from a single individual in need of a diagnosis. The PSAP-genomic-regions combines information on the predicted pathogenicity and frequency of variants in the context of functional regions of the genome. In this work, we compare the PSAP-genomic-regions strategy to other variant prioritization strategies on simulated and real data. We show the better performance of PSAP-genomic-regions over a classical approach based on variant pathogenicity scores alone. PSAP-genomic-regions provides a straightforward approach to prioritize causal pathogenic variants, especially non-coding ones, that are often missed with other strategies and could explain the cause of undiagnosed rare diseases.

Introduction

Each rare disease affects, by definition, a small number of individuals. However, as a whole, rare diseases affect about 350 million people world-wide (1). Approximately 80% of rare diseases have a genetic origin that mostly follows a Mendelian mode of inheritance (2–4). The advent of Next Generation Sequencing (NGS) and the development of variant pathogenicity prediction tools have allowed, in recent years, the identification of many genes involved in rare Mendelian diseases. Nonetheless, despite extensive efforts, the molecular diagnosis is still unknown for more than 50% of rare diseases cases (5–7). This can mainly be explained by the fact that many rare diseases are characterized by an extreme genetic heterogeneity, which results in only one individual carrying a specific pathogenic causal variant. This issue is referred to as the “n-of-one” problem (8).

With the advent of high throughput sequencing technologies in clinics, molecular diagnosis is now often sought through whole exome or whole genome sequencing (WES and WGS respectively). However, due to the large number of rare variants in each individual genome, causal variants are sought among very rare and highly pathogenic variants in genes relevant to the current known disease mechanism. The limited knowledge about gene functions and disease mechanisms can make this strategy unfruitful. To address the issue of variant prioritization at the level of an individual, the Population Sampling Method (PSAP) (8) was developed. PSAP computes, for each gene, a null distribution, which is the probability to observe in the general population a genotype with a CADD pathogenicity score (9) greater than or equal to the highest one to the highest one observed in the patient for this gene. This initial version of the PSAP method, which we will refer to as PSAP-genes, has been successfully applied to identify variants of interest in diverse phenotypes, including male infertility (10–12), recurrent pregnancy loss (13) and ciliary diskynesia (14).

A current hindrance to the application and generalization of PSAP-genes as a tool for diagnosis is its restriction to the coding parts of the genome. Indeed, the majority of variants reside in non-coding parts of the genome (15). Non-coding variants may contribute to explain part of the etiology of rare diseases (16), as suggested by the large number of GWAS hits located in non-coding regions of the genome (17). The involvement of non-coding pathogenic variants in rare diseases is further corroborated by the fact that non-coding regions are heavily involved in the regulation of gene expression. Several prediction tools have been developed to this end (18–20), but most of them lack a variant-based score for both coding and non-coding regions. In addition, to be performant, they often require multiple annotations like Human Phenotype Ontology (HPO) terms (21) to characterize the symptoms or disease of a patient . Thus, they rely on previous knowledge and rarely go beyond candidate genes.

To move beyond the gene as a natural unit of testing for the PSAP method, we need to use predetermined regions across the whole genome. These regions also need to be defined using functional information to be used as a cohesive unit for the construction of PSAP null distributions. This challenge of defining regions along the whole genome has been tackled by Bocher et al. in the context of rare-variant association testing (22): they describe CADD regions, which are characterized by a lack of observed variants with high functionally-Adjusted CADD Scores (ACS) in the gnomAD database (23). CADD regions are expected to reflect functional constraints. CADD regions present the key advantage of providing pre-defined and functionally-informed regions which can be used to construct PSAP null distributions.

We have made available a new implementation of the PSAP method using Snakemake (24) workflows, called Easy-PSAP (https://github.com/msogloblinsky/Easy-PSAP), which features null distributions constructed with up-to-date allele frequency data and pathogenicity scores. Here, we introduce PSAP-genomic-regions, an extension of the PSAP method to the non-coding genome by using the pre-defined CADD regions as testing unit instead of genes.

This is an innovative strategy to prioritize variants at the scale of an individual genome. PSAP-genomic-regions is now available in Easy-PSAP. We devised an evaluation protocol using artificially-generated disease exomes and genomes, obtained by inserting coding and non-coding ClinVar (25) variants in general population whole genomes from the 1000 Genomes Project (26) and exomes from the FrEnch EXome (FREX) project (27). We show the consistent improvement in prioritization by using PSAP-genomic-regions over pathogenicity scores alone for non-coding and then coding variants. For coding variants, we also demonstrate the good performance of PSAP-genomic-regions compared to PSAP-genes. On real-life data, we illustrate the power of PSAP-genomic-regions on WES data from six resolved cases of Cerebral Small Vessel Disease (CSVD) and WGS data from three families affected by male infertility. These two diseases are particularly relevant to test our method, monogenic forms of CSVD (28) and male infertility (29) being extremely heterogeneous.

Results

Construction of PSAP null distribution in coding and non-coding regions

The idea behind the original PSAP method, referred to as PSAP-genes, relies on the calculation of gene-specific null distributions of CADD pathogenicity scores. More precisely, for an individual exome or genome and in a given gene, PSAP-genes considers the genotype with the highest CADD score and evaluates the probability to observe such a high CADD score in this gene in the general population (see S1 File for a detailed explanation of the calculation of PSAP null distributions). PSAP-genes deals separately with heterozygote and homozygote variants in the autosomal dominant (AD) and the autosomal recessive (AR) models respectively. As a result, PSAP-genes gives a p-value to the genotype with the highest CADD score in the gene for each gene, model, and individual. This p-value allows the ranking of the genes for an individual exome or genome. The PSAP principle can be generalized to any genomic unit.

Here, with PSAP-genomic-regions, we extended the PSAP method to analyze whole-genome data using predefined CADD regions as testing units instead of genes (Figure 5.1). The same principle as before is employed, with the difference being that the genotype with the highest CADD score in the region can be coding or non-coding. We thus constructed PSAP-genomic-regions null distributions with two pathogenicity scores : the initial CADD score (PHRED scaled across the whole genome), or the ACS (22) (PHRED scaled CADD scores by “coding”, “regulatory” and “intergenic” regions) to mitigate the higher CADD scores of coding variants. Our two novel strategies will be referred to as PSAP-genomic-regions-CADD and PSAP-genomic-regions-ACS. They were compared to the initial PSAP-genes strategy, also referred to as PSAP-genes-CADD.

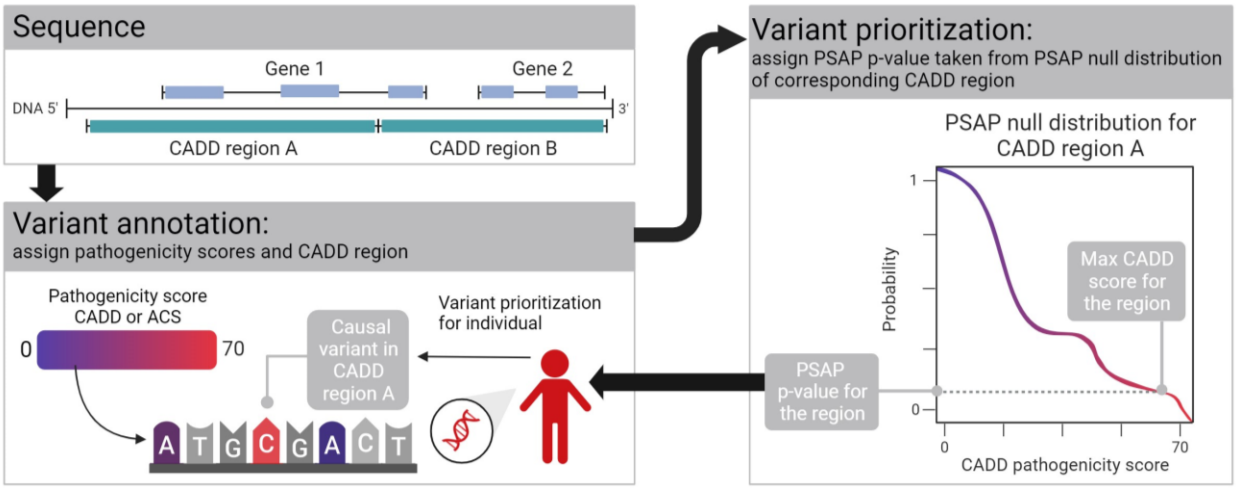


Figure 5.1. Description of the PSAP-genomic-regions strategy

We calculated PSAP null distributions for SNVs in genes and CADD regions, in the hg19 and hg38 assemblies of the human genome. In hg19, PSAP null distributions were obtained for 19,283 genes and 119,695 CADD regions. In hg38 PSAP null distributions were obtained for 18,395 genes and 123,991 CADD regions. PSAP null distributions and their parameters (unit of testing, allele frequencies and pathogenicity score) can be found in S1 Table.

Evaluating the performance of PSAP-genomic-regions on artificially-generated disease exomes and genomes using ClinVar variants

Prioritization of non-coding pathogenic variants

First, to evaluate how PSAP-genomic-regions performed to prioritize non-coding pathogenic variants, we used artificially-generated disease genomes created by inserting non-coding ClinVar variants in the NFE genomes from 1000G project (see Material & Methods and S2 File for the list of variants). Because the 1000 Genomes project is population-based, we expect that some individuals might carry one or a few pathogenic variants in their genome.

These pathogenic variants are characterized by a high CADD score and a low PSAP p-value. Indeed, there is large variation in the maximal CADD score or lowest PSAP p-value, whereas the rest of the distribution is extremely similar between individuals (S1 Fig). Thus, in order to summarize the rank of a ClinVar variant in an evaluation setting, we considered the best rank reached by the variant in at least 90% of the individuals.

Most of the NFE genomes carried a variant with a higher pathogenicity score or a lower PSAP p-value than most of the ClinVar variants (S2 Fig). We thus compared the percentage of the non-coding pathogenic variants ranked among the top N (N = 1, 10, 50 and 100) in at least 90% of the NFE genomes. The ranking at the individual level was done among all heterozygous variants for the ClinVar variants under the AD model, and across homozygous variants for the ClinVar variants under the AR model. (Figure 5.2A). With both CADD and ACS pathogenicity scores, PSAP-genomic-regions performed systematically better than using the pathogenicity scores alone. The improvement was especially large for the top 10 ranking: 24.6% and 79.2% of ClinVar variants reached the top 10 with PSAP-genomic-regions-CADD for the AD and AR models, respectively, while no ClinVar variant reached the top 10 with CADD scores alone.

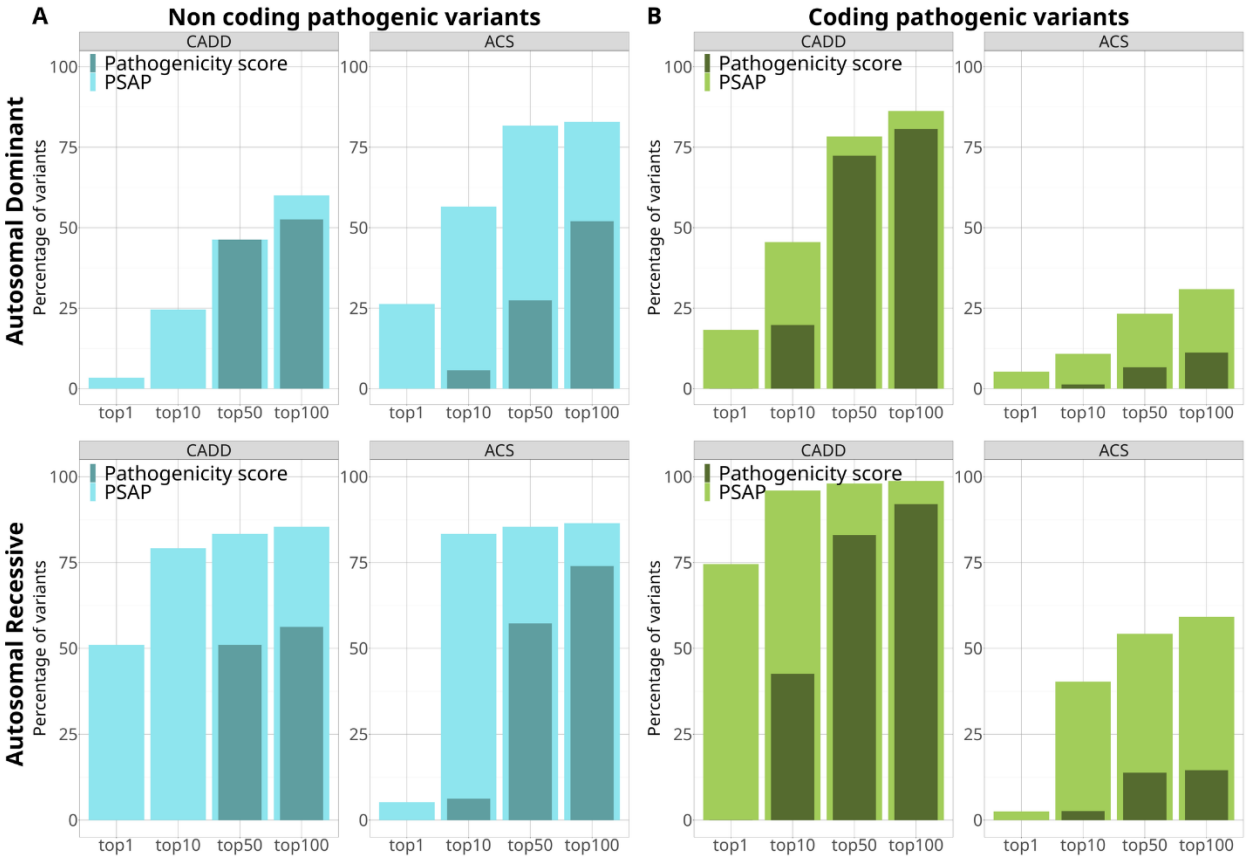


Figure 5.2. Comparison of the PSAP-genomic-regions strategy versus a pathogenicity score alone for in artificially-simulated disease genomes.

Percentage of non-coding and coding pathogenic ClinVar variants reaching the top N of variants in at least 90% of NFE genomes, with PSAP-genomic-regions (darker shade of blue or green) or the pathogenicity score alone (lighter shade of blue or green), CADD or ACS (A) N = 175 non-coding AD variants and N = 96 non-coding AR variants (B) N = 4,965 coding AD variants and N = 2,680 coding AR variants.

Using the ACS scores further improved the performance to detect non-coding-variants: 56.6% and 83.3% of variants reached the top 10 with PSAP-genomic-regions-ACS for the AD and AR models, respectively. Nonetheless, we can note the pattern is different for the top 1 for the AR model: 51% with PSAP-genomic-regions-CADD to 5.5% with PSAP-genomic-regions-ACS. Indeed, switching from CADD score to ACS score has lowered the PSAP p-value of non-coding variants shared by more than 10% of NFE genomes.

This led to a defect of the top rank reached by the ClinVar variants, as we considered the lowest rank reached in at least 90% of individuals. For instance, a variant in the CADD region R109138 shared by 70 of the NFE genomes went from a CADD score of 18.1 and a PSAP-genomic-regions-CADD p-value of 0.1 to an ACS of 22.2 and a PSAP-genomic-regions-ACS p-value of $5.18x10^{-10}$. Thus, the ClinVar variants inserted in these individuals having a higher p-value than $5.18x10^{-10}$ do not rank first.

We further explored PSAP results for splicing ClinVar variants versus other type of non-coding ClinVar variants. Indeed, we observed that splicing variants are the major type of non-coding ClinVar variants. These splicing variants often had a very good ranking, especially with PSAP-genomic-regions-ACS (n=115 splicing variants among 175 non-coding AD variants and n=72 splicing variants among 96 non-coding AR variants; S3 Table; Panel A in S3 Fig). Splicing ClinVar variants have a much higher ACS than CADD scores (Panel B in S3 Fig) which results in better ranking than for other types of non-coding ClinVar variants using PSAP-genomic-regions-ACS p-values (Panel C in S3 Fig). As a consequence, the percentage of splicing ClinVar variants ranked in the top 10 was largely improved when using PSAP-genomic-regions-ACS, for the AD model especially which was less powerful with PSAP-genomic-regions-CADD to begin with (Panel D in S3 Fig).

The full results of ranking by PSAP-genomic-regions-ACS for the non-coding non-splicing pathogenic ClinVar variants can be found in S3 File. With PSAP-genomic-regions-ACS, around half of the non-coding non-splicing variants are ranked in the top 100 of variants for more than 90% of NFE genomes (46 out of 73 variants for the AD model and 19 out of 31 variants for the AR model). The other half of variants present a less significant PSAP-genomic-regions-ACS p-value and a poorer ranking. To confirm this pattern of ranking for non-coding non-splicing pathogenic variants on another set of variants, we evaluated with our artificially generated disease genomes protocol 320 non-coding SNVs used to train Genomiser (30). These variants were not associated with a mode of inheritance.

Hence, we inserted them in the NFE genomes and scored them with both AD and AR PSAP-genomic-regions-ACS null distributions. Among the 320 non-coding variants, 169 reached the top 100 in at least 90% of NFE genomes, with either the AD or AR model (S4 File). This can be explained by the distributions of CADD scores compared to ACS scores for the ClinVar variants: the non-coding variants that do not reach the top 100 have a significantly lower CADD and ACS scores compared to all the other types of variants (S4 Fig). Overall, PSAP-genomic-regions-ACS prioritizes around half of non-coding ClinVar and Genomiser training variants in the top 100 of NFE genomes. The ones who have a higher ranking present much lower CADD and ACS scores and would never be well-ranked by any PSAP strategy.

PSAP-genomic-region is also relevant for the analysis of exome data. Indeed, exome sequencing captures variants outside of the bounds of coding regions (31), such as intronic variants. We explored the prioritization of non-coding ClinVar variants located within the WES-targeted regions of the FREX individuals using our artificially-generated disease exomes protocol (N=48 variants for the AD model and N=64 variants for the AR model, Panel A in S5 Fig). For both PSAP-genomic-regions-CADD and PSAP-genomic-regions-ACS, there was a large increase in prioritization performance compared to using only the pathogenicity scores. Because there are fewer variants in an exome background than in a genome background, the rankings of these non-coding ClinVar variants were better in FREX than in NFE genomes. The best ranking was achieved using PSAP-genomic-regions-ACS, with 82% and 90.3% of variants reaching the top 10 for the AD and AR models, respectively. Most of these non-coding pathogenic variants were splicing variants (40 out of 73 variants for the AD model and 56 out of 64 variants for the AR model), and half of them were considered as having a functional “HIGH IMPACT” (26 variants for the AD model and 22 variants for the AR model). Hence, prioritizing variants with PSAP on CADD regions allows identifying more variants even in exome data, that are in addition functionally-relevant.

Prioritization of coding pathogenic variants

Similar evaluations were performed for ClinVar coding variants inserted in either WGS from 1000G NFE individuals or WES from FREX. As observed for non-coding pathogenic variants, PSAP-genomic-regions outperformed the pathogenicity scores alone (Figure 5.2B, Panel B in S5 Fig). However, in the context of coding pathogenic ClinVar variants, we observed that the strategy of PSAP-genomic-regions-CADD provided better prioritization compared with the PSAP-genomic-regions-ACS strategy. We observed that 18.2% and 74.6% of the coding variants reached the top 1 in at least 90% of genomes backgrounds with the PSAP-genomic-regions-CADD for the AD and AR model respectively, against no variants with the CADD score alone, and against 5.3% and 2.5% reaching the top 1 with PSAP-genomic-regions-ACS. In the exome background and with PSAP-genomic-regions-CADD, 38.7% and 89.8% of AD variants reached the top 1 and top 50, respectively; 80.3% and 97.9% of AR variants reached the top 1 and the top 50, respectively.

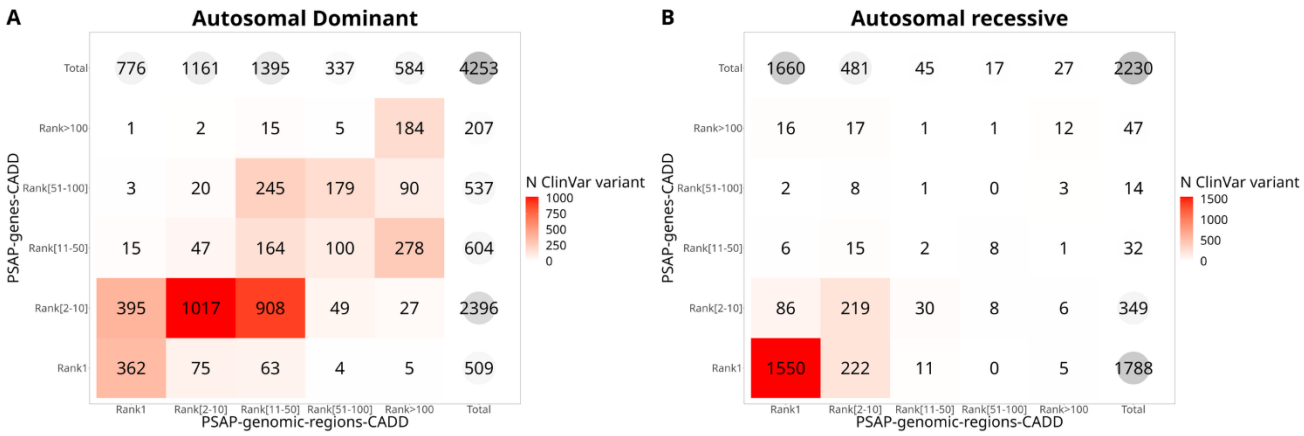


Figure 5.3. Comparison of PSAP-genomic-regions-CADD and PSAP-genes-CADD strategies in artificially-simulated disease genomes

Number of coding pathogenic ClinVar variants reaching rank [x-y] of variants in at least 90% of 1000 Genomes Project NFE individuals for each strategy.

We also compared the number of coding ClinVar variants reaching the tops in NFE genomes between PSAP-genomic-regions-CADD strategy and the initial PSAP-genes-CADD strategy (Figure 5.3). More differences were observed across the two PSAP strategies for the AD than for the AR model (Figure 5.3A). There were 362 variants ranked first and 1,017 variants ranked [2-10] in common between the two strategies. However, 908 variants that were ranked [2-10] with PSAP-genes-CADD were [11-50] with PSAP-genomic-regions-CADD, and 395 variants that were ranked [2-10] with PSAP-genes-CADD were ranked first with PSAP-genomic-regions-CADD. Regarding variants that are ranked more than a 100 with PSAP-genomic-regions-CADD, 278 of them are ranked [11-50] and are ranked [51-100] by PSAP-genes-CADD. Regarding the AR model (Figure 5.3B), PSAP-genomic-regions-CADD performed similarly to PSAP-genes-CADD, and the majority of variants were ranked first with both strategies (1,550 variants). Even more promising results can be found when looking at the same comparison of ranks within the FREX exomes (S6 Fig). For instance, in the AD model, 592 variants that were ranked [2-10] with PSAP-genes-CADD are ranked first with PSAP-genomic-regions-CADD, against 115 variants ranked [2-10] with PSAP-genomic-regions-CADD that become first with PSAP-genes-CADD.

Application of PSAP-genomic-regions to real data with different modes of inheritance

To illustrate our method in real-life settings, we analyzed two datasets (S4 Table), one with an AD mode of inheritance and the other with an AR mode of inheritance. The first dataset consisted of WES data for six individuals affected by monogenic forms of CSVD (32). Using PSAP-genomic-regions-CADD, all of the causal variants were ranked at least in the top 100 in each patient (Figure 5.4). The contribution of CADD regions as a unit of testing was especially visible for the variant in *COL4A2* and

one variant in *HTRA1* which were not well-ranked using genes as testing unit (rank 110 and 193 respectively with genes, and rank 3 and 69 with CADD regions).

Using their maximal CADD score by gene or CADD region alone, these variants would not have been prioritized in the top 100 for five out of six individuals.

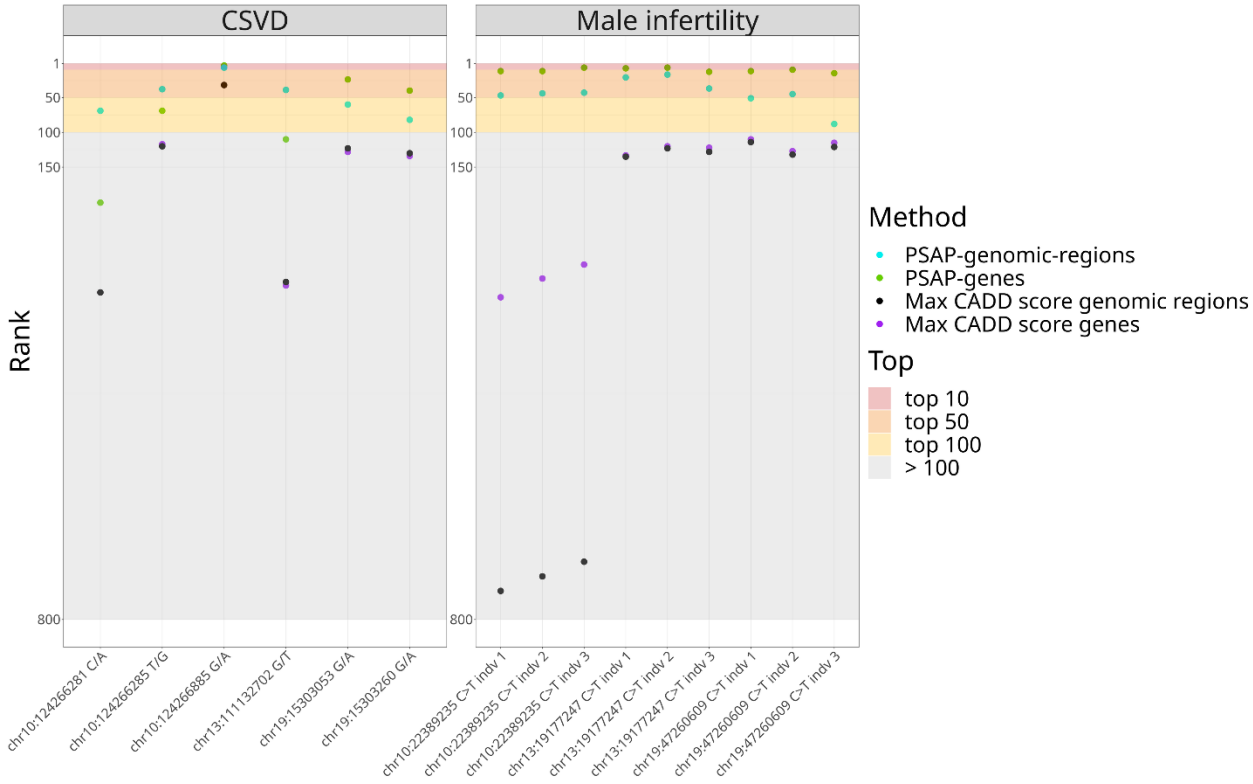


Figure 5.4. Prioritization of 6 known CSVD mutations and 3 male infertility candidate variants with PSAP-genomic-regions-CADD, PSAP-genes-CADD and the maximal CADD score on genes or CADD regions.

The second dataset consisted of WGS data for 9 individuals from three families with clinically diagnosed male infertility (33). All causal variants fell within the top 20 of variants with prioritization by PSAP-genes-CADD, and within the top 50 for at least one case per family with PSAP-genomic-regions-CADD (within top 100 for all cases, Figure 5.4). PSAP-genomic-regions-CADD did not improve the ranking of these coding variants, which was expected considering the large number of variants in a WGS analysis (see S4 Table for the total number of variants in each analysis).

The prioritization from PSAP-genomic-regions-CADD was still interesting to narrow the set of candidates for causal variants. In clinics when the CADD score alone is used, these variants would not have been prioritized (CADD score < 25, and rank > 100 with the maximal CADD score strategy). PSAP-genomic-regions-CADD thus allow a relevant prioritization of coding pathogenic variants in WGS sequencing and an unbiased exploratory analysis at the scale of the whole genome.

Using PSAP-genomic-regions-ACS or the ACS score alone, almost all of the CSVD and male infertility coding pathogenic variants had a rank greatly exceeding the top 100 (S4 Table). The only exception is one variant in *HTRA1* (10:124266885 G/A) that was ranked 3 by PSAP-genomic-regions-ACS and 10 by the maximal ACS score alone. This *HTRA1* variant was a splicing variant, which confirms the good performance of the PSAP-genomic-regions-ACS strategy on this type of variant.

Discussion

Variant prioritization, especially in the case of very heterogeneous rare diseases, is a clinically-relevant methodological challenge for both clinicians and researchers. Mounting evidence suggests that current methods of analysis and their restriction to the coding genome are a hindrance to the discovery of new genetic variants implicated in rare diseases (16). We have developed PSAP-genomic-regions, an extension of the PSAP method to the whole genome using functionally-relevant genomic regions. PSAP-genomic-regions broadens the scope of variants evaluated by PSAP and addresses the issue of variant prioritization at an individual whole-genome scale.

PSAP-genomic-regions has been thoroughly tested and validated by using simulations emulating real-life scenarios of causal variant prioritization. PSAP-genomic-regions achieves a prioritization of coding pathogenic SNVs in the top 100 variants of an exome or genome which is a relevant number of variants to analyze for clinicians. Without use of prior knowledge on the disease, PSAP-genomic-regions achieves relevant variant prioritization within millions of variants to analyze, which is illustrated by the ranking of 6 variants involved in CSVD and 3 variants involved in familial cases of male infertility in the top 100 of WES and WGS data respectively. PSAP-genomic-regions thus helps with the diagnosis of such heterogeneous diseases in conjunction with other relevant information like the mode of transmission, prevalence or type of variant involved.

PSAP-genomic-regions also allows the scoring of variants otherwise discarded from the analysis, like splicing variants with a high predicted functional impact, and other non-coding variants of proven clinical significance. The only scenario for which PSAP-genomic-regions is not advantageous compared to the PSAP-genes strategy is for prioritizing coding variants in WGS data. In that case, using coding CADD regions, i.e. the coding parts of CADD regions for the analysis still yields better results compared to PSAP-genes (S7 Fig). Our simulations using known pathogenic variants have shown which PSAP strategy performs the best depending on the type of data and variant expected to be involved in the disease mechanism (S8 Fig).

To effectively prioritize non-coding variants in WES and WGS, we advise the use of PSAP-genomic-regions-ACS. For coding variants, PSAP-genomic-regions-CADD gives the best results in WES, and PSAP-coding-genomic-regions-CADD performs best in WGS data. A two-step approach can also be carried out if there is no expected type of variant: first, the PSAP-genomic-regions-CADD or PSAP-coding-genomic-regions-CADD strategy is applied depending on the type of data, and if no coding variant of interest for the disease is found within the top results, PSAP-genomic-regions-ACS can be applied to look for non-coding variants of interest.

To the best of our knowledge, there is no other score of predicted pathogenicity for all possible SNVs comparable to CADD. Other methods have been developed to distinguish between coding pathogenic and neutral variants (34–39), but often restrict to non-synonymous variants. These methods were shown to perform better or have advantages compared to CADD for the limited set of variants they explore (34–39). Similar types of methods aim at prioritizing more constrained regions in the non-coding genome (18,20) or distinguishing deleterious non-coding variants from neutral ones (18,40). Other well-known methods for identification of pathogenic variants in exome and genome data rely on the use of HPO terms to make a prediction, like Exomiser (41) or Genomiser (30), making in comparison PSAP an unmatched prioritization tool. As any other bioinformatics variant prioritization method, it has to be used in conjunction with other lines of evidence to ultimately lead to any genetic diagnosis of a patient. PSAP-genomic-regions does not make assumption on the type of variants and does explore the whole genome. The ranking by p-values coming from the application of PSAP-genomic-regions to an individual's variants is a useful way to narrow-down the list of variants to further investigate for both researchers and clinicians in different scenarios.

The method most comparable to the strategy followed by PSAP-genomic-regions is the recently-developed machine-learning algorithm FINSURF (42). FINSURF aims to predict the functional impact of non-coding variants in regulatory regions and has been applied to known pathogenic variants inserted in WGS data like we did. Nonetheless it has been difficult to compare properly the two methods considering FINSURF only scores non-coding variants in predefined regulatory regions, and the set of variants used to train the method is not available.

The main limitation of PSAP-genomic-regions comes from the score used to calibrate null distributions, namely the CADD score. We have observed that known pathogenic non-coding ClinVar variants that were not well-ranked by PSAP-genomic-regions had significantly lower CADD and ACS scores compared to splicing and better-ranked non-coding variants. Because such CADD score is likely to be seen in the general population, PSAP-genomic-regions will not be able to prioritize such a variant with at a low rank. We also observed that some CADD regions were badly-calibrated and resulted in the assignment of very low PSAP-genomic-regions p-values to putatively neutral variants in the 1000 Genomes Project. As allele frequencies from larger databases and more accurate pathogenicity scores become available, this will lead to an improvement of the PSAP method as well. The most recent release of the CADD score v1.7 (43) notably integrates regulatory annotations and may further improve the prioritization of non-coding pathogenic variants when integrated in PSAP-genomic-regions.

Many avenues of further development and improvement are open for PSAP-genomic-regions, including the inclusion and scoring of InDel variations and structural variants. Exploring the combination of the PSAP-genomic-regions p-values with other metrics or information coming from omics analysis could also improve prediction. Finally, the flexibility of the PSAP method makes it potentially adaptable to other more complex models like digenic and oligogenic models of inheritance, considering the increasing availability of information coming from gene networks and biological pathways.

Materials and Methods

Construction of PSAP null distributions

The first parameter is the units in which to construct the PSAP null distribution. Here we considered two unit strategies: the genes and the CADD regions (S1 Table). For the genes, the coding regions of genes were defined based on the biomaRt R package: the gene coding sequences were retrieved from Ensembl (44) by requesting the "genomic_coding_start" and "genomic_coding_end", on both the hg19 and hg38 builds. To account for splicing regions, the coding regions were extended by two bases on both sides of the gene coding regions. In total, 19,780 genes were retrieved in hg19 and 23,163 in the hg38 build. For the CADD regions, their coordinates were downloaded from https://lysine.univ-brest.fr/RAVA-FIRST/ for the hg19 build and were lifted over to hg38 using the Ensembl Assembly Converter. CADD regions coordinates in hg38 are available on Easy-PSAP GitHub (https://github.com/msogloblinsky/Easy-PSAP). There were 135,224 CADD regions in hg19 and 131,970 in hg38. For the coding CADD regions, i.e. the coding parts of CADD regions, we considered the intersection of the CADD regions and the gene coding regions for each build, which yielded 37,978 coding CADD regions in hg19 and 52,340 in hg38.

The second parameter is the allele frequencies database. Here we considered the global allele frequencies from the gnomAD database to calibrate the PSAP null distributions: gnomAD genome r2.0.1 for hg19 and gnomAD V3 (45) for hg38. For our purpose, we considered only single nucleotide variants (SNVs) annotated as PASS by the Variant Quality Score Recalibration (VQSR) of GATK (46) and located in well-covered regions. Well-covered regions in gnomAD genome were defined as regions for which 90% of individuals have coverage at depth 10. Variants not seen in gnomAD genome, not annotated as PASS or not located in well-covered regions (gnomAD genome version according to the build) have a frequency of 0 and thus did not contribute to the construction of the null distributions.

To ensure reliability of PSAP null distribution, it is crucial that the units are well covered in the database from which the allele frequencies are taken. Thus, we only considered units for which at least half of the unit was well-covered (as defined previously) in gnomAD genome (version according to the build). Coding regions of genes and well-covered regions in gnomAD genome were intersected to get the percentage of each gene's coding regions that were well-covered in the database. The same steps were carried out with CADD regions as genomic units for PSAP, for hg19 and hg38 builds. PSAP null distributions were thus constructed for 19,283 and 18,395 genes in hg19 and hg38 respectively, 119,695 and 123,991 CADD regions, and 34,397 and 35,226 coding CADD regions in hg19 and hg38 respectively.

The third parameter is the pathogenicity score. Here, for the evaluation of PSAP on coding variants, we used the version 1.6 of CADD (47) for each build, accessible on the CADD website (https://cadd.gs.washington.edu/). For the evaluation on non-coding variants, which tend to have lower CADD scores than coding variants (48), we followed the strategy described in Bocher et al.(22) to adjust the RAW CADD score v1.6 of all possible SNVs on a PHRED scale stratifying by type of genomic regions: "coding", "regulatory" and "intergenic", resulting in "adjusted CADD scores", referred to as "ACS".

Easy-PSAP (https://github.com/msogloblinsky/Easy-PSAP) was used to generate null distributions according to the previously described input files and parameters. This resulted in 4 sets of null distributions for the AD and AR models for both hg19 and hg38 assemblies (S1 Table).

Evaluating the performance of PSAP-genomic-regions using artificially-generated disease exomes and genomes

To evaluate the ability of PSAP-genomic-regions to prioritize known pathogenic variants in an individual, we leveraged artificially-generated disease exomes and genomes using available general population cohorts. These different PSAP strategies (see Table S1) were compared in terms of their performances to prioritize the known pathogenic variants.

The pathogenic ClinVar (25) SNVs with coordinates in hg19 and hg38 were downloaded from the NCBI website (https://www.ncbi.nlm.nih.gov/clinvar/, accessed on the 3rd of June 2022). Some of these ClinVar variants had an annotated mode of inheritance ("moi autosomal recessive" and "moi autosomal dominant"). From ClinVar, there were 12,776 variants annotated as AD and 12,776 variants annotated as AR. Variants were filtered out to keep only autosomal pathogenic SNVs having as review status either “reviewed by expert panel” or “criteria provided, multiple submitters, no conflicts”, which are the two best review status in ClinVar. There were 1,518 AD and 1,118 AR variants meeting these criteria.

For variants which did not have an annotated mode of inheritance, we used a curated version of the database OMIM, hOMIM (49) to retrieve a mode of inheritance, and kept variants that were always associated with an AD or AR mode of inheritance in hOMIM. The same filtering was applied, which left 3,641 additional variants for the AD and 1,706 for the AR model. In total, we had a set of 5,159 variants for the AD model and 2,824 variants for the AR model. Among these ClinVar variants, 4,965 and 2,680 variants were coding SNVs respectively for the AD and AR models. Similarly, 175 and 96 variants were non-coding variants for the AD model and AR models, among which 48 variants for the AD model and 64 for AR model fell within the boundaries covered by FREX exomes. The list of pathogenic ClinVar variants and their mode of inheritance can be found in S2 File.

We inserted each variant from our curated list of pathogenic ClinVar variants successively in each of the 533 high coverage genomes of Non-Finnish Europeans (NFE) from the 1000 Genomes Project phase 4 (NFE genomes) and each of the 574 exomes from the FREX project. An individual-focused QC was applied on both datasets using the RAVAQ R package (50): we performed a genotype and variant QC with default parameters corresponding to standard GATK hard filtering criteria, mean allele balance computed across heterozygous genotypes and call rates, except for MAX_AB_GENO_DEV = 0.25, MAX_ABHET_DEV, MIN_CALLRATE and MIN_FISHER_CALLRATE "disabled".

We conducted the artificially-generated disease genome and exome evaluation with PSAP null distributions in hg19 and hg38 respectively, to match with the build of the data. We then applied the 3 PSAP strategies mentioned previously (PSAP-genes-CADD, PSAP-genomic-regions-CADD and PSAP-genomic-regions-ACS). For each strategy, we kept the maximal pathogenicity score (CADD or ACS) for each unit (gene or CADD regions) and then ranked the units according to their PSAP p-value or to their pathogenicity score alone within each genome or exome. We compared the PSAP-genes-CADD and PSAP-genomic-regions-CADD strategies to using the maximal CADD score alone by gene or CADD regions, respectively; and the PSAP-genomic-regions-ACS strategy to using the maximal ACS score by CADD region. For each ClinVar variant, we retrieved its rank within each genome or exome. Coding ClinVar variants were evaluated with the 3 PSAP strategies whereas non-coding ClinVar variants were evaluated with the novel PSAP-genomic-regions-CADD and PSAP-genomic-regions-ACS strategies (see S2 Table for more details).

Patient data analysis

The PSAP strategies were applied to real WES data from six unrelated patients affected by a CSVD for which the causal variant is known, which allowed a comparison of performance between the different strategies. The full description of the dataset can be found in [Aloui et al. 2021] (32), with the exception of the QC process. For this analysis, the same QC as for the FREX and 1000 Genomes Project datasets was performed. We applied PSAP-genes-CADD and PSAP-genomic-regions-CADD in hg19 to the six resolved CSVD patients' exome data. The other PSAP parameters were the ones by default as described previously. Two of the individuals had a causal pathogenic variant in the gene *NOTCH3* (19:15303053 G/A and 19:15303260 G/A), one individual in the gene *COL4A2* (13:111132702 G/T) and three individuals in the gene *HTRA1* (10:124266285 T/G, 10:124266281 C/A and 10:124266885 G/A). The rank of the known CSVD variants among other heterozygote variants in the patient's exome according to its PSAP p-value for the 2 strategies was then retrieved.

The PSAP strategies were also applied to WGS data of three families with clinically diagnosed forms of male infertility (33) and for which a pathogenic recessive variant was prioritized using a computational pipeline featuring the initial PSAP-genes implementation. Three affected individuals were analyzed for each family. The description of the whole dataset and candidate variant filtering process can be found in [Khan and Akbari et al. 2023] (33), except for the QC that was performed in the same way as for the CSVD data. Two other families were resolved from the same dataset, but considering that the causal variants were deletions we did not include them in the current analysis. The prioritized pathogenic variants were in the genes: *SPAG6* (chr10:22389235 C/T) for family 3, *TUBA3C* (chr13:19177247 C/T) for family 7 and *CCDC9* (chr19:47260609 C/T) for family 4. We applied PSAP–genes-CADD and PSAP-genomic-regions-CADD in hg38 to the 9 cases and retrieved the rank of the known male infertility variants among other homozygote variants in the patient's genomes according to its PSAP p-value for the 2 strategies.

References

1. Sequeira AR, Mentzakis E, Archangelidi O, Paolucci F. The economic and health impact of rare diseases: A meta-analysis. Health Policy and Technology. 2021 Mar 1;10(1):32–44.

2. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM®). Nucleic Acids Research. 2009 Jan 1;37(suppl_1):D793–6.

3. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Research. 2015 Jan 28;43(D1):D789–98.

4. Ehrhart F, Willighagen EL, Kutmon M, van Hoften M, Curfs LMG, Evelo CT. A resource to explore the discovery of rare diseases and their causative genes. Sci Data. 2021 May 4;8(1):124.

5. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet. 2018 May;19(5):253–68.

6. Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. The American Journal of Human Genetics. 2017 May 4;100(5):695–705.

7. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. The American Journal of Human Genetics. 2015 Aug 6;97(2):199–215.

8. Wilfert AB, Chao KR, Kaushal M, Jain S, Zöllner S, Adams DR, et al. Genomewide significance testing of variation from single case exomes. Nat Genet. 2016 Dec;48(12):1455–61.

9. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar;46(3):310–5.

10. Wyrwoll MJ, Temel ŞG, Nagirnaja L, Oud MS, Lopes AM, van der Heijden GW, et al. Bi-allelic Mutations in M1AP Are a Frequent Cause of Meiotic Arrest and Severely Impaired Spermatogenesis Leading to Male Infertility. The American Journal of Human Genetics. 2020 Aug 6;107(2):342–51.

11. Kasak L, Punab M, Nagirnaja L, Grigorova M, Minajeva A, Lopes AM, et al. Bi-allelic Recessive Loss-of-Function Variants in FANCM Cause Non-obstructive Azoospermia. The American Journal of Human Genetics. 2018 Aug 2;103(2):200–12.

12. Salas-Huetos A, Tüttelmann F, Wyrwoll MJ, Kliesch S, Lopes AM, Conçalves J, et al. Disruption of human meiotic telomere complex genes TERB1, TERB2 and MAJIN in men with non-obstructive azoospermia. Hum Genet. 2021 Jan;140(1):217–27.

13. Kasak L, Rull K, Yang T, Roden DM, Laan M. Recurrent Pregnancy Loss and Concealed Long-QT Syndrome. J Am Heart Assoc. 2021 Aug 16;10(17):e021236.

14. Bustamante-Marin XM, Horani A, Stoyanova M, Charng WL, Bottier M, Sears PR, et al. Mutation of CFAP57, a protein required for the asymmetric targeting of a subset of inner dynein arms in Chlamydomonas, causes primary ciliary dyskinesia. PLoS Genet. 2020 Aug 7;16(8):e1008691.

15. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences. 2009 Jun 9;106(23):9362–7.

16. Posey JE. Genome sequencing and implications for rare disorders. Orphanet Journal of Rare Diseases. 2019 Jun 24;14(1):153.

17. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019 Jan 8;47(Database issue):D1005–12.

18. Gussow AB, Copeland BR, Dhindsa RS, Wang Q, Petrovski S, Majoros WH, et al. Orion: Detecting regions of the human non-coding genome that are intolerant to variation using population genetics. PLOS ONE. 2017 Aug 10;12(8):e0181604.

19. Huang YF, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. 2017 Apr;49(4):618–24.

20. Vitsios D, Dhindsa RS, Middleton L, Gussow AB, Petrovski S. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. Nat Commun. 2021 Mar 8;12(1):1504.

21. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. Am J Hum Genet. 2008 Nov 17;83(5):610–5.

22. Bocher O, Ludwig TE, Oglobinsky MS, Marenne G, Deleuze JF, Suryakant S, et al. Testing for association with rare variants in the coding and non-coding genome: RAVA-FIRST, a new approach based on CADD deleteriousness score. PLOS Genetics. 2022 Sep 16;18(9):e1009923.

23. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.

24. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012 Oct 1;28(19):2520–2.

25. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4;46(Database issue):D1062–7.

26. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.

27. Génin E, Redon R, Deleuze J, Campion D, Lambert J, Dartigues J, et al. The French Exome (FREX) Project: A Population-based panel of exomes to help filter out common local variants. Genetic Epidemiology. 2017;41(7):691–691.

28. Rannikmäe K, Henshall DE, Thrippleton S, Ginj Kong Q, Chong M, Grami N, et al. Beyond the Brain. Stroke. 2020 Oct;51(10):3007–17.

29. Houston BJ, Riera-Escamilla A, Wyrwoll MJ, Salas-Huetos A, Xavier MJ, Nagirnaja L, 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. 2022 Feb 1;28(1):15–29.

30. Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, Spielmann M, et al. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. Am J Hum Genet. 2016 Sep 1;99(3):595–606.

31. Guo Y, Long J, He J, Li CI, Cai Q, Shu XO, et al. Exome sequencing generates high quality data in non-target regions. BMC Genomics. 2012 May 20;13:194.

32. Aloui C, Hervé D, Marenne G, Savenier F, Le Guennec K, Bergametti F, et al. End-Truncated LAMB1 Causes a Hippocampal Memory Defect and a Leukoencephalopathy. Annals of Neurology. 2021;90(6):962–75.

33. Khan MR, Akbari A, Nicholas TJ, Castillo-Madeen H, Ajmal M, Haq TU, et al. Genome sequencing of Pakistani families with male infertility identifies deleterious genotypes in SPAG6, CCDC9, TKTL1, TUBA3C, and M1AP. Andrology. 2023 Dec 10;

34. Niroula A, Urolagin S, Vihinen M. PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. PLOS ONE. 2015 Feb 3;10(2):e0117380.

35. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet. 2018 Oct 4;103(4):474–83.

36. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015 Aug 15;31(16):2745–7.

37. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812–4.

38. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018 Feb 1;34(3):511–3.

39. Li S, van der Velde KJ, de Ridder D, van Dijk ADJ, Soudis D, Zwerwer LR, et al. CAPICE: a computational method for Consequence-Agnostic Pathogenicity Interpretation of Clinical Exome variations. Genome Med. 2020 Aug 24;12:75.

40. Caron B, Luo Y, Rausell A. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. Genome Biology. 2019 Feb 11;20(1):32.

41. Smedley D, Jacobsen JOB, Jager M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015 Dec;10(12):2004–15.

42. Moyon L, Berthelot C, Louis A, Nguyen NTT, Crollius HR. Classification of non-coding variants with high pathogenic impact. PLOS Genetics. 2022 Apr 29;18(4):e1010191.

43. Schubach M, Maass T, Nazaretyan L, Röner S, Kircher M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. Nucleic Acids Research. 2024 Jan 5;52(D1):D1143–54.

44. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Research. 2022 Jan 7;50(D1):D988–95.

45. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes [Internet]. bioRxiv; 2022 [cited 2023 Aug 30]. p. 2022.03.20.485034. Available from: https://www.biorxiv.org/content/10.1101/2022.03.20.485034v2

46. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

47. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Medicine. 2021 Feb 22;13(1):31.

48. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2019 Jan 8;47(D1):D886–94.

49. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, et al. Natural selection on genes that underlie human disease susceptibility. Curr Biol. 2008 Jun 24;18(12):883–9.

50. Marenne G, Ludwig TE, Bocher O, Herzig AF, Aloui C, Tournier-Lasserve E, et al. RAVAQ: An integrative pipeline from quality control to region-based rare variant association analysis. Genetic Epidemiology. 2022;46(5–6):256–65.

5.3 Discussion - Calibration of PSAP null distributions

In this section, I discuss specific parameters and resources that I used to construct the updated PSAP null distributions, and ways to continue to improve upon what has already been done. PSAP is a powerful concept and methodological idea which was initially developed by the Pr Conrad and his team. All of the work I have done around PSAP highlights that by integrating more accurate pathogenicity scores, larger allele frequency databases and functionally-relevant testing units, PSAP gains in precision as a tool for variant prioritization.

5.3.1 Testing units

5.3.1.1 Choice of testing units

The issue of choosing testing units to construct PSAP null distributions is analogous to the issue of choosing specific units in RVATs (Chen, Coombes and Larson 2022) to aggregate rare variants. According to the PSAP principle, a variant's pathogenicity score has to be looked at in the context of a biologically-relevant genomic region. The natural unit for such region-based testing is the gene, as it also encompasses most of the potentially pathogenic variants affecting protein function. For RVATs, different strategies have been proposed as alternative to the standard gene-based testing, like using sliding genomic windows (Ionita-Laza *et al.* 2012; Schaid *et al.* 2013) as testing units or combining gene-level results are the scale of the pathway or the gene-set (Wu and Zhi 2013).

Although our primary goal was to extend PSAP to analyze the non-coding genome, hence the choice of CADD regions, we also considered other strategies with PSAP null distributions constructed on units of testing other than genes that we discuss hereafter. The sliding window strategy was not applicable for the PSAP method as PSAP null distributions need to be pre-computed on a defined set of genomic regions. Such strategy would be highly computationally and storage intensive, so we discarded it.

In the idea of looking at biologically-relevant genomic regions with a finer granularity than with genes, individual exons or protein functional domains could have been used as units of testing. We have mentioned that a number of functional units had been described in the non-coding genome (e.g. enhancers or promoters). These regulatory elements can have an impact on gene expression (Kolovos *et al.* 2012; Elkon and Agami 2017) due to the 3D organization of the genome, by their individual effect or in conjunction with the action of other regulatory elements. Thus, regulatory elements could also be used as units of testing for PSAP, by themselves or grouped together in a relevant region like a TAD. In the same idea as CADD regions, other type of functionally-informed regions had already been defined like Constrained Coding Regions (CCRs) (Havrilla *et al.* 2019) and could serve as units of testing. These approaches could allow a better evaluation of variant pathogenicity in specific sub-regions of a

gene or in a set of regulatory elements or constrained regions. However, they present the major drawback of not being defined over the whole genome. The use of these units of testing would have restricted the analysis by PSAP to a smaller number of variants and discarded systematically some types of variants like intergenic ones. In addition, regions like CCRs could be too small to construct PSAP null distributions from a sufficient number of variants in the allele frequency database (the maximum length of a CCR being 224 bp). Nonetheless, these strategies could be considered in other research projects for which they would be relevant. To that end, we have provided the pipeline to calculate PSAP null distributions for any unit of testing, which we will expand on further in Chapter 6.

5.3.1.2 CADD regions

Considering the limitations of the aforementioned units of testing, we chose to adapt the CADD regions developed for the RAVA-FIRST strategy (Bocher *et al.* 2022) in the context of PSAP to propose the PSAP-genomic-regions strategy and extend PSAP to the non-coding genome. These regions span the entire genome and their boundaries are based on variants with high CADD scores observed in the general population. By construction, these regions are depleted in observed highly pathogenic variants and thus reflect functionally constrained regions of the genome. CADD regions had initially been defined by our team in (Bocher *et al.* 2022) for the hg19 assembly, with their primary boundaries being variants with an Adjusted CADD Score v1.4 (ACS 1.4) ≥ 20 and that were seen at least two times in the database gnomAD v2 (AC > 2). To construct PSAP null distributions for the hg38 assembly, we thus lifted over the coordinates of CADD regions from hg19 to hg38 and used allele frequencies from gnomAD v3 and the CADD or ACS pathogenicity scores 1.6 in hg38 as well. Although some CADD regions were lost in the liftover process, using the hg38 liftover CADD regions ensured we used CADD regions that were already well documented and tested to construct our PSAP null distributions.

An observation that sheds light on the parameters that can affect CADD regions boundaries and that ties with the previous remark is the definition of new CADD regions directly in the hg38 assembly. With two interns, we worked on defining CADD regions in hg38 using gnomAD v3 and the newly-calculated ACS 1.6. A first result that we were able to point out is the repartition of the three types of genomic regions (“coding”, “regulatory” and “intergenic”) used to define the ACS in hg38, by using the latest versions available for the same resources as in (Bocher *et al.* 2022). In hg19, the regulatory regions accounted for 44% of the genome, whereas they covered 60% of the genome in hg38. This was mainly due to intergenic regions becoming regulatory in the ENCODE v3 (Moore *et al.* 2020) in hg38.

We also observed the impact of using a larger database to define the boundaries of CADD regions. Indeed, gnomAD v3 (Chen *et al.* 2022) includes more than 70,000 genomes against around 15,000 for gnomAD v2 (Karczewski *et al.* 2020). As expected, a lot more variants met the criteria for CADD regions boundaries with gnomAD v3 (242,220 variants with ACS ≥ 20 and AC > 2 in gnomAD v3 against 78,181 in gnomAD v2), which resulted in more CADD regions (195,447 in hg38 against 135,224 in hg19). These new CADD regions in hg38 were overall smaller, with a mean length of 14.10 Mb against 19.85 Mb for the hg19 CADD regions. Our next step will be to see if the boundaries of some of the longer and thus more conserved CADD regions were maintained with this new version, which would be a good way to test if CADD regions indeed reflect regions of the genome intolerant to deleterious variations. These new CADD regions in hg38 have not yet been tested in simulations or real data, or compared to the liftover version of CADD regions that we had been using until then. There is still ongoing work to include them in PSAP-genomic-regions as well as in the RAVA-FIRST strategy. Our aim has also not been to rethink the definition itself of CADD regions (criteria for boundaries, clustering of small regions), which could be discussed down the line depending on how the new hg38 CADD regions perform and to what aim they could be adapted.

5.3.2 Pathogenicity score

The PSAP method ranks variants within a gene according to a pathogenicity score. The initial version of PSAP used the CADD score as pathogenicity score. In the same vein, we have kept the same albeit updated CADD score to calibrate our PSAP null distributions. Indeed, the CADD score presents several advantages that are discussed in Part II - Chapter 2, including being defined over the whole genome which was crucial for our PSAP-genomic regions extension. In addition, it was also the score used to construct the ACS and define the CADD regions. We thus felt we had a good understanding and hindsight of this particular pathogenicity score as the basis for evaluating variant deleteriousness in the PSAP method.

The issue of choosing a variant pathogenicity score is complex and depends on number of factors. An extensive review on the subject of variant prioritization scores is the one from (Eilbeck, Quinlan and Yandell 2017). They argue that there is no one-size-fits-all approach when it comes to choosing a pathogenicity score to evaluate a variant's impact and that the different software and approaches have to be carefully compared and interpreted. No pathogenicity score reflects perfectly the true impact of a variant, which has to be put back into the context of a patient's genotype, family history and phenotype. Even benchmarks of pathogenicity scores report conflicting results (Ruscheinski *et al.* 2021; Liu *et al.* 2022) and no score gives the best results for all types of variants and scenarios. In practice, any pathogenicity score could have been used in place of CADD for our work, from the main ones described in Part II - Chapter 2 or new ones that will be developed in the coming

years as our knowledge on the genome continues to improve. Another option, in line with the idea of the ACS, could be to combine different scores depending on regions of the genome.

Comparing the performance of PSAP null distributions calibrated with different pathogenicity scores would have been an extensive work that was beyond the scope of this thesis. However, by making available the workflow to construct PSAP null distributions, we leave room for new PSAP strategies to be created and tested, perhaps with a new score or a blend of scores depending on the genomic region. We also proposed an evaluation protocol using simulated disease genomes and exomes that can be used to test the performance of PSAP null distributions constructed according to different parameters. To open up the discussion on the impact of using a new variant pathogenicity score to calibrate PSAP null distributions, I have used the recently published CADD 1.7 (Schubach *et al.* 2024) which is presented as improving the prediction of pathogenicity in the non-coding genome especially. In Table 5.1, we can see the comparison of prioritization performance between PSAP constructed on CADD 1.7 or the ACS 1.7 and the same strategies with the CADD 1.6 on artificially simulated disease genomes in build 38 (with inserted pathogenic coding and non-coding variants, for the AD and AR models). Surprisingly, the CADD or ACS 1.7 strategies give worse results compared to CADD or ACS 1.6 in almost all scenario. This could be due to overfitting in the CADD 1.7 model, due to the high number of features included for the prediction. This behavior has not been replicated by (Schubach *et al.* 2024), who found that CADD 1.7 slightly improved the classification compared to CADD 1.6 in a set of ClinVar benign and pathogenic variants, mostly in the coding genome. More in depth testing would have to be carried out to see if CADD 1.7 performs better for a specific type of variants compared to CADD 1.6, and why the performance drops for other types of variants.

(A) Pathogenic coding variants, AD model

Pathogenicity score	Top in controls	Pathogenicity score alone 1.7	PSAP 1.7	Pathogenicity score alone 1.6	PSAP 1.6
CADD	top1	0.282%	18.857%	0.071%	18.246%
CADD	top10	25.888%	44.980%	19.751%	45.544%
CADD	top50	61.274%	75.852%	72.349%	78.345%
CADD	top100	78.744%	83.024%	80.696%	86.269%
ACS	top1	0%	4.914%	0%	5.314%
ACS	top10	0.494%	11.498%	1.293%	10.816%
ACS	top50	4.867%	23.113%	6.654%	23.301%
ACS	top100	9.499%	27.933%	11.169%	30.919%

(B) Pathogenic coding variants, AR model

Pathogenicity score	Top in controls	Pathogenicity score alone 1.7	PSAP 1.7	Pathogenicity score alone 1.6	PSAP 1.6
CADD	top1	0.089%	72.695%	0.045%	74.566%
CADD	top10	43.296%	95.011%	42.628%	95.991%
CADD	top50	84.677%	97.105%	83.029%	97.996%
CADD	top100	92.472%	97.906%	92.027%	98.753%
ACS	top1	0%	2.004%	0%	2.494%
ACS	top10	2.450%	28.909%	2.628%	40.312%
ACS	top50	14.120%	51.982%	13.808%	54.254%
ACS	top100	14.610%	56.704%	14.521%	59.198%

(C) Pathogenic non-coding variants, AD model

Pathogenicity score	Top in controls	Pathogenicity score alone 1.7	PSAP 1.7	Pathogenicity score alone 1.6	PSAP 1.6
CADD	top1	0%	2.286%	0%	3.429%
CADD	top10	0%	14.286%	0%	24.571%
CADD	top50	17.714%	34.286%	46.286%	46.286%
CADD	top100	31.429%	41.143%	52.571%	60.000%
ACS	top1	0%	20.000%	0%	26.286%
ACS	top10	0%	46.857%	5.714%	56.571%
ACS	top50	17.714%	72.571%	27.429%	81.714%
ACS	top100	30.286%	76.000%	52.000%	82.857%

(D) Pathogenic non-coding variants, AR model

Pathogenicity score	Top in controls	Pathogenicity score alone 1.7	PSAP 1.7	Pathogenicity score alone 1.6	PSAP 1.6
CADD	top1	0%	38.542%	0%	51.042%
CADD	top10	1.042%	80.208%	0%	79.167%
CADD	top50	41.667%	82.292%	51.042%	83.333%
CADD	top100	50.000%	84.375%	56.250%	85.417%
ACS	top1	0%	0%	0%	5.208%
ACS	top10	6.250%	65.625%	6.250%	83.333%
ACS	top50	48.958%	80.208%	57.292%	85.417%
ACS	top100	60.417%	82.292%	73.958%	86.458%

Table 5.1 : Percentage of pathogenic variants in top of artificially simulated disease genome

(A) Pathogenic coding variants, AD model ; (B) Pathogenic coding variants, AR model ;

(C) Pathogenic non-coding variants, AD model ; (D) Pathogenic non-coding variants, AR model

5.3.3 Allele frequencies

Another important parameter for the calibration of PSAP null distributions is allele frequencies. All of our PSAP null distributions were constructed using the global allele frequencies from the gnomAD database. However, we know that allele frequencies differ between populations, which has been shown extensively by the multiple population genetic data panels including 1kGP and gnomAD. The version of the PSAP pipeline featured on the Conrad Lab GitHub (https://github.com/conradlab/PSAP) integrates an argument for the individual's ethnicity and the corresponding null distribution.

Using allele frequencies derived from the European Non Finnish gnomAD samples instead of global allele frequencies to construct PSAP null distributions gave very similar results in our simulations. This is not surprising given that the FREX individuals and 1kGP individuals chosen to create our disease exomes and genomes were of European descent. Nonetheless, population-specific PSAP null distributions could have been valuable for the Malakand analyses as strong founder effects are described in the South Asian(Wall *et al.* 2023) populations. Another useful metric could be to use the POPMAX allele frequency from gnomAD to calibrate PSAP null distributions, which is the allele frequency information for the non-bottlenecked population with the highest frequency, under the assumption that disease-causing variants would be uncommon in any population. The only caveat to using population-specific allele frequencies is that it reduces the number of individuals which are taken into account, resulting in the lack of observation of more rare variants.

More extensive simulations and comparisons would be needed to assess the impact of each parameter on the performance of PSAP null distributions. With the Easy-PSAP workflow to calculate PSAP null distributions, which we discuss in the next section, we have made it possible for other researchers to tune PSAP to their own specific needs or question. Understanding which known pathogenic variants are not well-prioritized by the framework could also open up new possibilities of improvement and exciting new avenues of development for PSAP.

Chapter 6 EASY-PSAP: AN INTEGRATED WORKFLOW TO PRIORITIZE PATHOGENIC VARIANTS IN SEQUENCE DATA FROM A SINGLE INDIVIDUAL

During the development of PSAP-genomic-regions, I proposed several updates and extensions of the PSAP pipeline and implemented efficient scripts to perform a fast calculation of PSAP null distributions with a flexible choice of input parameters. To make available an updated and easier to use version of the PSAP pipeline I created Easy-PSAP, a user-friendly, flexible and computationally efficient Snakemake (Köster and Rahmann 2012) workflow to create and apply all of our currently developed PSAP null distributions. Easy-PSAP is presented in the following manuscript, currently in review in BMC Bioinformatics. The supplementary materials for the Easy-PSAP manuscript can be found in Appendix II and the Easy-PSAP user guide in Appendix III.

6.1 BACKGROUND AND SUMMARY

As mentioned previously, the PSAP method was developed to tackle the issue of variant prioritization for a single patient, by leveraging allele frequencies from population databases and a variant pathogenicity score. However, the initial implementation of PSAP featured bash and R scripts to apply PSAP null distributions that were not easily adaptable for all users. The initial PSAP null distributions also used ExAC as a reference panel and CADD v1.0, and had not been updated since. The scripts to generate PSAP null distributions themselves had not been made available either.

Here, we describe Easy-PSAP, a new and update implementation comprising of two user-friendly and highly adaptable pipelines based on the PSAP principle, which can evaluate genetic variants at the scale of a whole genome using information from the latest population and annotation databases. In contrary to the initial PSAP that was restricted to the exome, Easy-PSAP allows the analysis of variants in the coding and non-coding genome by integrating both PSAP-genes and PSAP-genomic-regions in the parameters and available null distributions of the pipeline. Easy-PSAP features both a workflow to calculate PSAP null distributions and a workflow to apply them to patient data.

In the following application note, we tested the performance of Easy-PSAP on genes compared to the initial PSAP pipeline on simulated synthetic disease exomes. We inserted known pathogenic variants from the ClinVar database in healthy sequence data and evaluated their rank based on their PSAP p-value. Easy-PSAP showed a clear gain in performance compared to the initial PSAP. Overall, Easy-PSAP was able to capture more than 50% of causal coding pathogenic variants in the top 10 variants for an AD model of transmission and in the top 1 for an AR model of transmission.

These findings, along with the accessibility of the pipeline to both researchers and clinicians, make Easy-PSAP a state-of-the-art tool for NGS data analysis that is implemented to evolve as new frameworks and databases become available. In particular, the workflow to calculate new PSAP null distributions allows researcher to tailor PSAP to their research question and requirements.

6.2 RESULT

Easy-PSAP: an integrated workflow to prioritize pathogenic variants in sequence data from a single individual

Marie-Sophie C. Ogloblinsky[1,*], Daniel P. Lewinsohn[2], Mathilde Nguyen[1], Lourdes Velo-Suarez[1,3], Anthony F. Herzig[1], Thomas E. Ludwig[1,4], Helen Castillo-Madeen[2], Donald F. Conrad[2], Emmanuelle Génin[1,4,†], Gaëlle Marenne[1,*,†]

[1]*Univ Brest, Inserm, EFS, UMR 1078, GGB, F-29200 Brest, France*

[2]*Division of Genetics, Oregon National Primate Research Center, Oregon Health & Science University, Beaverton, Oregon, United States of America*

[3]*Centre Brestois d'Analyse du Microbiote (CBAM), Brest University Hospital, Brest, France*

[4]*CHU Brest, Brest, France*

**Correspondence: marie-sophie.ogloblinsky@inserm.fr; gaelle.marenne@inserm.fr*

†Emmanuelle Génin and Gaëlle Marenne have contributed equally to this work

Abstract

Background: Next-Generation Sequencing data analysis has become an integral part of clinical genetic diagnosis, raising the question of variant prioritization. The Population Sampling Probability (PSAP) method has been developed to tackle the issue of variant prioritization in the exome of a single patient, by leveraging allele frequencies from population databases and a variant pathogenicity score.

Results: Here, we present Easy-PSAP, a completely new implementation of the PSAP method comprising two user-friendly and highly adaptable pipelines. Easy-PSAP can evaluate genetic variants at the scale of a whole exome or genome using information from the latest population and annotation databases. Through simulations on synthetic disease exomes, we show that this new implementation is able to capture more than 50% of causal pathogenic variants in the top 10 variants for an autosomal dominant model of transmission and in the top 1 for an autosomal recessive model of transmission.

Conclusion: These findings, along with the accessibility of the pipeline to both researchers and clinicians, make Easy-PSAP a state-of-the-art tool for variant prioritization in Next Generation Sequencing (NGS) data that can continue to evolve as new frameworks and databases become available. Easy-PSAP is implemented in R and bash within an open-source Snakemake framework. It is available on GitHub alongside conda environments containing the required dependencies (https://github.com/msogloblinsky/Easy-PSAP).

Keywords: variant prioritization, Next Generation Sequencing, rare diseases

Background

With the advent of Next Generation Sequencing (NGS), genomic data has become more widely available and has opened up a new era for rare disease diagnosis. These rare diseases are characterized by an important genetic heterogeneity, that can result in only one individual carrying a specific pathogenic variant in a specific gene, also known as the “n-of-one problem”. Prioritization tools for causal pathogenic variants like the Population Sampling Method (PSAP) (1), which are applicable to sequence data from a single individual, have been developed to tackle this issue. PSAP uses allele frequencies from large population databases to construct gene-based null distributions of CADD pathogenicity scores (2), to allow the evaluation of a variant of unknown significance in the context of each gene, without the need for control individuals. PSAP gives a p-value by gene for each individual, which summarizes how unlikely it is to observe a variant in this gene with such CADD score in the general population.

The initial implementation of PSAP used allele frequencies from the ExAc database (3), comprising around 60,000 individuals at the time, and the version of CADD v1.0 to construct null distributions, and featured the use of a bash pipeline to apply PSAP to patient data. The annotation of the input Variant Call Format (VCF) file was conducted by Annovar (4). Since then, major updates in genomic sequence databases (gnomAD V2 (5), comprising 125,748 exomes and 15,708 genomes) and variant annotations (CADD v1.6 (6)) have proven to better capture the complexities of genetic variations in humans.

Here, we introduce Easy-PSAP a new implementation of the PSAP pipeline through computationally efficient and user-friendly Snakemake workflows (7). A first workflow allows the custom calculation of PSAP null distributions from allele frequencies data and a pathogenicity score, which had not been previously available (Figure 6.1A). A second workflow applies these null distributions to patient data, with the most up-to-date null distributions already available without calculations (Figure 6.1B). This highly flexible ecosystem is adaptable to the specific needs of researchers and the growing information arising from new databases, whilst already providing a strong framework for the scoring of genetic variants of an individual patient scalable to whole-genome data.

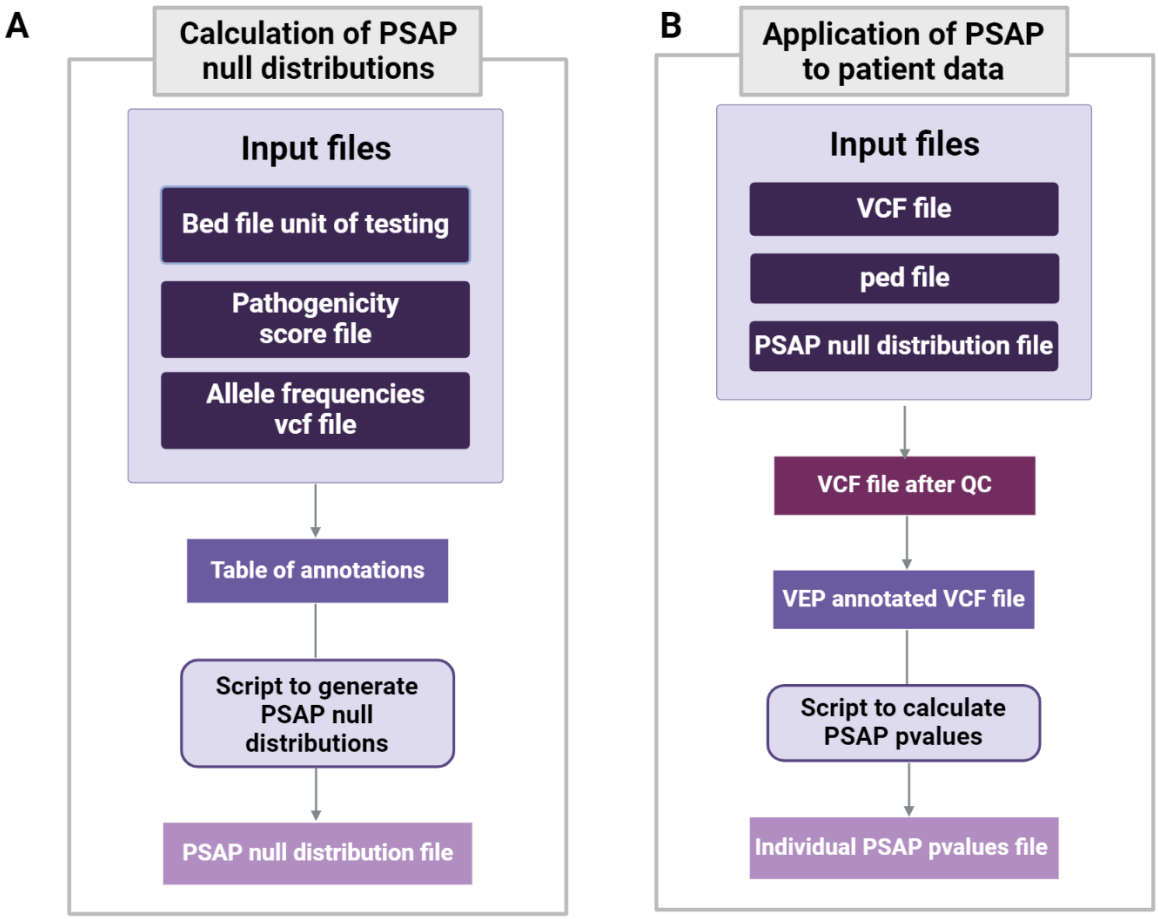


Figure 6.1. Description of Easy-PSAP

(A) Snakemake workflow to construct PSAP null distributions. (B) Snakemake workflow to apply PSAP to patient data.

Implementation

The new implementation of PSAP relies on the use of Snakemake workflows, which are a series of “rules” connecting input to output files. When executed, Snakemake infers the combination of rules allowing the creation of the desired output files. Snakemake can be tuned to the specific computational resources provided by the user, making the pipeline scalable. The pipeline is also efficient: it does not rerun steps for which the output file is already present, which allows the user to easily iterate the analysis over a set of different parameters. Input parameters and files are specified in a dedicated configuration file for each workflow. A contained software environment with all necessary dependencies can easily be created by command line using the conda package manager (https://conda.io).

Calculation of PSAP null distributions

To calculate PSAP null distributions, the user supplies a VCF file with allele frequencies for calibration and a file with the desired pathogenicity score. A bed file with the start and end of genes is supplied with the workflow, but any other specified genomic units can be used. It is recommended to provide an optional bed file containing a list of well-covered regions in the allele frequencies database. This file will be used to exclude variants that fall in regions that were not well-covered in the reference panel and for which the method cannot make an inference. Adding this information improves the reliability of PSAP’s results.

Parallelization allows the calculations to be run by chromosome, making the generation of null distributions efficient. First, tables containing all the necessary information to calculate the null distributions from the input files are created. Then, the main rule of the workflow “calculate_null_distributions_PSAP” is applied.

Temporary output files by chromosome are ultimately merged, which generates the two main output files of the workflow: PSAP null distribution tables for the autosomal dominant and autosomal recessive models. To improve the speed of PSAP, we also replaced the simulations of genotypes with exact probability computations.

Application of PSAP to patient data

Custom input parameters for this workflow include the desired version of CADD, the assembly, and the corresponding PSAP null distribution file, which can be an output from the previous workflow or the default provided null distributions with allele frequencies from gnomAD genome r2.0.1 and CADD v1.6.

Input files for the PSAP pipeline are: a VCF file and the corresponding pedigree (PED) file. It is preferable that the VCF provided as input for the PSAP pipeline has undergone Quality Control (QC). We recommend for this the R library RAVAQ (8), with its default parameters. RAVAQ also allows the split of multi-allelic variants on different lines of the VCF, which is necessary for the application of the PSAP pipeline. An optional input is a file containing the CADD annotation for the InDels from the VCF file, which can be obtained through a request on the CADD website (https://cadd.gs.washington.edu/).

Preprocessing steps include filtering the input VCF to keep only regions well-covered in the database used to construct the null distribution and annotation of the VCF using the VEP software (9). A bed file for the location of genes and the regions well-covered in the version of gnomAD genome used to construct our version of PSAP null distributions are supplied with the PSAP pipeline, but other custom bed file can be supplied. Variants that were not classified as PASS by GATK in gnomAD are also filtered out of the VCF.

The main rule of the workflow “apply_PSAP_calculations” calculates, for each individual, PSAP p-values for the dominant and recessive models. The pipeline parallelizes calculation by individual if the user supplies multiple cores for execution. One output file by individual recapitulates the variant with the maximal CADD score in the unit of testing chosen, its PSAP p-value, and other relevant information from the VEP annotation.

The last step of the pipeline is to create a report file merging individual results. If there are both cases and controls in the input VCF (status specified through the PED file), only case results are in the report file and variants are flagged as “validated” if they are absent from the controls.

Results

To illustrate the contribution of this new implementation of PSAP for the prioritization of pathogenic variants in the context of real NGS data, we inserted pathogenic well-reviewed coding SNVs from the ClinVar database (10) (n=4,593 variants for the autosomal dominant model and n=2,430 variants for the autosomal recessive model) into each of the 574 exomes of the FrEnch Exome Project (11) (Figure 6.2, see Additional files 1 and 2 for more details on the simulations and the complete list of variants respectively). We applied both the original and new implementations of PSAP to these simulated disease exomes, and ranked variants according to their PSAP p-value. The new implementation of PSAP shows a clear gain in performance, with 25% of variants reaching the top 1 compared to 0.4% previously for the autosomal dominant model, and 83% of variants reaching the top 1 compared to 24% previously for the autosomal recessive model (Figure 6.3). This gain in performance can mainly be explained by the update of null distributions using the latest versions of gnomAD and CADD, which makes them more accurate at evaluating pathogenic variants.

Table 6.1 also shows the efficiency and speed of Easy-PSAP when applied to the 574 FREX exomes, with parallelization on 20 cores.

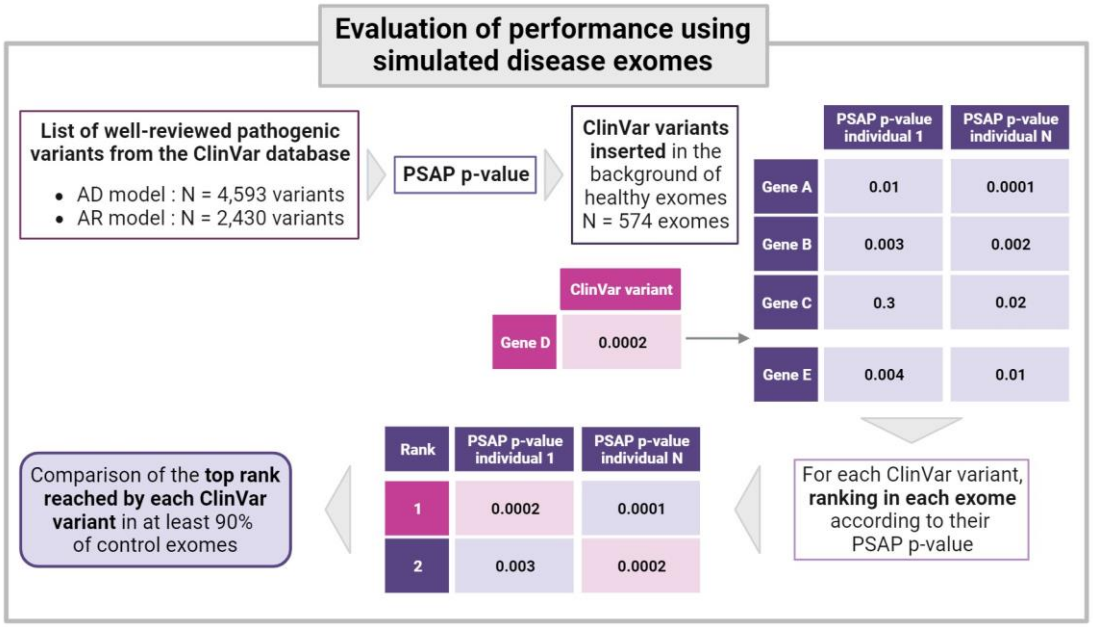


Figure 6.2. Flowchart of the evaluation of PSAP null distributions using simulated disease exomes

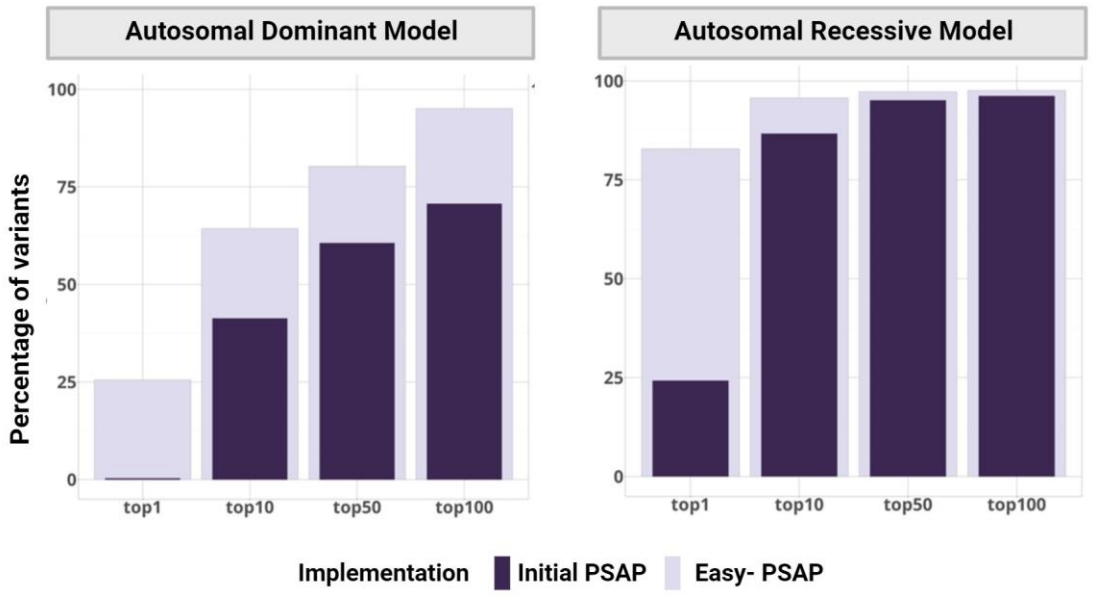


Figure 6.3. Gain in performance using the new implementation of PSAP (light purple) compared to the initial implementation (dark purple)

Percentage of pathogenic variants reaching the top N of variants according to the PSAP p-value in at least 90% of controls individuals, for the autosomal dominant and recessive models

Table 6.1. Time taken, memory consumption (RSS "Resident Set Size"), and CPU usage of Snakemake rules for the Easy-PSAP workflow to apply PSAP null distributions to 574 control exomes with parallelization on 20 cores

Rule name	**Description**	**Time (h:m:s)**	**Memory (GB)**	**CPU usage**
filter_regions	Filter out regions not well-covered in the database used to construct PSAP null distribution and sort VCF file	1:06:17	0.889	100% (1 core)
write_column_names	Write column names from VCF file	0:00:03	0.024	4% (1 core)
vep_annotation	Annotate VCF file with VEP software	0:25:08	67.980	2000% (20 cores)
filter_variants	Filter out low-quality variants in the database used to construct PSAP null distribution	0:03:29	9.485	100% (1 core)
apply_PSAP_calculations	Calculate PSAP p-values for each individual of the VCF file	0:03:02 by individual 1:26:00 for whole cohort	3.000 by individual	2000% (20 cores, 1 core for each individual)
make_report_file	Merges individual PSAP output files	0:21:07	3.111	100% (1 core)
compress_output_files	Compress PSAP output files	0:00:02 by file 0:01:00 for all files	0	1% (20 cores, 1 core for each file)

Conclusion

Easy-PSAP offers versatile pipelines to create custom null distributions of CADD scores and to score genetic variants of unknown significance in NGS data. These pipelines have been developed using the Snakemake workflow management system, which makes them user-friendly, computationally efficient, and reproducible. They come with pre-computed null distributions using up-to-date information from population databases and genetic annotations. PSAP null distributions in the GRCh38 assembly of the human genome, calibrated with allele frequencies from gnomAD V3(12) (comprising 76,156 genomes) and CADD v1.6 in GRCh38 are also made available with Easy-PSAP. These tools are adaptable to various scenarios, both in research and clinics and provide a strong framework for the prioritization of pathogenic variants at the scale of the genome.

Availability and requirements

Project name: Easy-PSAP

Project home page: https://github.com/msogloblinsky/Easy-PSAP

Operating system(s): Linux/MacOS

Programming language: R, Python, Shell

Other requirements: None.

License: None.

Any restrictions to use by non-academics: None.

References

1. Wilfert AB, Chao KR, Kaushal M, Jain S, Zöllner S, Adams DR, et al. Genomewide significance testing of variation from single case exomes. Nat Genet. 2016 Dec;48(12):1455–61.

2. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar;46(3):310–5.

3. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285–91.

4. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010 Sep 1;38(16):e164.

5. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.

6. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Medicine. 2021 Feb 22;13(1):31.

7. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Research; 2021

8. Marenne G, Ludwig TE, Bocher O, Herzig AF, Aloui C, Tournier-Lasserve E, et al. RAVAQ: An integrative pipeline from quality control to region-based rare variant association analysis. Genetic Epidemiology. 2022;46(5–6):256–65.

9. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016 Jun 6;17(1):122.

10. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4;46(Database issue):D1062–7.

11. Génin E, Redon R, Deleuze J, Campion D, Lambert J, Dartigues J, et al. The French Exome (FREX) Project: A Population-based panel of exomes to help filter out common local variants. Genetic Epidemiology. 2017;41(7):691–691.

12. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv; 2022

6.3 FUTURE DEVELOPMENTS

I plan to continue developing new features for Easy-PSAP in the future, with the help of people from my team at UMR1078 who will hopefully continue to use the pipeline for their own projects. For now, PSAP null distributions had been calculated using only SNVs. As the CADD scores of gnomAD InDels are now available, we would like to offer the possibility to include InDels in addition to SNVs for the calculation of PSAP null distributions which could improve the prediction of InDels pathogenicity by PSAP. A CADD pathogenicity score has also been recently developed for structural variants (Kleinert and Kircher 2022), so integrating them both to the construction and application of PSAP null distributions could be another step in the future as well.

Beyond the type of variants integrated in Easy-PSAP, another main limitation of the current state of the pipeline is the restriction to AD and AR models of inheritance. The initial PSAP article by (Wilfert *et al.* 2016) integrated the calculation of PSAP p-values for the compound heterozygote model. Other models, like the X-linked hemizygote model for males could be of interest for specific pathologies like male infertility. The scripts to calculate PSAP null distributions p-values for these additional models are already written and will soon be implemented within the Easy-PSAP workflows. An option to choose a specific allele frequency from a VCF file like gnomAD's to calculate population-specific PSAP null distributions, as it was mentioned above, would be another valuable addition to the pipeline.

Finally, most of the scripts featured in Easy-PSAP are written in R, which makes them quite computationally-intensive and long to run, especially for the calculation of PSAP null distributions on CADD regions. I will continue working on making them more efficient, mostly by switching to C++ for the most intensive calculations when possible.

Chapter 7 ANALYSIS OF CONSANGUINEOUS MALE INFERTILITY FAMILIES

To further explore the potential of PSAP in real-life cases of RDs, I applied the method to currently undiagnosed families affected by male infertility from Pakistan. Through this work, I was able to show how PSAP was able to perform and prioritize candidate variants, which led to identification of several candidate genes relevant to the disease.

7.1 BACKGROUND

As described in the section 4.3.2, male infertility is a prevalent issue affecting couples worldwide and research in the domain has shed light on the genetic origin of some cases of male infertility. While numerous genes have been linked to monogenic forms of male infertility, much remains unknown, including the identification of new genes and understanding of genotype-phenotype relationships. One obstacle lies in identifying patients with monogenic forms of male infertility. These conditions are typically caused by rare recessive gene variants, making them uncommon in randomly mating populations. However, populations practicing consanguineous marriages, especially among close relatives like in Pakistan, offer valuable insights for genetic studies. Indeed, in these populations, the increased frequency of homozygous genotypes in the population results in a higher incidence of recessive diseases. Consanguinity also influences fertility rates in Pakistan, correlating with longer times to first birth and smaller family sizes. In addition, increasingly available WGS data allows a thorough exploration of potential causative genetic variants, both coding and non-coding. Given these factors, our study sought to investigate Pakistani families with hereditary male infertility through WGS to uncover new genetic variants and genes implicated in this pathology.

In the article by (Khan *et al.* 2023), they describe the analysis of the 1st phase of the project, comprising of seven male infertility families from the Malakand of Pakistan. In five out of the seven families in question, they discovered potential genetic factors contributing to male infertility. These findings included a homozygous 10 kb deletion affecting exon 2 of the established male infertility gene *M1AP*. They also found biallelic missense substitutions in genes *SPAG6*, *CCDC9* and *TUBA3C*, as well as an in-frame hemizygous deletion in *TKTL1*, all genes with an emerging significance in the context of male infertility. Our aim was to replicate the findings in genes *SPAG6*, *CCDC9* and *TUBA3C* from this publication through our PSAP prioritization strategies, and analyze additional families from the 2nd phase of the project to contribute new candidate variants to explain the cases of male infertility. We also sought to corroborate our results by looking if our candidate variants were carried by homozygous

regions inherited from the same common ancestor present in cases but not controls, if the information was available.

7.2 Material & Methods

7.2.1 Datasets and quality control

Our data comprised of the two separate datasets for each phase of the project: the first phase will be referred to as Malakand phase 1 and the second phase as Malakand phase 2. WGS was carried out for all individuals. Overall, the Malakand phase 1 dataset included seven families with clinically diagnosed male infertility, among which five had a recent history of consanguinity. There was WGS data generated for 26 individuals (3-5 individuals per family, including cases and controls). For the Malakand phase 2, seven additional families with clinically diagnosed male infertility went through WGS, among which three families had a recent history of consanguinity. As for the phase 1, there were 26 WGS samples available (3-5 per family, including cases and controls). The pedigrees from the Malakand phase 1 families are included in (Khan *et al.* 2023), whilst the Malakand phase 2 pedigrees are featured in Figure 7.1.

The full description of the Malakand phase 1 WGS dataset can be found in (Khan *et al.* 2023). The same method has been applied to the Malakand phase 2, with an additional QC that was performed using R package RAVAQ (Marenne *et al.* 2022). To apply PSAP in optimal conditions, we performed a genotype and variant QC corresponding to standard GATK hard filtering criteria, genotypes with a depth < 10 or a genotype quality < 20 has been set to missing, as well as heterozygous genotypes with an allele balance outside [0.25-0.75]. No QC on variant call rates was applied. This corresponds to the default RAVAQ parameters for genotype and variant QC except that: MAX_AB_GENO_DEV = 0.25, MAX_ABHET_DEV, MIN_CALLRATE and MIN_FISHER_CALLRATE "disabled".

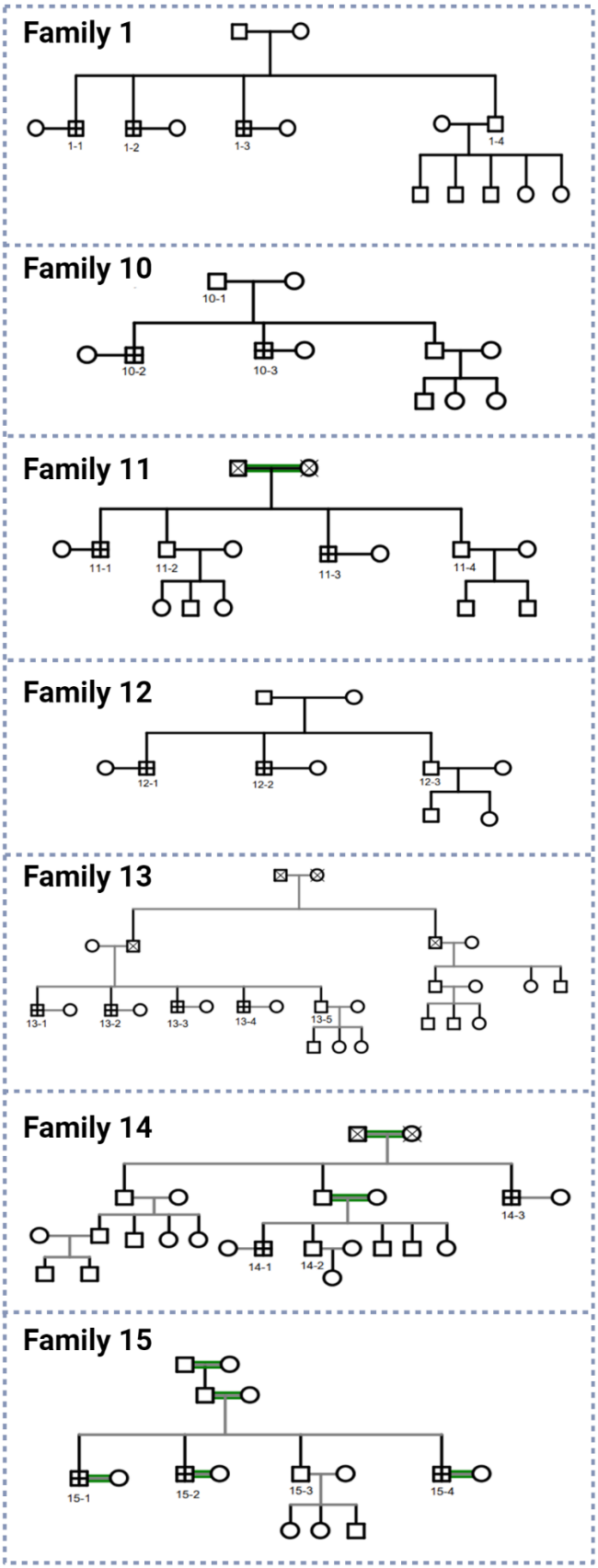


Figure 7.1 : Pedigrees of Malakand phase 2 families

Individuals that were sequenced are annotated with a number, affected individuals are marked with a cross

7.2.2 Prioritization strategy

To prioritize potentially candidate variants in the Malakand phase 1 and phase 2 datasets, we applied different PSAP strategies for hg38 data with our pipeline Easy-PSAP. InDels were not included in the analysis, considering the PSAP prioritization was less well calibrated than for SNVs (see section 5.2). Considering the initial version of PSAP on genes had been applied to the Malakand phase 1 data by the Conrad Lab team, we first applied PSAP-genes-CADD to both datasets in order to replicate the results. Then, as advised in our article describing the PSAP-genomic-regions strategy regarding WGS analysis (cf. section 5.2), we applied PSAP-coding-genomic-regions-CADD to prioritize potential coding pathogenic variants, and PSAP-genomic-regions-ACS to prioritize non-coding pathogenic variants. For each strategy, we focused on the top 100 variants according to PSAP p-values, which according to our simulations analyses would yield most of the causal pathogenic variants if they were included in the analyses.

Additional post-filtering was applied to provide only relevant candidates for male infertility for each phase separately. Only homozygous SNVs were kept (i.e. we only kept results from the AR model), which is in line with most causes of monogenic male infertility and the analysis of consanguineous individuals. Synonymous variants and variants seen in control individuals from any of the families were filtered out. To narrow down the pool of candidate variants even more and in accordance with the typical filtering process undergone at the Conrad Lab, variants were kept if they were not common or absent in gnomAD exomes v2.1.1 (maximum population-specific allele frequency - POPMAX < 0.01). Finally, we highlighted only variants in genes expressed in testis according to the GTEx tissue panel (GTEx Consortium 2020), and that were shared by cases from the same family if expected according to the pedigree.

7.2.3 Exploration of candidate genes

7.2.3.1 RNA sequencing data

Interpreting variants and the potentially affected genes in the context of their relevance to the specific disease at hand is a crucial part following the variant prioritization process. In order to give more weight or discard some of the prioritized variants from our analyses of the Malakand datasets, we exploited bulk and single-cell RNA-sequencing data. Traditional bulk transcriptomics measures the average gene expression levels across many cells, providing insights into the overall behavior of a tissue or population of cells. However, it lacks the ability to capture the variability and nuances present among individual cells. Single-cell transcriptomics, on the other hand, enables researchers to analyze the gene expression profiles of thousands of individual cells simultaneously within a heterogeneous

population. This technology allows for the identification of rare cell types, characterization of cellular heterogeneity within tissues, and elucidation of gene regulatory networks at a single-cell resolution.

For prioritized genes, we looked at the global expression of the gene in the body from bulk tissue expression as seen on GTEx (GTEx Consortium 2020) portal (https://www.gtexportal.org/). In line with the male infertility phenotype, we expected a plausible candidate gene to be expressed predominantly in testis. To explore even further candidate genes, we also used the Human Infertility Single-cell Testis Atlas (HISTA) (Mahyari *et al.* 2023) web-portal (https://conradlab.shinyapps.io/HISTA/). HISTA encompasses 26,093 high quality cells derived from testis biopsies (Mahyari *et al.* 2021) (2 juveniles, 6 normal adults, 1 adult with azoospermia, 1 adult with ejaculatory dysfunction, and 2 adults with Klinefelter Syndrome; Figure 7.2). This browser allows the exploration of gene expression signatures across different cellular populations that characterize normal testicular function and distinguish clinically distinct forms of male infertility. Both of these approaches allowed us to gain insight into the gene expression dynamics and the functional consequences of genetic alterations at different resolutions.

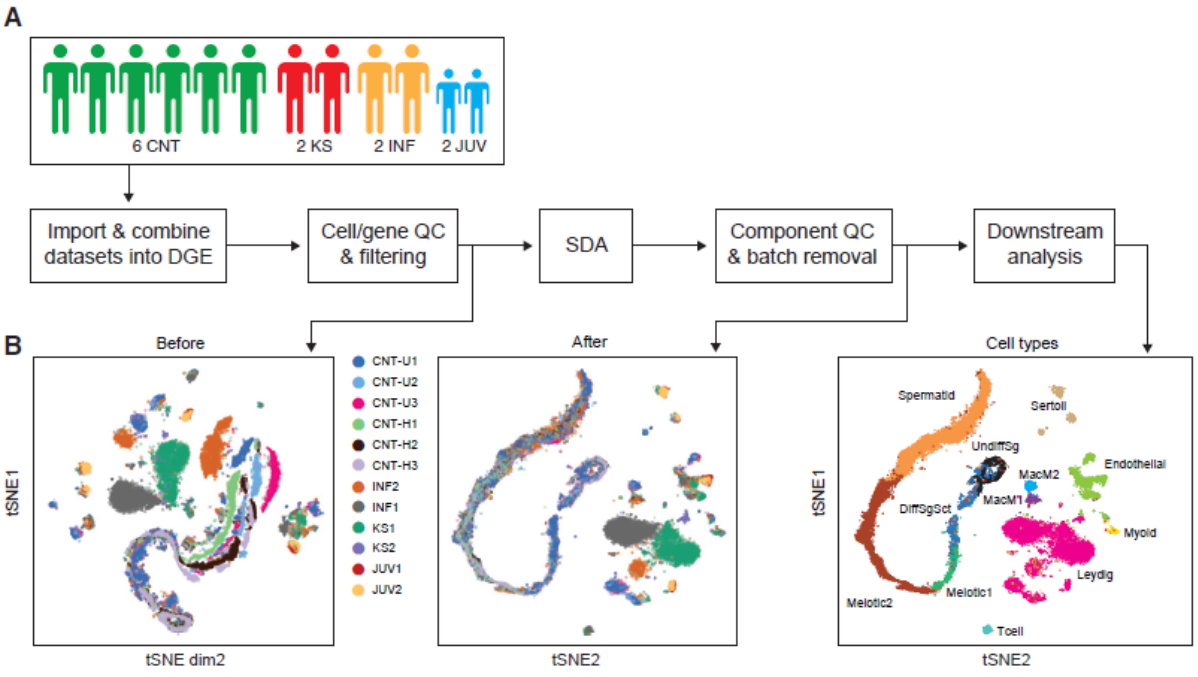


Figure 7.2 : Construction of the Human Infertility Single-cell Testis Atlas

From Mahyari et al. 2021 ; DGE: Digital Gene Expression matrix is produced; SDA: Sparse Decomposition of Arrays

7.2.3.2 Detection of HBD segments

To further confirm our variant prioritization results, we leveraged the background consanguinity from the Pakistani population through the detection of Homozygosity by Descent segments on the genome of the Malakand individuals. Indeed, in genetics, an inbred individual can inherit at a locus two copies of the same allele from a common ancestor of their parents. These alleles are said to be Identical by Descent (IBD), which results in Homozygosity by Descent (HBD) at that locus in the inbred individual. In case of a recessive disorder and known consanguinity, the initial assumptions are that the disease-causing variant is likely to reside within a HBD region.

However, information about IBD in the genome is not directly observable; it must be inferred from genotypes at markers. To do this, one must distinguish between IBD SNPs and homozygous but non-IBD SNPs. According to simulations with known pedigrees conducted using different estimation methods, it is recommended to use Hidden Markov Models (HMMs) on multiple subsets of markers to estimate and detect consanguinity in large samples (Gazal *et al.* 2014). A HMM is a statistical model where you have a sequence of observations, and each observation results from an unobserved state, referred to as the hidden state. The goal is to infer the sequence of hidden states based on the observations and a set of parameters. HMMs assume that, conditional on the IBD status, there is independence of the observed genotypes, which is not the case in the presence of LD between alleles of different markers. Taking subsets of markers (submaps) is a method to circumvent this problem (100 submaps recommended). There are two strategies for determining subsets: selecting subsets of markers separated by a certain genetic distance or subsets of markers between recombination hotspots (1 SNP per LD block).

For the analysis of the Malakand samples, we used the R package Fantasio (https://github.com/genostats/Fantasio). Fantasio leverages a statistical model to estimate the inbreeding coefficient f of an individual, and a parameter a, where af is the instantaneous rate of change per cM from no HBD to HBD. An individual was considered inbred by Fantasio if the median p-value of Likelihood Ratio Tests on good quality submaps is inferior to 0.05. Several factors had to be taken into account when applying Fantasio to the two Malakand datasets. First of all, our data was in the hg38 build. We thus needed to use a specific file with hotspots in the hg38 build, which we then used to calculate 100 submaps. Another specificity of our dataset was that we were working with WGS data instead of the SNP genotyping data usually used to carry HBD analysis. This meant we had a lot more variants than expected from the method. To mitigate that issue, we restricted our VCF files to variants with an allele frequency of at least 5% in the South Asian population of gnomAD genome v3.1.2.

7.3 RESULTS

7.3.1 Malakand phase 1 variant prioritization analysis

After QC, there were 13,449,466 out of 13,944,569 variants kept for analysis. Our prioritization and post-filtering strategy yielded 5 variants in 3 families, with their ranking from the PSAP-gene-CADD and PSAP-coding-genomic-regions-CADD strategies showed in

Table 7.1. Among these variants, 3 were the ones already identified in *SPAG6* (in family 3), *TUBA3C* (in family 7) and *CCDC9* (in family 4) in this dataset. In addition, we can put forward here additional variants that could have been considered for two of the families: family 4 and 3. Regarding the HBD regions analysis, according to the inbreeding coefficient calculated by Fantasio, the control individuals from family 3 and 4 were not inbred (individuals 3-5 and 4-4), as well as two cases from family 6 (individuals 6-1 and 6-2). This was surprising due to the history of consanguinity in both families. All other individuals for which WGS data were available were considered inbred.

NT Genomic Change (GRCh38)	Consequence	Amino acids	Protein position	Gene	CADD region	Family	ID	Rank PSAP-genes	Rank PSAP-genomic-regions	Rank PSAP-coding-genomic-regions
chr10: 22389235 C>T	missense variant	R/W	310	*SPAG6*	R083368	3	Mal3-1, Mal3-2, Mal3-3	**12,12,7**	**47,44,43**	**11,9,6**
chr22: 32864976 G>T	missense variant	S/Y	217	*SYN3*	R134775	3	Mal3-1, Mal3-2, Mal3-3	**26,19,14**	**1,2,2**	**122,115,107**
chr19: 45805018 T>C	splice acceptor variant	-	-	*RSPH6A*	R128508	4	Mal4-1, Mal4-2, Mal4-3	**2,1,4**	**2,3,4**	**1,1,1**
chr19: 47260609 C>T	missense variant	R/W	78	*CCDC9*	R128563	4	Mal4-1, Mal4-2, Mal4-3	**12,10,15**	**51,45,88**	**9,8,17**
chr13: 19177247 C>T	missense variant	G/R	246	*TUBA3C*	R100895	7	Mal7-1, Mal7-2, Mal7-3	**8,7,13**	**21,17,37**	**8,4,16**

Table 7.1 : Candidate diagnostic variants identified in Malakand phase 1 families

The *TUBA3C* variant in family 7 was the only prioritized variant in this family, and the low ranking for all cases with the three PSAP strategies (top 20 with PSAP-genes-CADD and PSAP-coding-genomic-regions-CADD, top 50 with PSAP-genomic-regions-CADD) confirmed the plausible association of this variant with the disease in this family. The *TUBA3C* gene encodes a protein called Alpha-tubulin 3C, which is a member of the alpha-tubulin family. Tubulins are structural components of microtubules, which are essential components of the cytoskeleton in eukaryotic cells. Microtubules play critical roles in various cellular processes, including cell division, intracellular transport, and maintaining cell shape. There is no prior mention in the literature of an association between *TUBA3C* and male infertility, and no ortholog of this gene exists in mice. The *TUBA3C* overlapped with HBD segments detected in cases but not in controls in family 7.

Regarding family 4, our analysis highlighted two potentially disease-associated variants. The first one was a missense variant in gene *CCDC9* that was in the top 20 with PSAP-gene-CADD and PSAP-coding-genomic-regions-CADD and which was the one already found in (Khan *et al.* 2023). *CCDC9* had already been identified as a candidate gene for severe asthenozoospermia (Sha *et al.* 2019) in an inbred case. In addition to the *CCDC9* variant, we prioritized a splice acceptor variant in gene *RSPH6A* with a high predicted VEP impact and a rank within the top 5 across the whole genome by all PSAP strategies. Rsph6a is a testis-specific protein essential for sperm flagellar assembly and flagellar motility. *RSPH6A* has not been associated with male infertility in humans, but its depletion has been reported as associated with male infertility in mice (Abbasi *et al.* 2018). Both *CCDC9* and *RSPH6A* are expressed in testis (Figure 7.3), although *CCDC9* is broadly expressed in other tissues too. While *RSPH6A* expression is enriched in spermatocytes, *CCDC9* expression in testis seems to be specific to very early spermatogonia and spermatids. Both variants are rare or absent from gnomAD. Interestingly, the two candidate variants are situated at 2 Mb of distance on chromosome 19. In family 4, genes *CCDC9* and *RSPH6A*, located 2 Mb apart on chromosome 19, fell within the same HBD region shared by all cases, which did not help prioritize one gene over the other. This confirmed the potential implication of *SPAG6* and *CCDC9* in the disease in family 4. However, we were not able to check if the segments were present in control individuals in this family as they were not inbred.

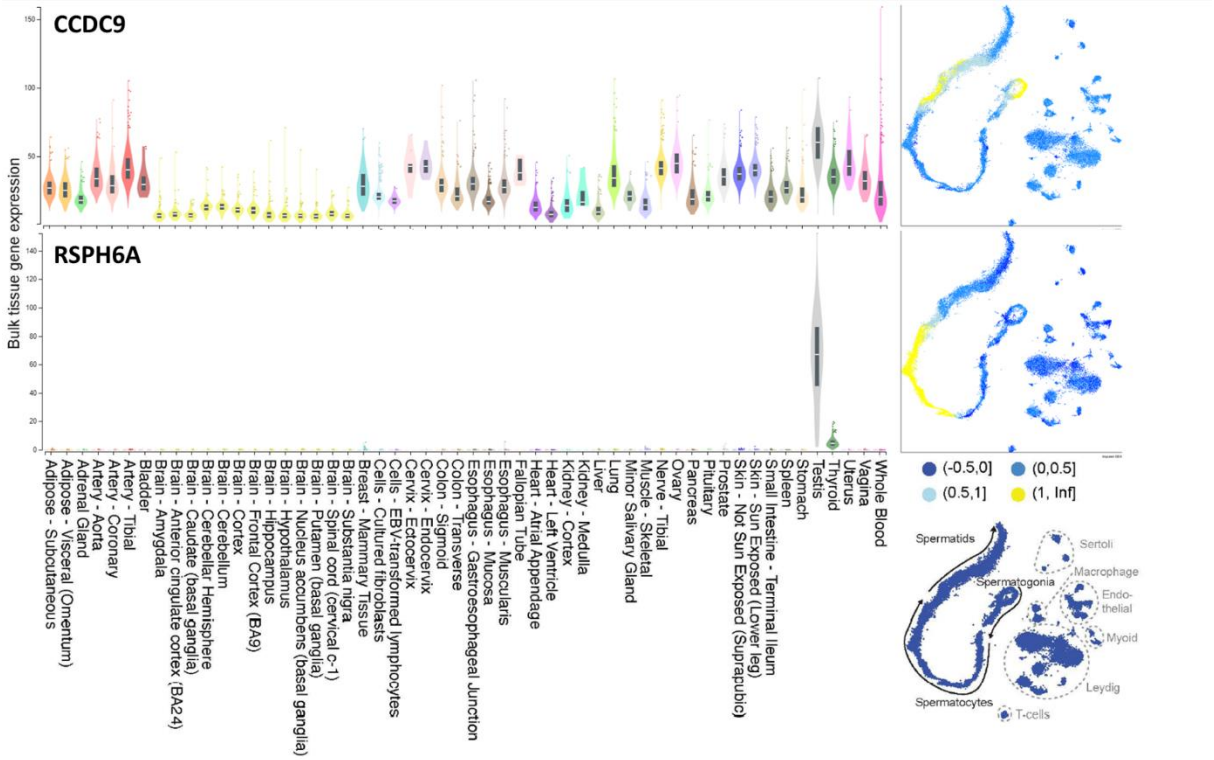


Figure 7.3 : Expression profiles of CCDC9 and RSPH6A candidate genes in family 4 from Malakand phase 1

The expression level of the gene (left column) across 54 tissues of the body, as defined by bulk RNA-sequencing by the GTEx project, and (right column) at a single-cell level across cells of the adult testis according to the HISTA browser are featured

In family 3, there were two candidate variants prioritized as well. One of the variants, a rare missense variant in *SPAG6*, had already been reported in (Khan *et al.* 2023) and was prioritized in the top 20 with PSAP-genes-CADD and PSAP-coding-genomic-regions-CADD, and in the top 50 with PSAP-genomic-regions-CADD. As mentioned in the article, *SPAG6* has been associated with male infertility in mice (Sapiro *et al.* 2002) as well as in humans (Wu *et al.* 2020; Xu *et al.* 2022), particularly in three separate cases of severe asthenoteratospermia. The second candidate variant was a missense variant in gene *SYN3,* absent from the gnomAD database. The variant in SYN3 was highly prioritized by PSAP-genomic-regions-CADD at first or second rank, but was not ranked in the top 100 in any of the cases by PSAP-coding-genomic-regions-CADD due to the higher ranking of other coding CADD regions. *SYN3* encodes the Synapsin-3 protein, which belongs to the synapsin family. The function of Synapsin-3 is primarily associated with regulating the release of neurotransmitters from synaptic vesicles. Synapsins are involved in tethering synaptic vesicles to the cytoskeleton. They also play roles in the formation and maintenance of synapses during development, as well as in synaptic plasticity. *SYN3* had no reported involvement in male infertility related phenotypes, in human or mice. *SPAG6* and *SYN3* expression was high in testis (Figure 7.4), with *SYN3* also largely expressed in the brain. *SPAG6* showed

a specific testis expression in spermatocytes and early spermatids, whilst *SYN3* expression seemed more restricted to spermatids. In addition, the *SPAG6* gene overlapped an HBD region shared by all cases on chromosome 10, while the *SYN3* was situated after but not within an HBD segment shared by all cases as well on chromosome 22, which confirmed the potential implication of SPAG6 in the disease.

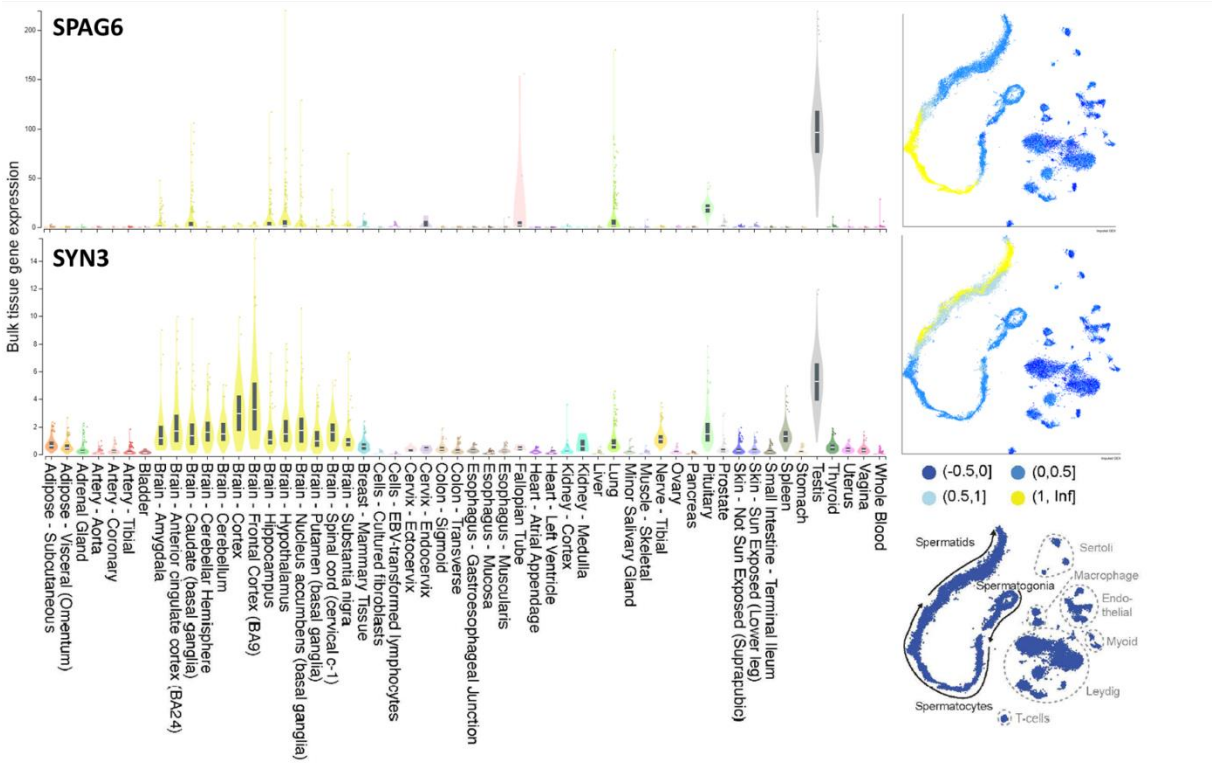


Figure 7.4 : Expression profiles of SPAG6 and SYN3 candidate genes in family 3 from Malakand phase 1

From the Malakand phase 1 dataset, family 6 and 9 had candidate variants as well that could not be analyzed with our pipeline as they were not autosomal SNVs: an in-frame mutation in the *TKTL1* gene and a deletion of *M1AP* exon two, respectively. This left families 5 and 8 without any candidate variants in the coding genome. For these two families specifically, we considered the prioritization by PSAP-genomic-regions-ACS without filtering on testis expression to not restrict to coding regions, i.e. to explore non-coding regions. No variant segregated with the disease phenotype in family 5 among the top 100 of PSAP results; two variants segregated with the disease phenotype in family 8. The interpretation of both variants was difficult, as one them was intergenic and the other an intronic variant. There were 3 homozygote carriers of the intergenic 12-38957630-A-G variant in the South Asian population of gnomAD (AF = 0.03), and 4 carriers for the intronic 3-15965671-A-C variant (AF = 0.04). They did not colocalize with regulatory elements described by ENCODE. Neither variants had significantly high scores with other *in sillico* prediction tools (philoP, GERP, SpliceAI).

7.3.2 Malakand phase 2 variant prioritization analysis

For the Malakand phase 2 dataset analysis, the QC process left 14,200,927 out of 14,702,449 variants for analysis. We found a total of 5 candidate variants that segregated with the phenotype and were plausible candidates to explain male infertility in cases (Table 7.2). One variant was observed in both cases from family 12, two variants in one case from family 14 (the two cases from this family were a nephew and an uncle) and one variant in two cases from family 15 (we only had WGS data from these two cases and were unable to confirm our findings in the last case from this family). Regarding the HBD regions analysis, the inbreeding coefficient from Fantasio indicated that all individuals from family 12 (control individual 12-1 and cases 12-2 and 12-3) and three individuals from family 11 (case 11-1 and controls 11-2 and 11-3) were not inbred. This result was more unexpected for family 11 for which, in contrary to family 12, there was recent consanguinity in the pedigree.

NT Genomic Change (GRCh38)	**Consequence**	**Amino acids**	**Protein position**	**Gene**	**CADD region**	**Family**	**ID**	**Rank PSAP-genes**	**Rank PSAP-genomic-regions**	**Rank PSAP-coding-genomic-regions**
chr13: 19174098 C>T	missense variant	R/Q	373	*TUBA3C*	R100895	12	Mal12-1, Mal12-2	**6,6**	**3,2**	**3,3**
chr10: 96016313 A>C	missense variant	E/A	1210	*CC2D2B*	R087600	14	Mal14-1	**NA**	**6**	**6**
chr10: 97002188 G>A	missense variant	P/S	1446	*SLIT1*	R087640	14	Mal14-1	**11**	**7**	**8**
chr16: 80549652 G>A	splice donor variant	-	-	*DYNLRB2*	R118954	15	Mal15-1, Mal15-2	**1,1**	**1,1**	**1,1**

Table 7.2 : Candidate diagnostic variants identified for in Malakand phase 2 families

The most striking finding from this analysis is a second distinct variant in *TUBA3C* found in family 12, in the same CADD region R100895 as for the variant found in family 7 from the Malakand phase 1. This *TUBA3C* variant was ranked in the top 10 with PSAP-genes-CADD, PSAP-genomic-regions-CADD and PSAP-coding-genomic-regions-CADD.

As Khan et al. pointed out, *TUBA3C* expression is restricted to testis, with a broad expression throughout spermatogenesis, from spermatogonia to round spermatids (Figure 7.5). They also mention a case from the GEMINI cohort which carried a homozygous missense change of the conserved residue of *TUBA3C* that was absent from gnomAD. Family 12 did not have a recent history of consanguinity in the pedigree which could explain this result, but it meant we could not explore HBD segments around the *TUBA3C* gene.

Another strong candidate variant was highlighted for family 15 in gene *DYNLRB2*. This splice donor variant in *DYNLRB2* with a predicted high impact was absent from gnomAD and prioritized at first rank with all PSAP strategies and in both cases of the family, whilst being absent in controls. The *DYNLRB2* gene encodes a protein called Dynein light chain roadblock-type 2. This protein is a component of the dynein complex, which is a motor protein involved in intracellular transport along microtubules. Mutations in the dynein genes have already been associated with asthenozoospermia of variable severity (Zuccarello *et al.* 2008). A recent article has shown that *DYNLRB2* was indispensable for spindle formation in meiosis I, and that its KO in mouse testes results in an arrest of meiosis progression (He *et al.* 2023). As for *TUBA3C*, *DYNLRB2* expression is specific to testis, from spermatocytes to spermatids (Figure 7.5). We were able to see through our analysis that both cases 15-1 and 15-2 from family 15 shared an HBD segment on chromosome 16 that encompassed gene *DYNLRB2* and that was not shared with control individual 15-3 from the same family.

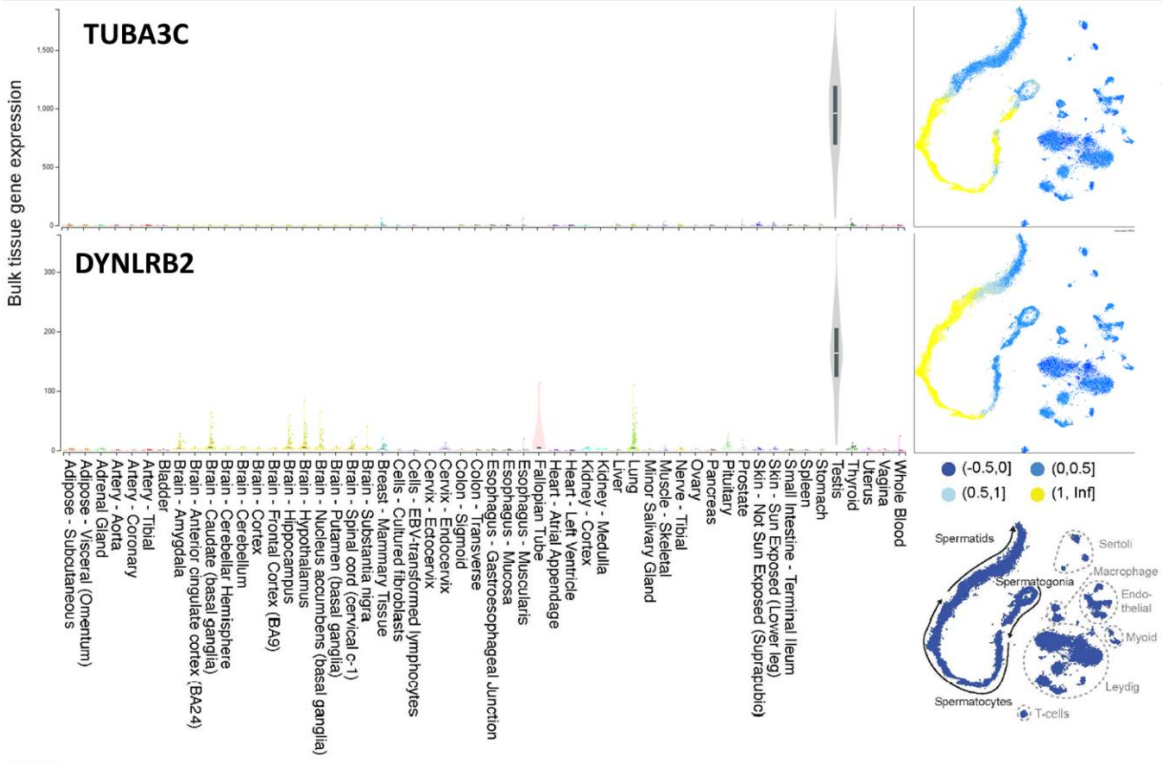


Figure 7.5 : Expression profiles of TUBA3C and DYNLRB2 candidate genes from Malakand phase 2

Finally, we were also able to prioritize two candidate variants in individual 1 from family 14 (individual 14-1), which were not present in controls but neither in the other affected individual from the family (individual 14-3, uncle of individual 14-1). The two variants were missense variants, ranked in the top 10 will all PSAP strategies, in genes *CC2D2B* and *SLIT1*, which are located 2 Mb apart on chromosome 10. The *CC2D2B* gene encodes a protein known as coiled-coil and C2 domain-containing protein 2B. This protein is involved in various cellular processes, particularly in ciliary function and regulation. Mutations in the *CC2D2B* gene have been implicated in a spectrum of ciliopathies, which are a group of disorders characterized by defects in cilia structure or function. Although *CC2D2B* has not been associated with male infertility in human or mice, studies have shown the link between ciliopathies and male infertility. Indeed, sperm flagella and motile cilia share similarities in structure and function including the axonemes, which are internal structure necessary for motility. Defects in the axonemal structure can thus disrupt the motility of both cilia and flagella (Sironen *et al.* 2020). The *SLIT1* gene encodes a protein called Slit homolog 1. *SLIT1* plays a crucial role in various developmental processes, particularly in axon guidance and neuronal migration during embryonic development. *SLIT1* functions by binding to its receptors, particularly roundabout (ROBO) receptors, which are present on the surfaces of neurons and other cell types. The binding of SLIT1 to ROBO receptors triggers intracellular signaling pathways that regulate cytoskeletal rearrangements and cellular responses. Mutations or dysregulation of the *SLIT1* gene have been implicated in various developmental disorders, including certain congenital malformations of the brain and spinal cord. Although we have not found direct reported links between *SLIT1* and male infertility phenotypes, Slit/Robo has recently been identified in mice as a novel signaling mechanism that regulates Leydig cell steroidogenesis which can have impacts on fertility (Martinot and Boerboom 2021). Although these two variants could be suitable candidates to explain the disease in individual 14-1, we will not expand on them further considering the evidence for their pathogenicity was less strong than for the prioritized variants in the coding genome.

7.4 Discussion – Interpreting Easy-PSAP outputs in a diagnostic setting

An important aspect of any variant prioritization strategy is the interpretability and accuracy of its results. Below, I discuss some limitations and strength of Easy-PSAP when applied in a real-life setting.

7.4.1 Post-Easy-PSAP filtering strategies

As evidenced by our evaluation of Easy-PSAP on artificially-generated disease exomes and genomes in Part III - Chapter 5, there is no universal cutoff or evident significance threshold for PSAP p-values. 1kGP Non-Finnish Europeans individuals from the general population carried variants with low PSAP p-values in their genome. The lowest PSAP p-value for Non-Finnish Europeans individuals ranged from 10^{-4} to 10^{-5} for the AD model, and from 10^{-3} to 10^{-10} for the AR model.

Thus, we decided to use PSAP as a prioritization tool, which allowed us to rank variants based on their PSAP p-value and not to use a threshold on PSAP p-values. This strategy proved less restrictive than filtering on PSAP p-values which distributions can differ drastically from one individual to another, but did not reduce the number of variants to analyze. In our simulations, we looked sequentially to the top reached by the ClinVar variants, going from the top 1 (most prioritized variant, with the lowest PSAP p-value), to the top 10, top 50 and finally top 100. Not going beyond the top 100 was an arbitrary choice which was also based on the number of variants one could feasibly explore and interpret in a clinical setting. For our analysis of the Malakand datasets, we combined the top 100 variants prioritized by PSAP with a range of filters that fit with the expected disease-causing variant and leveraged familial data, which helped narrow down significantly the number of candidate variants. We advise a similar approach when applying PSAP to other pathologies.

Finally, we have to point out that PSAP performs better to prioritize pathogenic variants for the AR than for the AD model. Indeed, our simulations also show that if the pathogenic variant is a SNV with an AR model of inheritance, it will almost always be prioritized in the top 10 of the genome by PSAP-genomic-regions. However, if the model is an AD model of inheritance, looking at the top 100 will not encompass all of the pathogenic SNVs. There is not clear explanation for that pattern, which has already been observed in (Wilfert *et al.* 2016) for the initial PSAP. However, it has been shown that AR disorders are overwhelmingly linked with Loss-Of-Function (LOF) mechanisms, whereas AD disorders can cause disease through different mechanisms (Gerasimavicius, Livesey and Marsh 2022). These molecular mechanisms can include LOF, but can also be dominant-negative (expression of a mutant protein interfering with the activity of a wild-type protein) or gain-of-function effects (e.g. constitutive activation of the protein, shift of substrate or binding target specificity, protein aggregation). The latter two mechanisms have limited effects on the protein structure and may not be captured as accurately by the CADD score. In addition, although PSAP is able to score InDels if their

CADD score is provided, we have not constructed PSAP null distributions with InDels and thus preferred to not include them in the analysis. Indeed, InDels often have a higher predicted deleteriousness than SNVs and were thus overly prioritized with PSAP compared to SNVs when we included them.

7.4.2 Male infertility candidate variants

In our analysis of the Malakand phase 1 and phase 2 datasets, we were able to both confirm candidate variants and put forward new candidate variants. Regarding the Malakand phase 1 families, the *TUBA3C* variant was the only one we identified in family 7, in line with previous reporting by (Khan *et al.* 2023). In families 3 and 4, we prioritized the already candidate variants in *SPAG6* and *CCDC9*, respectively. For family 3, we identified another variant in *SYN3* that passed our filtering criteria. However, the *SPAG6* variants seems like a stronger candidate due to the large evidence in the literature of the repeated associations of this gene with male infertility and the more favorable PSAP rankings. Another candidate variant was identified in family 4, in gene *RSPH6A*. Both *RSPH6A* and *CCDC9* seem good candidates to explain the male infertility phenotype, although the *RSPH6A* variant has a more severe predicted impact. The exploration of HBD segments in these individuals showed that both *RSPH6A* and *CCDC9* variants, which were 2 Mb apart on chromosome 19, were carried by the same HBD segment shared by all cases and not controls. It was also the case for the *RSPH6A* variant but not for the SYN3 variant in family 3. This confirmed the potential implication of *SPAG6* and *CCDC9* or *RSPH6A* in the disease in family 3 and family 4, respectively. Two other families (family 6 and family 9) from the Malakand phase 1 had candidate variants that were not included in our analyses, and none of our PSAP strategies prioritized a candidate variant in these families. This supports the idea that a careful processing of PSAP results can remove false positive results, especially when familial data is available. Due to the challenge of detecting HBD segments in WGS data and the fact that some cases and controls were not identified as inbred, we did not try to pursue further homozygosity mapping strategies. However, the HBD segments we were able to detect in the data for some families confirmed the interest of some of our candidate variants from our variant prioritization strategy.

Our analysis of the Malakand phase 2 dataset allowed us to prioritize another variant in *TUBA3C*, in the same CADD region than the previous variant from the Malakand phase 1 analysis. This CADD regions overlaps with the Tubulin/FtsZ family, GTPase domain of the protein, which is an evolutionary conserved protein domain (Nogales *et al.* 1998). Overall, this second highly prioritized variant confirms the potential implication of alterations of *TUBA3C* in male infertility in these two families. In addition, we highlighted a strong candidate variant for family 15 in gene *DYNLRB2*. Although we were not able to detect HBD segments in family 12, we were able to show that all

sequenced affected individuals from family 15 shared an HBD segment that encompassed *DYNLRB2* and that was not shared with the control individual from the same family.

A striking observation that could tie together multiple of our previous results comes from (Jumeau *et al.* 2017). They curated the human sperm microtubulome, which encompasses genes encoding proteins present in the sperm and that are associated with cytoskeleton of the microtubule. When looking at protein–protein interactions, they produced a disease-interaction network composed of 50 genes, which also integrated various transcriptomics, proteomics, and interactomics data (Figure 7.6). Three of the genes previously described in our analysis of the Malakand datasets could be found in this network: *DYNLRB2*, *TUBA3C* and *SPAG6*. Interestingly, *SPAG6* and *DYNLRB2* were highlighted for their association with the HPO term "Abnormal sperm motility" (HP:0012206), and were described as having improved the disease-interaction network by highlighting other critical nodes. Then, when trying to extract cluster of densely connected nodes in the network, the authors found a cluster comprising of 10 genes centered around *CUL3*, including *DYNLRB2* and *TUBA3C*. This cluster significantly associated with several Gene Ontology (Ashburner *et al.* 2000) terms pertaining corresponding processes, like "signal transduction", "cell communication", "transport", "spindle", "vesicle", and "extracellular region". This gives even more weight to some of our candidate variants, which could lead to perturbation of the microtubule cytoskeleton that impact sperm motility and thus lead to asthenozoospermia or other male infertility phenotypes.

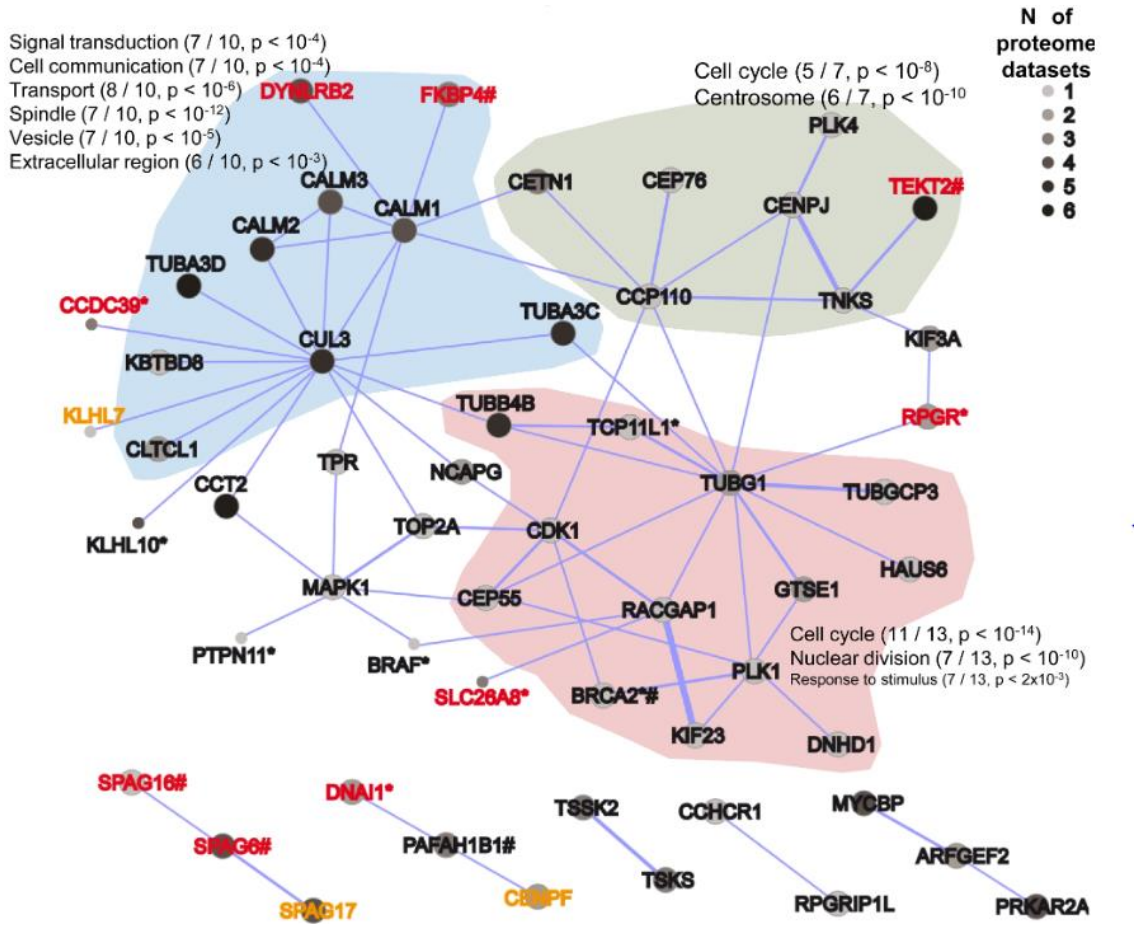


Figure 7.6 : Disease-interaction network for the human sperm microtubulome

From Jumeau et al. 2017 ; Genes associated with the HPO terms “Abnormal sperm motility” or “Abnormal ciliary motility” were color-coded in red and orange, respectively. An asterisk (*) after a gene symbol indicates an association with the HPO term “Male infertility”. Human genes having mouse orthologs associated with abnormal sperm motility, abnormal motile cilium morphology and physiology and male infertility are indicated with a hash sign (#)

Functional analyses of these variants would be needed to conclude with certainty about their implication in the male infertility phenotype in these individuals. Sequencing of other family members, like affected individual 4 from family 15, could also confirm the segregation of variants and give more weight to candidate variants in some families. Exploration of the impact of the variants in genes currently not associated with male infertility in the literature (e.g. *TUBA3C*) could help unravel new pathways and etiologies for this disease, about which there is still a lot to uncover through innovative strategies (Ding and Schimenti 2021).

Part IV

Assessing a complex mode of inheritance: the digenic model

Chapter 8 METHODS TO DETECT DIGENISM IN SEQUENCING DATA

We have described in Part II - Chapter 4 the complex genetic architecture of RDs, including the possibility that some RDs could be characterized by digenic inheritance (DI) at a molecular level. Although detecting digenic inheritance could help diagnose a number of RDs, it is no easy task and no gold standard method exists. In order to ultimately provide our own answer to the question, I put together in the last part of this thesis a review of the literature and benchmark of the currently published methods to detect digenic inheritance in sequencing data. The supplementary materials for this manuscript can be found in Appendix IV. This allowed an overview of the strength and limitations of the methods developed to date, and led to the proposition of new methodological developments that will be mentioned in the discussion.

8.1 BACKGROUND AND SUMMARY

Most of the currently described cases of DI were detected through familial analyses, and often involved genes already associated with the disease in a monogenic way. The databases DIDA (DIgenic diseases DAtabase) and its updated version OLIDA (OLIgenic diseases Database) report all of the known digenic and oligogenic combinations from the literature. Since 2018, a number of computational methods have thus been developed to approach the issue of detecting DI without relying as much on pedigrees and prior knowledge on the disease. In the articles describing each method, little to no comparison is made to the other methods developed to date achieving the same objective. Here, we offer an overview of all of the published and available methods to detect DI in sequencing data, and offer insight on the strength and limitations of the methods in a benchmark setting.

First, through a review of the literature, we categorized the methods to detect DI in three distinct categories: network-based, statistical and machine-learning (ML). The only method in the network-based category is OligoPVP (Boudellioua *et al.* 2018) which uses PVP (Boudellioua *et al.* 2019), a monogenic predictor of pathogenicity based on HPO terms, to score variants only in pairs of genes connected in a protein-protein interaction network. Two methods fall under the statistical method category: the Digenic Method (DM) (Kerner *et al.* 2020) and RareComb (Pounraja and Girirajan 2022). On one hand, the DM uses an adapted burden tests to detect DI or a common modifier variant for a monogenic disease. On the other hand, RareComb uses the Apriori algorithm (Agrawal *et al.* 1996) to enumerate simultaneous combinations of variants in cases and controls. Both methods involve prefiltering variants to limit the number of combinations to test, and keep only rare variants predicted as pathogenic. All of the other methods to detect DI can be classified as ML methods and are trained

on pathogenic pairs from DIDA or OLIDA and negative neutral pairs from the general population with different features. Among these methods, VarCoPP (Papadimitriou *et al.* 2019) and its successor VarCoPP2.0 (Versbraegen *et al.* 2023) which are run through the platform ORVAL (Renaux *et al.* 2019) take as input pairs of variants, whilst DiGePred (Mukherjee *et al.* 2021), DIEP (Yuan *et al.* 2022) and ARBOCK (Renaux *et al.* 2023) are applied on pairs of genes.

As these ML methods were the most easily applicable, especially in the case of RDs, and had similar input and output types, we sought to compare them through a benchmark. We selected known pathogenic gene pairs from the OLIDA database that were not used to train the ML methods, and different scenarios of neutral gene pairs by using variants from the FREX database. ARBOCK and DIEP categorized more of the OLIDA pairs as digenic (more than 80%), but ARBOCK especially also classified around 30% of pairs observed in the general population as pathogenic. In contrary, DiGePred and VarCoPP2.0 classified slightly less OLIDA pairs as pathogenic (64% and 49%, respectively), whilst keeping the false positives number much lower (less than 1% and 3% respectively).

Overall, we were able to show that network-based and statistical methods make strong assumptions on the interaction between the genes or the type of variants to detect DI. The statistical methods also expect cohort types of datasets as input to have sufficient statistical power. In contrast, ML methods presented key advantages as they could be applied at an individual scale to detect DI, which is relevant in the case of very heterogeneous RDs, although no method could exhaustively analyze all potential pairs of genes from an individual. Our benchmark of the ML methods to detect DI allowed us to emphasize the DiGePred method, which stood out as having by far the smaller number of false positives in any scenario, whilst keeping a substantial amount of true positive predictions.

8.2 RESULTS

Computational methods to detect digenism in sequencing data: a comprehensive review and benchmark

Marie-Sophie C. Ogloblinsky[1,*], The FrEx Consortium, Donald F. Conrad[2], Emmanuelle Génin[1,3,¶], Gaëlle Marenne[1,*,¶]

[1]Univ Brest, Inserm, EFS, UMR 1078, GGB, Brest, France

[2]Division of Genetics, Oregon National Primate Research Center, Oregon Health & Science University, Portland, Oregon, United States of America

[3]Centre Hospitalier Régional Universitaire de Brest, Brest, France

*Corresponding authors:

Email: marie-sophie.ogloblinsky@inserm.fr (M-S.O.); gaelle.marenne@inserm.fr (G.M.)

¶These authors contributed equally to this work

Abstract

Digenic inheritance is characterized by the combined alteration of two different genes leading to a disease. It could explain the etiology of many currently undiagnosed rare diseases. With the advent of next-generation sequencing technologies, the identification of digenic inheritance patterns has become more easily feasible, yet still poses significant challenges without any gold standard method.

Here, we present a comprehensive review of the existing methods developed to detect digenic inheritance in sequencing data. We systematically categorize methods by their type, and discuss their availability, output and scalability to inform potential users. Specifically, focusing on machine learning approaches to detect digenic inheritance, we propose a benchmark using different real-life scenarios involving known digenic and putative neutral pairs of genes.

We provide a classification of the methods to detect digenic inheritance in sequencing data in the following categories: network-based, statistical, and machine learning methods. The latter two types of methods appeared the most applicable to rare diseases. When assessing the performance of the machine learning methods, DiGePred stood out as the method with the highest predictive performance, followed by VarCoPP2.0.

By synthesizing the state-of-the-art techniques and providing insights into their practical utility, this review and benchmark serve as a valuable resource for researchers and clinicians in selecting suitable methodologies for detecting digenic inheritance in a wide range of disorders using sequencing data.

Keywords: digenism, machine learning, benchmark, rare diseases

Introduction

A classical paradigm in the study of human diseases has been that rare diseases were caused by a single variant; either in the heterozygous or homozygous state, altering a single gene, also known as monogenic inheritance. The wide-spread use of high throughput-sequencing has allowed the diagnosis rate of rare diseases to reach 30 to 50% of cases, corresponding almost entirely to monogenic cases (Boycott *et al.* 2017, 2019). However, in the last 30 years, the monogenic inheritance hypothesis has been proven wrong by several cases of rare diseases following more complex genetic inheritance patterns (Lupski 2012). The simplest case of complex genetic inheritance is the digenic model, under which the presence of damaging variants in two different genes is necessary for the disease to develop (Schäffer 2013). Each variant taken separately is not sufficient for the disease manifestation. Epistasis refers to the interaction between genes or loci where the effect of one gene or allele modifies the phenotypic outcome of another gene or locus (Cordell 2002). Initially, the term “epistatis” was used to describe some forms of digenic inheritance although it came to include a broader range of locus-locus interactions in polygenic diseases, including those identified through genome-wide association studies. In this review, we will only look at the specific case of digenism and not epistasis in general. DI has to be distinguished from the monogenic inheritance with a modifier gene scenario, according to which a variant in one gene is sufficient to cause the disease and the severity is modulated by a variant in a second gene (Génin, Feingold and Clerget-Darpoux 2008; Rahit and Tarailo-Graovac 2020).

The first description of digenic inheritance (DI) in the literature was for retinitis pigmentosa (Kajiwara, Berson and Dryja 1994). The first instance of DI described in the literature was for retinitis pigmentosa (Kajiwara, Berson and Dryja 1994) in 1994 and involved two photoreceptor-specific genes: *ROM1* and *peripherin/RDS*. Data from three pedigrees and the known interaction between the proteins produced by the two genes in an intermolecular complex supported this conclusion.

There were not many other examples of DI described until 2001, when an afflux of reports on DI were published, notably in Barbet-Biedl (BBS) syndrome (Katsanis *et al.* 2001, 2002; Beales *et al.* 2003; Fauser, Munz and Besch 2003). Once again, the authors based their conclusion on pedigree data from eight families, and proposed that three variants in two known BBS genes (*BBS2* and *BBS6*) could explain the phenotype in some of the families. Digenism thus has been involved in the etiology of several rare diseases (Cerrone *et al.* 2019; Kim *et al.* 2019; Jiang *et al.* 2020; Teles e Silva *et al.* 2022), and has been proposed as an hypothesis to some of the currently unsolved rare disease cases (Rahit and Tarailo-Graovac 2020). Recently, DIDA (Gazzo *et al.* 2016) (DIgenic diseases DAtabase) and its successor OLIDA (Nachtegael *et al.* 2022) (OLIgenic diseases Database), for oligogenic models involving more than two genes, were created as comprehensive curated resources dedicated to cataloging and documenting digenic and oligogenic diseases reported in the scientific literature. These databases play a crucial role in cataloging and organizing relevant information, thereby facilitating further investigation into the molecular mechanisms of digenic and oligogenic diseases. A common feature of the great majority of reports on DI to this day is the reliance on familial data from multiple pedigrees and also the detection DI involving at least one gene already known as implicated in the disease.

Despite great advances in next-generation sequencing (NGS) technologies as well as the availability of databases like DIDA and OLIDA, the detection of DI remains a challenging issue. Traditional approaches for rare disease variant discovery like variant filtering and classification through ACMG guidelines (Richards *et al.* 2015) are not suited for digenic pairs discovery as they were developed for highly-penetrant monogenic disorders. The relatively small number of digenic or oligogenic pairs described in the literature to this day can be owned partly to the limited disease population sizes, genetic heterogeneity, low frequency at which any particular pair of alleles is present in the population, and incomplete clinical descriptions. Moreover, assessing the interaction, which is the extent and manner in which two causes of a disease modify the strength of one another, between variants within different genes requires sophisticated bioinformatic and functional validation methods (Kousi and Katsanis 2015).

These variants may act through diverse molecular mechanisms, including protein-protein interactions, shared signaling pathways, or complementary functional roles within cellular processes. This challenge of detecting DI has been tackled by several computational approaches over the years (Boudellioua *et al.* 2018; Papadimitriou *et al.* 2019; Renaux *et al.* 2019, 2023; Kerner *et al.* 2020; Mukherjee *et al.* 2021; Yuan *et al.* 2022; Versbraegen *et al.* 2023; Gravel *et al.* 2024) with various and sometimes analogous strategies.

Here, we present a comprehensive review of the modalities, strengths and limitations of computational methods published to date to detect DI in NGS data. We separated the methods in three categories depending on the way they infer DI: network-based, statistical and machine learning (ML) methods. We also performed a benchmark of the methods falling into the machine-learning category which have become the most prominent in recent years (Okazaki and Ott 2022). These methods have never all been compared to each other. We devised an evaluation protocol to test the sensibility and specificity of each method, as well as their ability to distinguish between a digenic and a monogenic model. Our evaluation protocol leverages known digenic pairs from OLIDA, neutral variants from the French general population FrEnch EXome (FREX) project and pathogenic variants from the ClinVar (Landrum *et al.* 2018) database. The results from this review and benchmark will help users to make informed decisions about their choice of methods to detect DI. Appreciating the limitations of current methods to detect DI is a crucial step towards uncovering the molecular mechanisms behind digenic inheritance. A growing understanding of DI will likely provide researchers with new insights into the genetic basis of human rare diseases.

Currently available computational methods to detect digenic inheritance in sequencing data

This review is focused on computational methods applicable to sequencing data that do not need large pedigrees to make inferences on DI. We categorized existing methods to detect DI in three categories, depending on the way they assess DI: network-based, statistical and ML methods (Table 8.1). The only network-based method is OligoPVP. It assesses digenism depending on the connection between genes in a network. Statistical methods leverage different statistical modelling to find a burden of rare variants or enrichment in pairs of genes, with or without comparison to control data. ML methods learn features from known digenic pairs described in the literature (catalogued in DIDA or OLIDA) and observed pairs from the general population to predict if a new pair or variant or genes is likely to be pathogenic. The reliability of theses known digenic pairs used to train the methods will be discussed below. The ML method ARBOCK differs slightly from the other ML methods as it uses a decision set model to predict potential pathogenic gene interactions based on a heterogeneous knowledge graph.

Table 8.1. Main characteristics of the computational methods developed to detect DI

Category	Name	Availability	Output type	Scalable	Reference
Network-based method	**OligoPVP**	Command line tool	score {0-2}	√	Boudellioua et al. (2018)
Statistical method	**Digenic Method**	R scripts	p-values for each variant-gene or gene-gene pair	√	Kerner et al. (2020)
	RareComb	R package	list of statistically significant combinations that meet the user-specified input criteria	√	Pounraja et al. (2022)

Machine Learning	**VarCoPP**	Online platform (ORVAL)	**final class**: "neutral" or "disease-causing" **SS**: percentage of the classifiers agreeing about the pathogenic class, score {0-100} **CS**: median probability among individual predictors that pair is pathogenic, score {0-1}	×	Gazzo et al. (2017) Renaux et al. (2019)
	DiGePred	Precomputed list	score {0-1}	√	Mukherjee et al. (2021)
	DIEP	Precomputed list	score {0-1}	√	Yuan et al. (2022)
	VarCoPP2.0	Online platform (ORVAL)	score {0-1}	×	Versbraegen et al. (2023)
	ARBOCK	Command line tool	score {0-1}	√	Renaux et al. (2023)

Network-based method to detect digenic inheritance

One of the very first computational method aimed at detecting DI in sequencing data is OligoPVP (Boudellioua *et al.* 2018), which we categorized as a network-based method. This method combines a monogenic predictor of pathogenicity, the PhenomeNET Variant Predictor (PVP) version 2.0 ("DeepPVP") (An *et al.* 2022) with the information from a gene network database, STRING (Szklarczyk *et al.* 2017). Briefly, if the genes are connected in the STRING network, the PVP scores of their variants will be added to make a score by pair of genes. It is also possible to explore triplets of variants. The input file is a VCF file, along with an OMIM (Amberger *et al.* 2015) ID of the disease or a list of phenotypes (HPO (Robinson *et al.* 2008) or MPO (Smith, Goldsmith and Eppig 2005) terms).

The method was tested on a set of 189 synthetic whole genome sequences created by inserting known digenic pairs from DIDA v1 (Gazzo *et al.* 2016) in the background of genomes from the 1000 Genomes Project (Auton *et al.* 2015) and compared to monogenic predictors of variant pathogenicity, including PVP alone, CADD v1.3 (Kircher *et al.* 2014), DANN v1.0 (Quang, Chen and Xie 2015), GWAVA v1.0 (Ritchie *et al.* 2014) and Genomiser v7.7.2.1 (Smedley *et al.* 2016). In the article describing the method, OligoPVP performs less well than PVP to rank all biallelic pairs in the top 1, and less well or equally to PVP, CADD and DANN to rank them in the top 10. OligoPVP only outperforms all other methods when considering the 71 pairs for which there is a known biological interaction between genes. Biological interaction can be defined in this case as a situation in which the qualitative nature of the mechanism of action of a factor is affected by the presence or absence of the other. The performance of OligoPVP also drops when predicting 45 new DIDA pairs with their HPO terms that were added after OligoPVP was trained. The strength of this method is its applicability to all variants from a patient's exome, through a user-friendly command line tool. A main limit is its reliance on HPO terms and background interaction knowledge. Indeed, if the pair of genes is not connected through the 989,998 interactions present in the network, it will not be prioritized. In addition, the method relies on the phenotypic characterization of the patient and not only the genetic data, and thus is biased towards known phenotype-gene or phenotype-disease associations.

Statistical methods to detect digenic inheritance

Two methods can be classified as statistical methods developed to detect DI. The first one is the Digenic Method (Kerner *et al.* 2020) (DM), which adapts the idea of a burden test to the case of detecting true digenism or finding a modifier variant for known monogenic conditions. Rare candidate variants selected based on different parameters like their predicted impact or Minor Allele Frequency (MAF) are collapsed within the unit of a gene. Individuals are "carriers" for a specific gene if they carry at least one candidate variant in the gene under the studied mode of inheritance.

Then, a case-only logistic regression model is applied to test for interaction between the two genes carrying rare variants, or a gene carrying rare variants and a common variant. The method was evaluated through simulations using the 1000 Genomes Project and Human Genetics of Infectious Diseases (HGID) databases and on real data from a craniosynostosis WES dataset, in which the only significant result of the DM was the known digenic pair between gene *SMAD6* harboring rare variants and common variant rs1884302. R and bash scripts are available to run the method, but not in the form of a R package. Input files include two VCF files with the two sets of variants, e.g. one set of rare and one set of common variants, which makes it a flexible and scalable method.

The second statistical method is called RareComb (Pittman *et al.* 2022). It combines the Apriori algorithm (Agrawal *et al.* 1996) and statistical modelling to detect combinations of genes associated with a phenotype. The Apriori algorithm enumerates, in cases and controls independently, combinations of variants that meet a frequency threshold and that are seen simultaneously in one individual. Then, a binomial test is used to compare the observed frequency of each combination against the frequency expected under the hypothesis of independence, in cases and controls independently. Finally, RareComb identifies combinations that are enriched in cases and not in controls. The method was tested on patients and controls from the Simons Foundation Powering Autism Research (SPARK) cohort. RareComb identifies 148 pairs of genes with a significant enrichment in cases and not in controls, with adequate statistical power and moderate to high effect sizes. Unlike the two previous methods, RareComb is available as a R package and takes as input a sparse Boolean matrix with rare variant information for all individuals.

Overall, these network-based and statistical methods can be applied in very specific scenarios to detect digenism and make a number of assumptions on the type of variants kept for the analysis (prefiltering steps in the DM and RareComb) or the link between the two impacted genes (direct connection in STRING for OligoPVP).

Machine Learning methods to detect digenic inheritance

To approach the challenge of detecting DI in sequencing data from another angle, a number of ML algorithms have been developed. The characteristics of these ML methods are summarized in Table 8.2. Most of them use a Random Forest (RF) algorithm, with slight differences in the exact model used. VarCoPP and DIEP are both RF ensemble predictors, the first comprising of 500 RF classifiers of 100 trees each and the second of 26 top RF classifiers with the highest 10-fold cross-validation scores. DiGePred consists of a unique RF classifier of 500 trees. One of the latest ML method VarCoPP2.0 involves a Balanced RF comprising of 400 trees. The ML method ARBOCK differs from the other methods by the model used to predict digenic pairs and will be discussed separately at the end of this section, as it is less comparable to the other ML methods.

Table 8.2. Parameters of the ML methods to detect DI

Method	Model	Positive training set	Negative training set	Features
VarCoPP	RF ensemble predictor	DIDAv1	1000 Genomes Project	11 features
DiGePred	RF classifier	DIDAv1	Unaffected relatives Undiagnosed Diseases Network	12 features
DIEP	RF ensemble predictor	DIDAv1	Random pairs filtered using 1000 Genomes Project	17 features
VarCoPP2.0	Balanced RF	OLIDA	1000 Genomes Project	20 features
ARBOCK	Decision set	OLIDA	1000 Genomes Project	BOCK knowledge graph

A main characteristic of ML methods is their positive and negative training sets. VarCoPP, DiGePred and DIEP were all trained on the pairs from DIDAv1. DIDAv1 contains 200 digenic variant pairs, which in turns fall into 140 unique gene pairs, Because VarCoPP considers pairs of variants, it used the 200 DIDAv1 digenic variant pairs as a positive training set, while DiGePred and DIEP consider pairs of genes and thus used the 140 DIDAv1 digenic gene pairs as a positive training set. VarCoPP2.0 was the only RF method trained on OLIDA.

In OLIDA, each variant pair is given a score to indicate the strength of evidence of oligogenicity, which summarizes the genetic and functional evidence found in the literature to support the disease-causing character of the combination (FINALmeta score). VarCoPP2.0 was trained only on the 301 digenic pairs from OLIDA with significant enough evidence of oligogenicity (FINALmeta score ≥ 1) (Papadimitriou *et al.* 2023). Regarding the negative training set, VarCoPP and VarCoPP2.0 selected respectively 100,000 and 150,500 random variant pairs from the 1000 genomes project (Auton *et al.* 2015). DIEP also used the 1000 Genomes Project data for negative gene pairs. First, DIEP's developers selected 50,000 randomly constructed gene pairs from all possible genes (Random set) as well as 13,390 gene pairs obtained by randomly combining DIDA unique genes, excluding true DIDA pairs (DIDA_NI set). Then, these two negative gene sets were filtered by only keeping gene pairs which involved pairs of variants with the following criteria: rare (frequency ≤ 1%) and non-synonymous variant observed in 2 or more individuals of the 1000 Genomes Project. In total, DIEP had 8,400 pairs of genes in its negative training set, including 7,000 pairs from the Random set and 1,400 pairs from the DIDA_NI set. Following a different approach, DiGePred used as its final negative training set around 8,400 pairs from genes with variants in sequencing data from unaffected relatives of the Undiagnosed Diseases Network (UDN) individuals (Ramoni *et al.* 2017).

Finally, regarding the features used to train the ML algorithms, VarCoPP was trained on 11 features and VarcoPP 2.0 on 15 features at the variant, gene and gene-pair level. DIEP and DiGePred were both trained only on gene and gene-pair level features, with 12 and 17 total features respectively. A summary of all the features shared and specific to each method can be found in Supplementary Table S1. Interestingly, two features were included in all of the methods: the haploinsufficiency of each gene and the distance between the two genes in a protein-protein interaction (PPI) network. Another feature was present in all methods except VarCoPP: the coexpression of genes. VarCoPP and VarCoPP2.0 also shared a variant level features: the CADD scores of the 4 alleles of the pair. At the variant level, VarCoPP included in addition to CADD the flexibility and hydrophobicity of each variant.

The last feature included in VarCoPP was the recessiveness of each gene, which was also taken into account by DIEP. DIEP and DiGePred shared some features: the intolerance to loss-of-function variants of each gene, and the phenotype similarity between the genes. DiGePred shared one more feature with VarCoPP2.0: the selection pressure of each gene. The rest of all the other ML methods training features were method-specific.

In the description of each ML method, the Gini importance of each feature, which represents its contribution to the prediction (positive or negative) was computed. For VarCoPP, the feature importance was by far the greatest for the CADD score of the first variant allele of each gene, followed by the recessiveness of each of the genes. A similar pattern was observed in VarCoPP2.0 for CADD, but the second most important feature group was gene pair features (PPI distance, coexpression and knowledge graph distance). Gene pair level features also played an important role in prediction for DiGePred and DIEP. In DiGePred, the phenotype similarity had by far highest importance, followed by pathway similarity, number of phenotypes and PPI distance. For DIEP, the feature with the most importance was by far the PPI distance, followed by protein functional interaction effects, biological distance, semantic similarity of disease ontology (phenotype similarity) and gene GO annotations.

The last method to detect DI that falls into the machine learning category takes a different approach from the RF-based methods. The method uses an association rule based on a knowledge graph, which makes inferences by leveraging the structured information stored within a knowledge graph. This method is called ARBOCK for Association Rule learning Based on Overlapping Connections in Knowledge graphs. ARBOCK is based on the BOCK heterogeneous knowledge graph, which comprises 158,964 nodes of 10 different types and 2,659,064 edges of 17 different types. ARBOCK uses a two-step approach to predict pathogenic gene interactions. First, it generates a rule set from local patterns of the known pathogenic gene pairs in the knowledge graph. Then, these rules are combined into a decision set classifier to allow the detection of new pathogenic gene pairs. For training, ARBOCK used 441 gene pairs from the latest version of OLIDA with sufficient evidence of pathogenicity, and 44,100 putatively neutral gene pairs from the 1000 Genomes cohort.

ARBOCK cannot be compared as directly to the other ML methods based on RF algorithms but it can still be noted that in ARBOCK, high-confidence rules are characterized by metapaths (sequence of node and edge types) related to similar phenotypes (50%), biological processes (31.2%), and molecular functions (12.5%) which indicates key information taken into account by the classifier.

Overall, the machine learning methods to detect DI have a minority of strictly common features, although some of them are shared by many of the methods. They all take into account information at the gene and gene pair level. The gene pair level features seem to matter greatly in most methods for the prediction, in addition to variant level features when they are present.

Benchmark of existing machine learning algorithms to detect digenic inheritance

Overview of the benchmark

ML methods do not need as input a cohort of patients or pedigrees to make predictions, and are thus suited for the detection of DI in the context of very rare and heterogeneous diseases. All of these methods output a score between 0 and 1, and provide a threshold above which the pair is classified as pathogenic. The recommended score threshold for each method was used to categorize pairs as digenic: ≥0.496 for DiGePred, ≥0.5 for DIEP, ≥0.891 for VarCoPP2.0's 99.9%-confidence zone, ≥0.788 for ARBOCK.excl.pheno and ≥0.929 for ARBOCK.incl.pheno.

For a comprehensive benchmark of existing methods to find DI in sequencing data, we focused on the ML methods currently published and for which a program or implementation is available. The initial VarCoPP method was not included in this benchmark as it is no longer available on the ORVAL platform and was replaced by VarCoPP2.0. The specific threshold and software parameters used for each method can be found in the following sections.

Specifically for ARBOCK, we make a distinction between two possible predictive models: a model excluding phenotypic information (ARBOCK.excl.pheno strategy) and a model including phenotypic information (ARBOCK.incl.pheno strategy). All of the methods except VarCoPP2.0 take as input pairs of genes. VarCoPP2.0 is run through the web platform ORVAL and takes as input a list of variants, creating all possible pairs from the provided list.

The aim of this benchmark is to test the classifying performance of the aforementioned methods under different scenarios, leveraging pairs for which the classification as pathogenic or neutral is known or very likely. We were not able to screen a full individual exome in order to test all possible gene pairs which would be the most agnostic approach, considering the ORVAL platform especially limits the number of input variants. We thus devised three scenarios that address a range of potential combination occurrences from true datasets. The full breakdown of all scenarios and the pairs for evaluation can be found in Supplementary Table S2.

Collection of digenic gene pairs for benchmarking

The gene pairs used in the benchmark as known digenic pairs were taken from the OLIDA database V3, which contains data published up to February 2023. We retrieved a total of 1,610 oligogenic variant combinations in 1,075 unique genes from the OLIDA website (https://olida.ibsquare.be/). Among these oligogenic variant combinations, 1,076 were digenic pairs and kept for benchmarking. VarCoPP2.0 was the only ML method trained with OLIDA. We thus excluded from the set of digenic pairs the 301 pairs of genes used to train VarCoPP2.0, which were downloaded from their GitHub page (https://github.com/oligogenic/VarCoPP2.0). Among the digenic pairs not used to train VarCoPP2.0, 652 had a FINALmeta confidence score of 0 and were discarded. This left 157 digenic pairs after quality filtering. Finally, we also removed the digenic pairs involving CNVs as they could not be analyzed by the ORVAL platform. This left a total of 106 digenic pairs that were included in the benchmark (OLIDA_total pairs), among which 69 could be scored by all ML methods.

Another set of gene pairs was used specifically to evaluate the performance of the ARBOCK method. Indeed, in the article describing the ARBOCK method (Renaux *et al.* 2023), a set of 15 pathogenic gene pairs that were held-out as independent test set was provided (ARBOCK_held_out pairs), among which 14 would be scored by all ML methods. Whereas some of the 106 previously described digenic pairs could have been selected in the method, this set was explicitly not used to train ARBOCK and could thus be used as a test set for the method.

Collection of non-digenic gene pairs for benchmarking

To collect plausible non-digenic gene pairs for benchmarking, we selected ten random individuals from the French general population among the 574 exomes of the FREX (Génin *et al.* 2017) project. An individual-focused QC was applied on the exome data using the RAVAQ R package (Marenne *et al.* 2022) : we performed a genotype and variant QC with default parameters corresponding to standard GATK (McKenna *et al.* 2010) hard filtering criteria: genotypes with a depth < 10 or a genotype quality < 20 were set to missing, as well as the heterozygous genotypes with an allele balance outside [0.25-0.75]. No QC on variant call rates was applied. This corresponds to the default RAVAQ parameters for genotype and variant QC except that: MAX_AB_GENO_DEV = 0.25, MAX_ABHET_DEV, MIN_CALLRATE and MIN_FISHER_CALLRATE "disabled".

The ORVAL platform specifies that a full exome cannot be analyzed with the VarCoPP2.0 method. In order to reduce the number of variants and thus gene pairs to be analyzed, we first kept the heterozygote and/or homozygote variant (SNVs or Indels) with the maximal CADD v1.6 (Rentzsch *et al.* 2021) score for each gene, which left around 11,000 variants for each FREX individual. Then we applied our in-house variant prioritization method Easy-PSAP (https://github.com/msogloblinsky/Easy-PSAP) to score variants with their PSAP p-values (null distribution “latest_gnomadgen_string_ensembl_cadd1.6_af_nosing_lookup_*” in hg19 on genes).

We applied two strategies to select the neutral variants: for each individual, either selecting 100 random variants from the 11,029 variants (random100_FREX pairs) or selecting the top 100 variants with the lowest PSAP p-value, corresponding to the variants with the highest predicted pathogenicity (top100PSAP_FREX pairs). For each of the ten individuals, all possible unique gene pairs involving two different genes associated with these variants were then used for benchmarking. We also shuffled the genes implicated in the 14 gene pairs held out from ARBOCK to test as negative control all possible pairs of these genes except the initial true digenic pairs (ARBOCK_shuffled pairs).

Finally, we retrieved a curated list of well-reviewed pathogenic variants (Ogloblinsky *et al.* 2024) from the ClinVar (Landrum *et al.* 2018) database with an associated autosomal dominant (AD) or recessive (AR) mode of inheritance. We randomly selected 100 variants in distinct genes for each mode of inheritance and paired them with a variant either from the random100_FREX or the top100PSAP_FREX variant lists. This allowed us to evaluate the performance of methods aimed at detecting DI on pairs involving a known monogenic variant and a neutral variant with four different scenarios to construct the pairs (random100_FREX_clinvar100_AD, top100PSAP_FREX_clinvar100_AD, random100_FREX_clinvar100_AR, top100PSAP_FREX_clinvar100_AR).

Software parameters and evaluation

DiGePred (Mukherjee *et al.* 2021) and DIEP (Yuan *et al.* 2022) scores were downloaded for all possible pairs of human genes. DiGePred scores were split in four files for which the download links were provided on their GitHub (https://github.com/CapraLab/DiGePred). We concatenated the four files in one file, and used the score “unaffected_no_gene_overlap_score” as advised in the DiGePred article. For DIEP scores, the full table for all gene pairs was downloaded from https://github.com/pmglab/DIEP.

The ORVAL (Renaux *et al.* 2019) platform was used to run VarCoPP2.0 (Versbraegen *et al.* 2023). All filtering options were removed: no minimum threshold on the MAF was applied, and no intergenic, intronic and synonymous variants were removed. No gene filtering was applied. The VarCoPP2.0 model in hg19 was used to match the build of the variants' coordinates. If multiple heterozygote variants were present in the same gene for different variant pairs, they were split in different ORVAL submissions as they would be considered compound heterozygote variants by ORVAL otherwise. ORVAL constructs all possible gene pairs and outputs their VarCoPP2.0 score, which we filtered to keep only the pairs were truly featured in our digenic and non-digenic gene pair set.

To run the ARBOCK method (Renaux *et al.* 2023), the BOCK knowledge graph was first downloaded from https://doi.org/10.5281/zenodo.7185679. Then, ARBOCK was installed following the specifications on the GitHub of the method (https://github.com/oligogenic/ARBOCK). The parameters by default were used and two "predict" models were tested: "ds_model_with_pheno" (ARBOCK.excl.pheno strategy) and "ds_model_no_pheno" (ARBOCK.incl.pheno strategy).

Despite the previously cited methods described as being able to analyze any gene pair, it was not actually the case in practice. Thus, the methods were compared only on pairs which could be scored by all methods, as for instance the pre-computed files with DiGePred or DIEP scores did not feature all possible gene pairs The performances of the methods were evaluated through different scenarios using different gene pair sets that are detailed in Supplementary Table S2.

Classification of digenic gene pairs held-out from ARBOCK method

The latest published method to detect DI being ARBOCK, we first tested the 14 ARBOCK_held_out pairs as well as the 364 ARBOCK_shuffled pairs. In the article describing ARBOCK, the performance of DiGePred had also been tested on these ARBOCK_held_out pairs. Overall, DiGePred and VarCoPP2.0 had the worst performance in terms of classifying the ARBOCK_held_out pairs as pathogenic (4 for DiGePred and 2 for VarCoPP2.0, Figure 8.1, Supplementary Table S3). DIEP accurately classifies more than half of the true digenic pairs.

The best results on ARBOCK_held_out pairs were achieved by ARBOCK: ARBOCK.excl.pheno and ARBOCK.incl.pheno classify 11 and 9 pairs as pathogenic, respectively. However, when looking at the negative testing set comprising of 364 ARBOCK_shuffled pairs, ARBOCK exhibits an extremely strong misclassification rate. ARBOCK.excl.pheno and ARBOCK.incl.pheno classify 236 (65%) and 206 (57%) of the ARBOCK_shuffled pairs as pathogenic, respectively. In contrast, DiGePred has the lowest rate of misclassification of the negative pairs (2%), followed by DIEP (12%) and VarCoPP2.0 (16%).

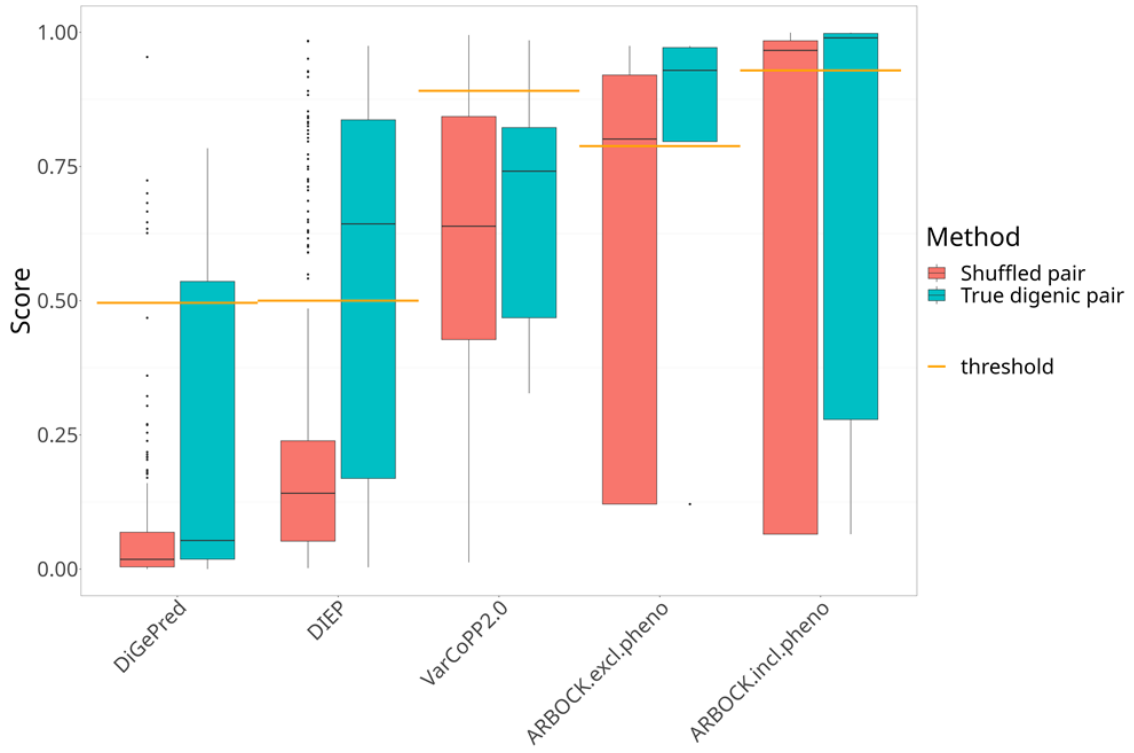


Figure 8.1. Distribution of ML methods scores for the pathogenic and shuffled pairs held out from ARBOCK training

Classification of digenic gene pairs from OLIDA compared to neutral pairs from FREX

We then extended our benchmark to a larger number of pairs and calculated evaluation metrics for the classification performance of each method. For the positive pair test set, we focused on the 69 digenic pairs from OLIDA that were not used to train VarCoPP2.0 with a confidence score above 0 and not involving CNVs that cannot be scored with VarCoPP2.0. For our neutral pairs test sets, we had 32,953 pairs in the random100_FREX and 18,348 pair in the top100PSAP_FREX set (see

Supplementary Table S2). The latter scenario corresponds to a possible clinical practice workflow that would involve shortlisting the variants with the highest predicted pathogenicity in an individual and searching for DI within that shortlist, which can challenge methods like VarCoPP2.0 that use variant level data. All possible unique combinations from these variants or their associated genes were tested for DI classification.

All methods classified around half or more of the OLIDA_total pairs as pathogenic (Figure 8.2, Supplementary Table S4), the best performance being achieved by ARBOCK.incl.pheno and then ARBOCK.excl.pheno with 64 and 60 well-classified pairs, respectively. The results were more contrasted between methods for the FREX neutral pairs, as observed for the ARBOCK_shuffled pairs. For the random100_FREX pairs, DiGePred and VarCoPP2.0 badly classified as pathogenic only 16 and 10 pairs, respectively. The number of random100_FREX pairs classified as pathogenic was higher for DIEP (828 out of 32,953 pairs) and ARBOCK.incl.pheno (2,525 out of 32,953 pairs) but still represented a fairly low percentage (2.5% and 7.6%, respectively). In contrast, ARBOCK.excl.pheno classified 10,056 random100_FREX pairs (30%) as pathogenic. This trend of classification is found again for the top100PSAP_FREX pairs, with DiGePred keeping the lowest number of badly classified pair. The only main difference was found for VarCoPP2.0 percentage of badly classified top100PSAP_FREX pairs which increased from 0.03% to 2.4%.

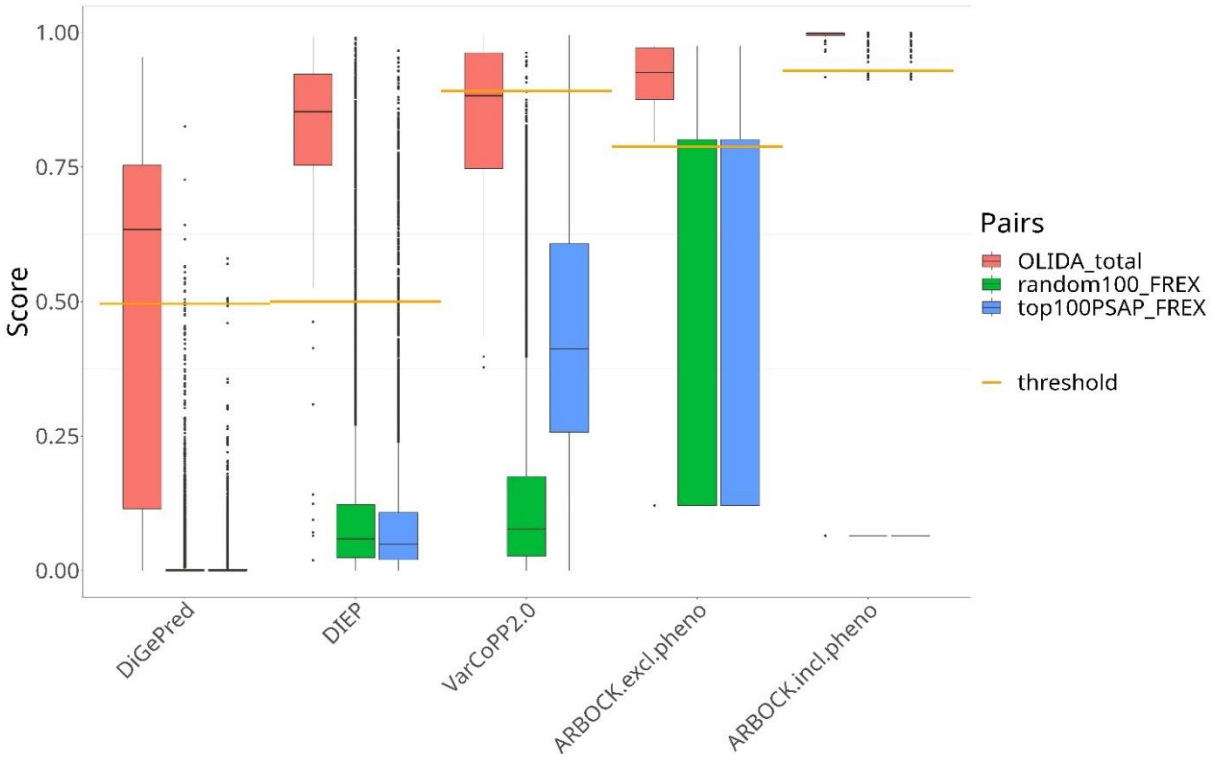


Figure 8.2. Distribution of ML methods scores for the OLIDA pairs and the two scenarios with FREX variant pairs

We calculated evaluation metrics for each ML method (Table 8.3), by using OLIDA_total pairs as true positives and either all of FREX pairs (all), only random100_FREX pairs (random) or only top100PSAP_FREX pairs (topPSAP) as true negatives. There was not a lot of differences in the evaluation metrics between the "all", "random" and "topPSAP" metrics so we focused on the "all" metrics for evaluation. All methods have a high predictive negative value but only DiGePred has a high predictive positive value (0.657). When also looked at the F1 score that measures the predictive performance of a classifier. The F1 score is the harmonic mean of precision, the ratio of true positive predictions to the total number of positive predictions made by the model, and recall, the ratio of true positive predictions to the total number of actual positive instances. The highest F1 score was achieved by far for DiGePred (0.647). All methods achieved a high specificity, except ARBOCK.excl.pheno which had a lower specificity of 0.695. Finally, sensibility of ARBOCK.excl.pheno, ARBOCK.incl.pheno and DIEP was high as well, and lower for DiGePred and VarCoPP2.0.

Table 8.3. Evaluation metrics for all benchmarked methods to detect DI

Method	Sensibility	Specificity	Positive predictive value	Negative predictive value	F1
ARBOCK.excl.pheno	0.87	0.695	0.004	1	0.008
ARBOCK.incl.pheno	0.928	0.929	0.017	1	0.034
DIEP	0.855	0.978	0.049	1	0.092
DiGePred	0.638	1	0.657	1	0.647
VarCoPP2.0	0.493	0.991	0.07	0.999	0.123

Classification of pathogenic ClinVar variant and neutral FREX variant pairs

A final aspect that we included in this benchmark is the ability of these methods to distinguish between DI and monogenic inheritance. To construct a test set for this scenario, we combined a curated list of well-review pathogenic ClinVar variants with a known autosomal dominant (AD) or autosomal recessive (AR) mode of inheritance (100 variants from different genes for the AD and AR models, respectively) with the previously described two sets of variants from FREX. This led to four scenarios involving a monogenic variant and neutral variant in pairs:

- random100_FREX_clinvar100_AD
- random100_FREX_clinvar100_AR
- top100PSAP_FREX_clinvar100_AD
- top100PSAP_FREX_clinvar100_AR.

In every scenario, DiGePred did not classify any of the pairs as pathogenic (Figure 8.3, Supplementary Tables S5 and S6). DIEP and VarCoPP2.0 gave almost similar results, with less than 10 pairs wrongly classified for any scenario.

VarCoPP2.0 performed slightly worse for the two sets using top100PSAP_FREX compared to random100_FREX, going from 2 to 6 and from 1 to 7 wrongly classified pairs for the AD and AR models, respectively. The worst classification performances were observed for ARBOCK.excl.pheno for the scenarios involving ClinVar AD variants, with 32 pairs (random100_FREX_clinvar100_AD) or 36 pairs (top100PSAP_FREX_clinvar100_AD) classified as pathogenic. ARBOCK.incl.pheno performed slightly better than ARBOCK.exl.pheno but worse than all of the other methods. The results of the two ARBOCK methods were better for the scenarios involving ClinVar AR variants, but underperformed compared to the other aforementioned methods.

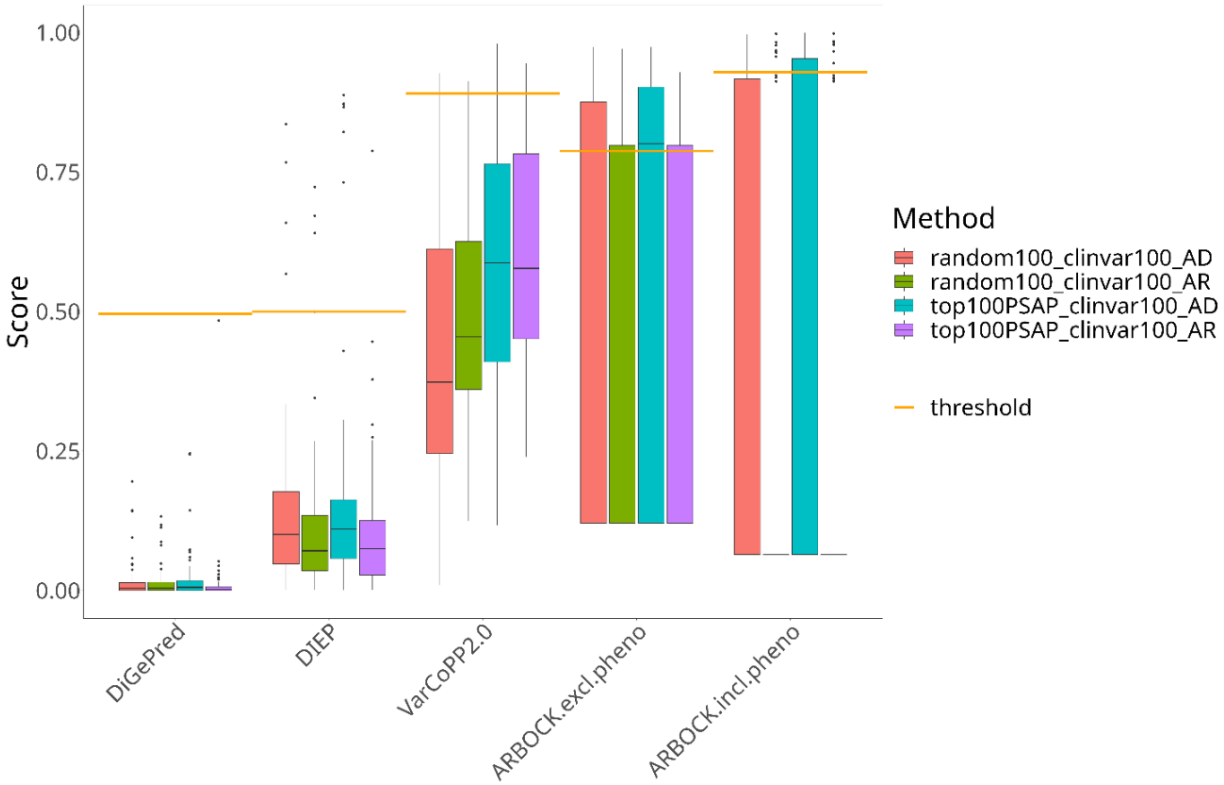


Figure 8.3 : Distribution of ML methods scores for the four scenarios involving ClinVar and FREX variant pairs

Discussion

Overall, several methods have been developed recently to tackle the complex issue of detecting DI in sequencing data, with different strategies. Our review highlights that network-based and statistical methods to detect DI are largely applicable to sequencing data but have several limitations. Regarding network-based methods, OligoPVP relies heavily on known biological interactions and phenotypic data and fails to predict most digenic pairs at high ranks as described in the article describing the method. Statistical methods need a substantial number of individuals and are not suited for individual-based analysis like in the case of many rare diseases. They also require a strong prefiltering step of variants to ensure adequate power. In contrary, ML methods can be applied at the individual scale. Although exhaustively evaluating all of the possible variant or gene pairs is still not feasible with that type of methods, they can in theory be applied to any pair or genes or variants. Although it is difficult to pinpoint which level of information contributes the most to accurate predictions of digenism, the performance of the benchmark ML and knowledge-graph based methods seems to indicate that gene-level and gene-pair level features play a key role in predicting DI.

A significant issue in the application of the ML methods to detect DI is their false positive rate. The method the most appropriate in that regard is DiGePred, which manages to control the number of false positives. A downside of DiGePred is that phenotype-related features play the most important role in DiGePred predictions (44% of feature importance), which can hinder the generalization of the model when these features are missing or incomplete for a gene. That is evident from the ARBOCK article and confirmed by our benchmark: DiGePred was only able to predict correctly 4 out of 14 ARBOCK_held_out pairs, against 11 pairs for ARBOCK.excl.pheno which was not trained using phenotype-based features. Therefore, DiGePred is a more conservative method that tend to fail to predict gene pairs as pathogenic if they lack complete phenotype annotation or have limited common phenotype terms. DIEP was described as a better predictor of DI than DiGePred, as it included a larger and more finely filtered negative training set and broader range of features.

Indeed, DIEP predicts more of the OLIDA_total (59 for DIEP against 44 for DiGePred) and ARBOCK_held_out pairs (8 for DIEP against 4 for DiGePred) as digenic. However, DIEP also produces a lot more false positive predictions across all the negative testing sets. In contrast to DiGePred, the most important features in DIEP are protein–protein association, functional interaction, and semantic similarity of gene DO and GO annotations. The two ARBOCK predictive models have both the highest sensitivity and lowest specificity. This behavior of the method had already been described in the article by [Renaux et al., 2023], which states that the ARBOCK pathogenicity predictor excluding phenotypic information, exhibits a relatively high false positive rate. In our benchmark, even when ARBOCK included phenotypic information, the number of false positives was still very high compared to the other methods, which can partly be explained by the uncertainty surrounding the selection of neutral gene pairs to train the method. It can also be noted than only DiGePred did not misclassify any of the pathogenic ClinVar variant and neutral FREX variant pairs

One observation already highlighted in the DiGePred article when comparing DiGePred to the initial VarCoPP model is that VarCoPP's prediction were strongly-affected by variant-level features, especially if one of the gene carries a variant predicted as highly pathogenic, which may hide or confound purely digenic effects. This behavior could also be seen in our benchmark, as VarCoPP2.0 wrongly classified as pathogenic only 10 pairs of the randomly selected neutral random100_FREX pairs and 440 pairs of the neutral top100PSAP_FREX pairs that include for each individual the variants predicted as the most pathogenic. The most important feature of VarCoPP2.0 being the CADD scores of the variants, this observation was expected. Still, the VarCoPP2.0 model was enriched in gene-pair level features compared to VarCoPP, which improved the capture of the gene-gene synergistic relationship inside a variant pair that is crucial for DI prediction. In the article describing VarCoPP2.0, the authors highlight that the method has a high sensitivity, which is important in a clinical setting, as well as a 5% false positive rate, and should be used in conjunction with variant filtering and/or restriction to a panel of genes. The currently advised variant filtering for VarCoPP2.0 in ORVAL is to remove variants with a MAF ≥ 0.035, intergenic, intron and synonymous variants.

We did not select variants according to these criteria as it would have drastically reduced our negative and positive testing sets (160 pairs in the random100_FREX set and 2,012 pairs in the top100PSAP_FREX set), but following this approach could have improved VarCoPP predictions.

A main limitation of our benchmark, and of all of the aforementioned benchmarked methods, is the small number of known digenic pairs both used as positive training and testing sets. The OLIDA database is updated regularly, and the performance of the methods to detect DI will have to be tested on the new pairs that are described in the literature. Our positive testing set can also be biased against specific disease and genes that are overly represented in OLIDA due to their known digenic or oligogenic nature. Regarding the applicability of the benchmarked methods, both DiGePred and DIEP provide pre-calculated tables of prediction for all human gene pairs. DIEP even presents the advantage of providing a high-efficiency java package to search for specific digenic pairs or genes. These two methods, as well as ARBOCK, present the advantage of separating the selection and effect of variants from the identification of potentially digenic gene pairs. Indeed, the question of variant impact has been extensively explored by different methods, whereas reflecting the potential impact of a simultaneous alteration of two genes is very specific to digenism. VarCoPP2.0 is not suitable for whole exome analysis, especially due to the challenge of running the web server on a large scale.

Methods continue to be published on the subject of DI detection regularly, and leverage either machine-learning (Paoli *et al.* 2023) or statistical approaches (Pittman *et al.* 2022; Zhang *et al.* 2023). The most recent method to detect DI that was published during the curation of this review is called the High-throughput oligogenic prioritizer (Hop) (Gravel *et al.* 2024) and leverages both the VarCoPP2.0 predictor and the BOCK knowledge graph. The input of the method is a VCF file, along with either a combination of HPO terms describing the patient's phenotype or a panel of genes known to be involved in the disease or both. The VCF file is filtered to remove variants with a MAF ≥ 0.035, synonymous variants that are further than 195 bp from exon edges and intronic variants. Then, Hop calculates two scores. The first score is calculated using VarCoPP2.0, which evaluates the pathogenicity of all possible variant combinations from the filtered VCF file.

The second score is a disease-relevance score for all gene pairs that is calculated using a random-walk with restart algorithm with the provided disease-related terms as seeds on the BOCK heterogeneous knowledge graph. These two scores are then averaged to provide a final score for ranking all of the variant pairs of the individual. The performance of Hop was tested on synthetic disease exomes created by inserting known pathogenic pairs from OLIDA, either used to train VarCoPP2.0 or not, in exomes from the 1000 Genomes Project and the UK10K project (UK10K Consortium *et al.* 2015). Hop was confronted to popular monogenic variant prioritization tools CADD (Rentzsch *et al.* 2021) and Exomiser (Smedley *et al.* 2015) as well as oligogenic predictor OligoPVP which uses the same type of input. Hop outperformed all of the other methods on the synthetic disease exomes and managed to rank around 70% of pathogenic gene pairs in the top 50 of pairs, against around 20% for Exomiser and OligoPVP and less than 10% for CADD. Despite promising results on the synthetic disease exomes, Hop is still limited by its reliance on prior knowledge, especially regarding the link between HPO terms and disease genes in the knowledge graph. The Hop method also works better if a relevant gene panel for the disease is provided, which limits the use of the method towards well-studied disease.

These methods are not yet at a stage allowing an agnostic exploration of real-world cases of undiagnosed diseases, but can serve as valuable information combined with other lines of evidence like familial data. They also lack interpretability in their predictions. ORVAL and ARBOCK tried to mitigate the issue of interpretability by offering more context at the gene and gene pair level such as visual mappings of the cellular location and pathway information and explanation in the form of subgraphs, respectively. Overall, the appropriate integration of variant selection and gene pair prediction will in the future lead to more accurate prediction of DI, and facilitate the detection of this complex mode of inheritance.

References

Agrawal R, Mannila H, Srikant R *et al.* Fast discovery of association rules. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, et al. (eds.). *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press, 1996, 307–28.

Amberger JS, Bocchini CA, Schiettecatte F *et al.* OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* 2015;**43**:D789–98.

An U, Pazokitoroudi A, Alvarez M *et al. Deep Learning-Based Phenotype Imputation on Population-Scale Biobank Data Increases Genetic Discoveries*. Bioinformatics, 2022.

Auton A, Abecasis GR, Altshuler DM *et al.* A global reference for human genetic variation. *Nature* 2015;**526**:68–74.

Beales PL, Badano JL, Ross AJ *et al.* Genetic Interaction of BBS1 Mutations with Alleles at Other BBS Loci Can Result in Non-Mendelian Bardet-Biedl Syndrome. *The American Journal of Human Genetics* 2003;**72**:1187–99.

Boudellioua I, Kulmanov M, Schofield PN *et al.* OligoPVP: Phenotype-driven analysis of individual genomic information to prioritize oligogenic disease variants. *Sci Rep* 2018;**8**:14681.

Boycott KM, Hartley T, Biesecker LG *et al.* A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers. *Cell* 2019;**177**:32–7.

Boycott KM, Rath A, Chong JX *et al.* International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *The American Journal of Human Genetics* 2017;**100**:695–705.

Cerrone M, Remme CA, Tadros R *et al.* Beyond the One Gene–One Disease Paradigm. *Circulation* 2019;**140**:595–610.

Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* 2002;**11**:2463–8.

Fauser S, Munz M, Besch D. Further support for digenic inheritance in Bardet-Biedl syndrome. *Journal of Medical Genetics* 2003;**40**:e104–e104.

Gazzo AM, Daneels D, Cilia E *et al.* DIDA: A curated and annotated digenic diseases database. *Nucleic Acids Research* 2016;**44**:D900–7.

Génin E, Feingold J, Clerget-Darpoux F. Identifying modifier genes of monogenic disease: strategies and difficulties. *Hum Genet* 2008;**124**:357–68.

Génin E, Redon R, Deleuze J *et al.* The French Exome (FREX) Project: A Population-based panel of exomes to help filter out common local variants. *Genetic Epidemiology* 2017;**41**:691–691.

Gravel B, Renaux A, Papadimitriou S *et al.* Prioritization of oligogenic variant combinations in whole exomes. *Bioinformatics* 2024:btae184.

Jiang H, Liang S, He K *et al.* Exome sequencing analysis identifies frequent oligogenic involvement and FLNB variants in adolescent idiopathic scoliosis. *J Med Genet* 2020;**57**:405–13.

Kajiwara K, Berson EL, Dryja TP. Digenic Retinitis Pigmentosa Due to Mutations at the Unlinked Peripherin/RDS and ROM1 Loci. *Science* 1994;**264**:1604–8.

Katsanis N, Ansley SJ, Badano JL *et al.* Triallelic Inheritance in Bardet-Biedl Syndrome, a Mendelian Recessive Disorder. *Science* 2001;**293**:2256–9.

Katsanis N, Eichers ER, Ansley SJ *et al.* BBS4 Is a Minor Contributor to Bardet-Biedl Syndrome and May Also Participate in Triallelic Inheritance. *The American Journal of Human Genetics* 2002;**71**:22–9.

Kerner G, Bouaziz M, Cobat A *et al.* A genome-wide case-only test for the detection of digenic inheritance in human exomes. *PNAS* 2020;**117**:19367–75.

Kim A, Savary C, Dubourg C *et al.* Integrated clinical and omics approach to rare diseases: novel genes and oligogenic inheritance in holoprosencephaly. *Brain* 2019;**142**:35–49.

Kircher M, Witten DM, Jain P *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**:310–5.

Kousi M, Katsanis N. Genetic Modifiers and Oligogenic Inheritance. *Cold Spring Harb Perspect Med* 2015;**5**:a017145.

Landrum MJ, Lee JM, Benson M *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;**46**:D1062–7.

Lupski JR. Digenic inheritance and Mendelian disease. *Nat Genet* 2012;**44**:1291–2.

Marenne G, Ludwig TE, Bocher O *et al.* RAVAQ: An integrative pipeline from quality control to region-based rare variant association analysis. *Genetic Epidemiology* 2022;**46**:256–65.

McKenna A, Hanna M, Banks E *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303.

Mukherjee S, Cogan JD, Newman JH *et al.* Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network. *The American Journal of Human Genetics* 2021, DOI: 10.1016/j.ajhg.2021.08.010.

Nachtegael C, Gravel B, Dillen A *et al.* Scaling up oligogenic diseases research with OLIDA: the Oligogenic Diseases Database. *Database* 2022;**2022**:baac023.

Ogloblinsky M-SC, Bocher O, Aloui C *et al.* PSAP-genomic-regions: a method leveraging population data to prioritize coding and non-coding variants in whole genome sequencing for rare disease diagnosis. 2024:2024.02.13.580050.

Okazaki A, Ott J. Machine learning approaches to explore digenic inheritance. *Trends in Genetics* 2022;**38**:1013–8.

Paoli FD, Nicora G, Berardelli S *et al.* Digenic variant interpretation with hypothesis-driven explainable AI. 2023:2023.10.02.560464.

Papadimitriou S, Gazzo A, Versbraegen N *et al.* Predicting disease-causing variant combinations. *PNAS* 2019;**116**:11878–87.

Papadimitriou S, Gravel B, Nachtegael C *et al.* Toward reporting standards for the pathogenicity of variant combinations involved in multilocus/oligogenic diseases. *Human Genetics and Genomics Advances* 2023;**4**:100165.

Pittman M, Lee K, Srivastava D *et al.* An oligogenic inheritance test detects risk genes and their interactions in congenital heart defects and developmental comorbidities. 2022:2022.04.08.487704.

Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015;**31**:761–3.

Rahit KMTH, Tarailo-Graovac M. Genetic Modifiers and Rare Mendelian Disease. *Genes* 2020;**11**:239.

Ramoni RB, Mulvihill JJ, Adams DR *et al.* The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. *Am J Hum Genet* 2017;**100**:185–92.

Renaux A, Papadimitriou S, Versbraegen N *et al.* ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Research* 2019;**47**:W93–8.

Renaux A, Terwagne C, Cochez M *et al.* A knowledge graph approach to predict and interpret disease-causing gene interactions. *BMC Bioinformatics* 2023;**24**:324.

Rentzsch P, Schubach M, Shendure J *et al.* CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Medicine* 2021;**13**:31.

Richards S, Aziz N, Bale S *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;**17**:405–24.

Ritchie GRS, Dunham I, Zeggini E *et al.* Functional annotation of noncoding sequence variants. *Nat Methods* 2014;**11**:294–6.

Robinson PN, Köhler S, Bauer S *et al.* The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *Am J Hum Genet* 2008;**83**:610–5.

Schäffer AA. Digenic inheritance in medical genetics. *J Med Genet* 2013;**50**:641–52.

Smedley D, Jacobsen JOB, Jager M *et al.* Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc* 2015;**10**:2004–15.

Smedley D, Schubach M, Jacobsen JOB *et al.* A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *Am J Hum Genet* 2016;**99**:595–606.

Smith CL, Goldsmith C-AW, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol* 2005;**6**:R7.

Szklarczyk D, Morris JH, Cook H *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research* 2017;**45**:D362–8.

Teles e Silva AL, Glaser T, Griesi-Oliveira K *et al.* Rare CACNA1H and RELN variants interact through mTORC1 pathway in oligogenic autism spectrum disorder. *Transl Psychiatry* 2022;**12**:1–11.

UK10K Consortium, Walter K, Min JL *et al.* The UK10K project identifies rare variants in health and disease. *Nature* 2015;**526**:82–90.

Versbraegen N, Gravel B, Nachtegael C *et al.* Faster and more accurate pathogenic combination predictions with VarCoPP2.0. *BMC Bioinformatics* 2023;**24**:179.

Yuan Y, Zhang L, Long Q *et al.* An accurate prediction model of digenic interaction for estimating pathogenic gene pairs of human diseases. *Comput Struct Biotechnol J* 2022;**20**:3639–52.

Zhang Q, Bhatia M, Park T *et al.* A multi-threaded approach to genotype pattern mining for detecting digenic disease genes. *Front Genet* 2023;**14**, DOI: 10.3389/fgene.2023.1222517.

8.3 Discussion – Perspectives for the detection of digenism in rare disease cases

As evidenced by our review of the methods to detect DI in sequencing data, there is no current gold standard method to achieve this aim and digenism remains a difficult mechanism to detect, especially in a clinical setting. However, ML methods and in particular DiGePred have emerged as promising methods to evaluate the potential pathogenicity of a combined alteration of a pair of genes. In this section, we will briefly mention some future development that have stemmed from our aforementioned work.

First, we have seen that evaluating all pairs of genes of an individual is not yet feasible due to the computational explosion that would ensue. A straightforward way to circumvent this issue would be to first filter variants for an individual. We do not recommend filtering on the frequency of the variant, as the frequency cutoff for observing the variant in the general population might not be as low as for a monogenic variant. However, the variant still needs to be deleterious for the gene to lead to digenism. We thus advise shortlisting variant depending on their PSAP p-value (top 100 of variants for instance), which is still tied to the frequency of predicted pathogenic variant in the general population in this gene, or keep variants with a high predicted impact (VEP classification). The latter criteria, when applied to the cohort of 235 CSVD patient from (Aloui *et al.* 2021) leads to 71 to 133 variants by individual and partially overlaps with the PSAP top 100. The number of combinations of genes with these variants is drastically lowered compared to all possible combinations of genes for an individual, although it is still extensive work to score them with the ML methods.

Another avenue we have explored is the integration of PSAP p-values in place of CADD scores for the training of the VarCoPP2.0 method, considering PSAP consistently outperformed CADD in our simulations. However, we were not able to compute PSAP p-values for all variant combinations used to train VarCoPP2.0 as some of the variants were in non-coding parts of genes that were not included in PSAP-genes null distributions or were located in a gene not well-covered in the gnomAD v3 database. When tested by the Lenaerts' team which developed VarCoPP2.0, this resulted in a slightly decreased performance of VarCoPP2.0 with PSAP p-values instead of CADD. Variants for which there was a missing value were imputed using the median of the value for that feature in the two sets and thus did not provide any information to the model.

Finally, another way to adapt the ML scores to detect DI would be to implement the idea of PSAP but for pairs of genes: instead of looking at variants with the CADD score within a gene, look at pairs of genes with the same ML score. This idea was prompted by the observation that, in the FREX database, pairs of genes that had a high ML score (here DIEP and DiGePred) were less likely to both carry a deleterious variant with a high CADD score in many of the FREX individuals (Figure 8.4), when

looking at a random sample of 10,000 pairs of genes. It appears that pairs of genes with a high ML score are less tolerant to a combined alteration of both genes in the general population. The next step would be to find a way to collapse PSAP null distributions for genes with the same ML score and evaluate the probability of seeing such deleterious variants in such a connected pair of genes, in the general population.

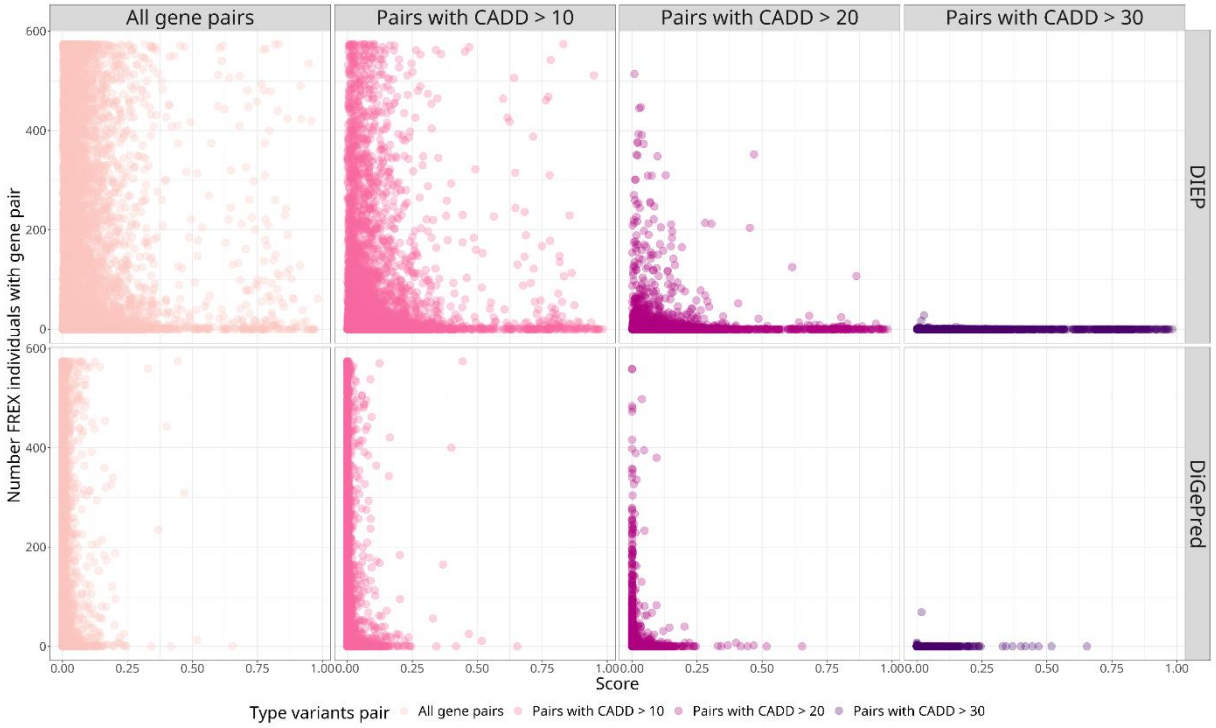


Figure 8.4 : Number of FREX individuals carrying a pair of genes both with variants meeting a CADD threshold depending on their DIEP or DiGePred score

(CADD threshold = 10, 20 or 30 from left to right)

Part V

Discussion

In this discussion, I outline the main results from my work and their impact in the field of RD diagnosis. Then, I touch on some additional subjects that have not been mentioned in previous sections. I show how our implementation of Easy-PSAP integrated within the broad issue of reproducibility of research results. Finally, I highlight some perspectives of this work which revolve around the challenge of variant interpretation in the non-coding genome and the extension of PSAP to integrate more complex models of inheritance and multi-omics data. I also discuss the application of emerging technologies like AI in the field of genomics, the access and use of genomic data and most importantly, the impact of methodological developments in RD diagnosis for patient's lives.

Chapter 9 CONTRIBUTION AND PERSPECTIVES FOR THE EASY-PSAP STRATEGIES

9.1 MAIN RESULTS

In summary, the main developments and results from my PhD thesis are the following:

i. The PSAP-genomic-regions method was developed, extending the PSAP method to prioritize coding and non-coding variants in whole-genome data from a single individual affected by a RD.

ii. On artificially-simulated disease genomes and exomes, the method was able to prioritize known pathogenic variants better than using a pathogenicity score alone. Coding pathogenic variants and non-coding splicing variants were the types of variants best prioritized by the method.

iii. On real exome or genome data from patients with known CSVD and male infertility variants respectively, PSAP-genomic-regions systematically prioritized the known variant within the top 100 variants of the exome or genome.

iv. The application of PSAP-genomic-regions combined to a disease-relevant post-filtering strategy to families affected by male infertility resulted in the proposition of candidate gene variants in some of the families, in known and novel genes.

v. PSAP-genomic-regions and the initial PSAP-gene strategy were implemented within the Easy-PSAP pipeline, which comprised of two user-friendly and flexible workflows: one to calculate custom PSAP null distributions and one to apply the PSAP strategies to patient data.

vi. A review of the current methods to detect digenic inheritance showed the promises and limitations of ML methods especially, which could be integrated with PSAP to detect digenism in RD cases.

9.2 CONTRIBUTION TO SUSTAINABLE DATA ANALYSIS

With Easy-PSAP, we made available the exact scripts, parameters and datasets we used to create all of the PSAP null distributions used in the article presenting PSAP-genomic-regions. This is a step into the field of Open Science (OS), especially regarding Open Data and Open Software. As researchers, we have the responsibility to make our research accessible and reproducible, which is one of the main goals of OS. There have been recent warnings by researchers themselves about a "reproducibility crisis" in science. A recent review on the subject pointed out that almost 80% of researchers in biology had failed to reproduce someone else's experiments and 60% failed to reproduce their own experiments (Baker 2016). Several propositions were made by (Munafò *et al.* 2017) to combat the issue of reproducibility in science and improve the scientific process, around the themes of methodologies, reporting, reproducibility of results, evaluation and incentives for OS.

In the field of computer sciences, analyses imply the processing of large datasets, often many times, with modifications to the methods, parameters and sometimes even the data itself, which can make the description of methods complex and difficult to follow (Mesirov 2010). We first addressed the issue of reproducibility by integrating Easy-PSAP within conda environments (https://conda.io). Conda is a reproducible environment that contains a specific collection of packages necessary to run the pipeline. In addition, Easy-PSAP is written as Snakemake workflows (Mölder *et al.* 2021), which allow the adaptability and transparency of our results and thus their sustainability as illustrated by Figure 9.1. Sharing our codes with the scientific community also becomes an opportunity for the code to be improved upon and adapted for other research projects.

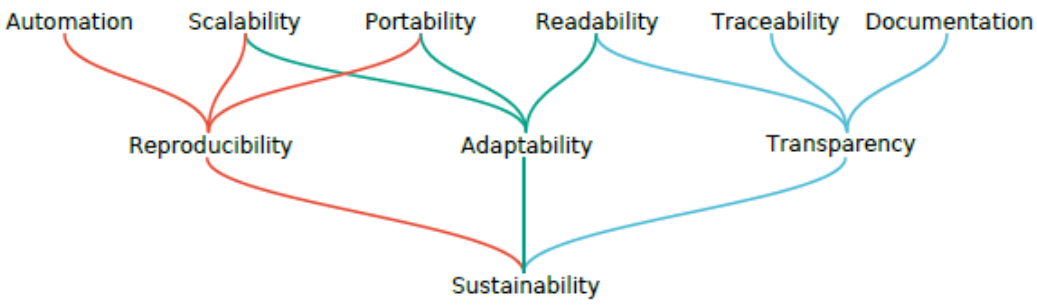


Figure 9.1 : Hierarchy of aspects to consider for sustainable data analysis

From Mölder et al. 2021: by supporting the top layer, a workflow management system can promote the center layer, and thereby help to obtain true sustainability

9.3 CHALLENGES OF VARIANT PRIORITIZATION IN THE NON-CODING GENOME

We have extensively motivated in this thesis the importance of analyzing the non-coding genome, especially in the case of RDs, and made available our PSAP-genomic-regions strategy to do so through the Easy-PSAP pipeline. The exploration of the Malakand phase 1 and phase 2 datasets underscored limitations in the applicability of PSAP-genomic-regions to detect candidate variants in the non-coding genome, especially around the interpretation of potentially pathogenic variants in the non-coding genome. First, the non-coding pathogenic variants from ClinVar were less well prioritized with PSAP-genomic-regions-ACS than the coding ones in our simulations. We still chose to look at the top 100 of variants prioritized by PSAP-genomic-regions-ACS to have a manageable number of variants to explore, but some of them could have been missed due to the restriction to the top 100. PSAP still performs better at prioritizing pathogenic non-coding variants compared to the CADD or ACS score alone which confirms the results of a previous studies that found the CADD score had limited clinical applicability to prioritize non-coding pathogenic variants, in the case of a hereditary cancer panel(Mather *et al.* 2016).

Some of the filters we used to prioritize the coding candidate variants in Malakand phase 1 and phase 2 families were not applicable for non-coding variants, like the expression of a gene in testis for variants that were not directly linked to the regulatory regions of a gene. Thus, we had a lot more variants to look through and less straightforward arguments to imply them or not with the pathology. Families 5 and 8 from the Malakand phase 1 dataset were good candidates for the involvement of a non-coding variant in the pathology, considering a deep exploration of the coding genome yielded no candidate variants for either families. However, there were no variants that stood out from our analysis as potentially disease-causing, either because it was not included in the analysis itself or we were not able to prioritize it over other variants.

Now that WGS data is becoming more available, the importance and difficulty of predicting the impact of these non-coding variants is starting to be more and more described in the literature (Rojano *et al.* 2019; Barbosa *et al.* 2023). For regulatory variants, a number of integrated annotation and sometimes prioritization tools exist, including RegulomeDB (Dong *et al.* 2023), HaploReg (Ward and Kellis 2016), FunciSNP (Coetzee *et al.* 2012) or the recently developed GREEN-DB (Giacopuzzi, Popitsch and Taylor 2022) (with associated annotation tool GREEN-VARAN) which all use the ENCODE project as their main data source. As we have mentioned in Part II - Chapter 2, several pathogenicity scores span the whole genome, like CADD, and some have been developed specifically for non-coding variant prioritization. They serve a purpose similar to our prioritization by PSAP-genomic-regions, but they are limited by the current understanding of the regulatory machinery encoded in the non-coding genome. A comprehensive review (Barbosa *et al.* 2023) of all computational tools to predict the impact of

variants affecting introns showed that the performance of such tools depends heavily on the type of variant: deep intronic variants outside of splice sites were poorly predicted whilst splicing-altering variants were better prioritized, and even the prioritization of splicing variants depended on the molecular mechanism altered by the variant. Other types of non-coding variants, like intergenic variants not associated with any molecular function, are even more difficult to interpret in a clinical context.

9.4 MULTI-OMICS INTEGRATION

No matter the approach used to analyze sequencing data and the potential contribution made by detecting non-coding variants or more complex modes of inheritance, the increased diagnosis rate in unsolved RD cases using WES or WGS seems to plateau between 35% and 50% (Frésard and Montgomery 2018). It is thus necessary to develop other approaches that can highlight in a more holistic way the signature of a disease to prioritize disease-causing variants and evaluate the impact of variants on various cellular products, like gene transcripts, proteins, or metabolites. These approaches are referred to as multi-omics approaches, as they integrate multiple high-throughput screening technologies like genomics, transcriptomics, proteomics or metabolomics (Figure 9.2).

A number of recent studies highlight the increase in molecular diagnostic for RDs when combining DNA and RNA sequencing (RNA-seq) analyses (Frésard *et al.* 2019; Prokisch 2019). RNA-seq delivers quantitative data on RNA expression levels which allows a better understanding and interpretation of a variant's effect, and is complementary to the variant information provided by DNA sequencing. Whilst complementing genetic data with transcriptomic data is starting to become more widespread, the increased utility of introducing additional omics approaches like proteomics into diagnostic workflows is getting traction (Stenton *et al.* 2020).

Methods like PSAP can evolve along this trend by integrating new layers of information. This extension to integrative omics analysis had already been mentioned in the initial (Wilfert *et al.* 2016) article. They calculated the likelihood that each gene in the genome would play an important role in a particular individual's disease based on GTEx expression. Combined with the gene-based PSAP-value, the expression information improved the ranks and thus the identification of disease-causing genes. This idea shows that PSAP could be adapted by integrating RNA-seq or even protein-protein interaction data to boost prediction and narrow down the list of candidate variants.

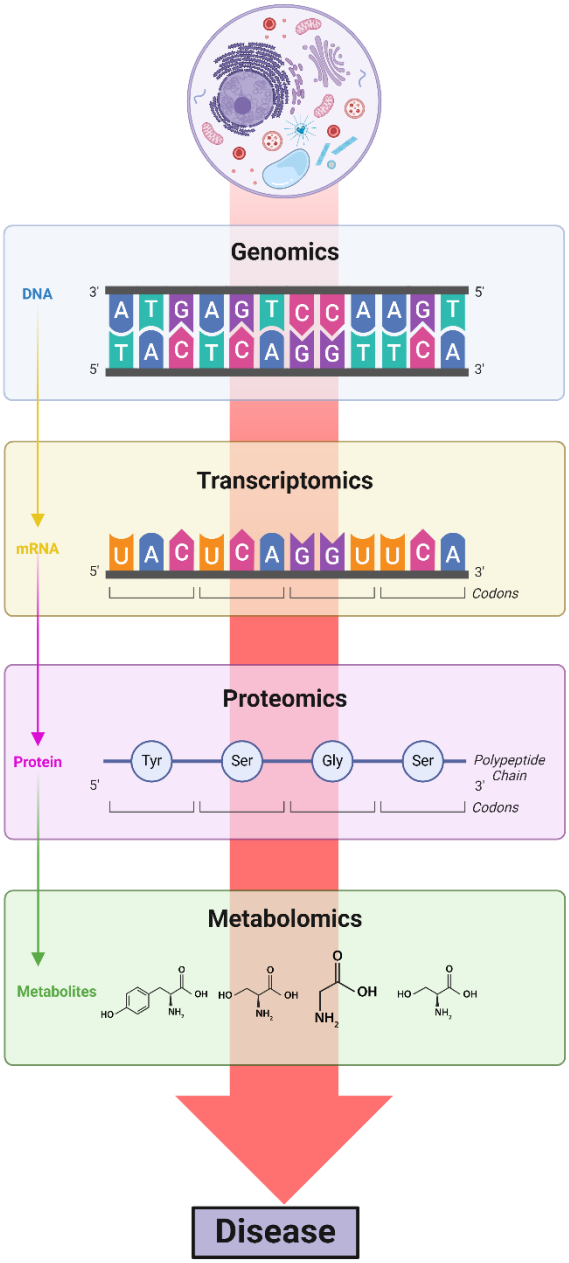


Figure 9.2 : Multi-omics approaches

Adapted from BioRender

9.5 CLINICAL APPLICABILITY IN THE FIELD OF RARE DISEASES

9.5.1 New technologies and artificial intelligence

The cost of sequencing has never been as low, with a single genome costing as little as 100$. In addition, new technologies like long-read sequencing continue to allow a better detection of variants, especially structural variants. As a consequence, the volume of sequencing data is exponentially increasing and so does the challenge of interpreting such data. This new wealth of data and increasing knowledge about the genome has allowed the training of Artificial Intelligence (AI) models, which is the development of computer systems able to perform tasks normally requiring human intelligence.

In the context of genome annotation and variant classification, the CADD score as well as other ML scores are already used to predict the impact of a genetic variation on functional elements of the genome. Deep-learning based approaches are starting to emerge in the field and seem to outperform other ML approaches for variant pathogenicity prediction and the detection of regulatory elements among other tasks (Alharbi and Rashid 2022). For instance, the DANN score (Quang, Chen and Xie 2015) is trained on the same set of features than CADD but uses a deep neural network (type of deep learning model) and appears to perform better than CADD. In addition, other deep learning-based models can achieve state-of-the-art predictions of pathogenicity by leveraging protein domains and conserved amino acid positions. These models include PrimateAI (Sundaram *et al.* 2018), which uses cross-species information, or the recently developed AlphaMissens e(Cheng *et al.* 2023), which is based on the protein 3D structure prediction tool AlphaFold2(Jumper *et al.* 2021). A better in-silico evaluation of variant pathogenicity by AI-based models would boost the predictions of a method like PSAP, which relies on a pathogenicity score for its calibration. AI-models and deep learning in particular also allow the integration of multi-omics data in unprecedented ways to generate predictive models (Sharma *et al.* 2024).

Although the promises of AI in the field of genomics are remarkable, key problematics have to be kept in mind when using AI tools. Some problematics are general in the field of AI but especially in a clinical setting, interpretability of AI models and accounting for machine bias due to substructure in the input data is crucial (Dias and Torkamani 2019). Indeed, the current focus on generating massive data to train AI models needs to be nuanced by the necessity of interpreting their results and using them to improve patient outcome.

9.5.2 Genetic data availability and data sharing

As generating data becomes easier, new challenges are arising regarding data storage and data access. Indeed, the specificity of genomic data, especially in a healthcare setting, makes it the subject of very strict national regulatory rules. In a clinical setting, genetic data are generated to provide a diagnosis for a RD patient and meet their immediate heath needs. Outside of this aim, the reanalysis of already-generated data tests the boundaries of the duty of care, especially if incidental findings are discovered for a patient. Structuring data access and consent processed is fundamental to allow global data sharing and analysis without breaching the public's trust (Stark *et al.* 2019). This globalized access to WGS data is crucial in a field like RDs research, to allow for federated analysis of patients, comparison and scalability of analyses (Umlai *et al.* 2022). In addition, the increasing availability of genomic data is bound to improve the ML algorithms for variant prioritization, as we have mentioned in the previous section. If carried out in a proper way that protects the privacy of patient data, this will benefit RDs patients both at the stage of molecular diagnosis and eventually, personalized treatment. For instance, initiatives like the virtual platform of the European Joint Programme on Rare Diseases currently gather resources (e.g. relevant knowledge databases or registries) for RD research and will eventually give access to patient data to create a federated environment for discovery whilst preserving patient privacy.

9.5.3 Impact for patient care

The impact for patients is at a forefront of any methodological developments in the field of RD diagnosis. PSAP is still a tool under development, which has proven to add valuable information in real-life scenarios of genetic diagnosis. Nonetheless, it had not been tested in a clinical setting and would benefit from further evaluation to really assess how it could be used in the context of patient clinical diagnosis. With proper information on the significance of PSAP p-values, the method could be integrated to clinical VCF annotation pipelines and be used along other pathogenicity scores to prioritize potentially disease-causing variants. A study has described that it takes on average 17 years for research evidence to be implemented in the clinics (Morris, Wooding and Grant 2011). It thus depends on researchers as well as other actors in the system to ensure that our tools are really translational and that patients benefit from developments in genomic medicine. By taking a step towards improving whole-genome analysis, PSAP-genomic-regions participates to the global effort to use of the information from whole sequencing to improve RD diagnosis (Umlai *et al.* 2022).

To move beyond the issue of diagnosis in RDs, the next frontier is patient care and therapeutics. Indeed, elucidating RDs remains a crucial issue both for patients and for the mechanistic insights they can provide on disease etiology (Antonarakis *et al.* 2010) but the diagnosis is only a first

albeit crucial step for patients. A genetic diagnosis can help stratify therapies for RDs (Wright, FitzPatrick and Firth 2018) and often results in tangible clinical actions that can have a significant impact of patient's lives (Pagnamenta *et al.* 2023) although RDs therapeutics remain scarce. However, a new kind of "odyssey" can emerge once the "diagnostic odyssey" will become less prevalent: the patient care odyssey. This patient care odyssey is defined by the inability to coordinate patient care as a whole and taking into account the familial element, despite the presence of a genetic diagnosis. The new challenge will be to refer as adequately as possible a patient from a specialized center for diagnosis to a specialized center for patient care. In France, the problematics of RDs diagnosis and therapeutics are the focus of National Plans for Rare Diseases (Plan National Maladies Rares, PNMR). While the first plans helped to reduce diagnostic wandering and deadlock, the 4th PNMR plan to reinforce the development of therapies and innovation, so that access to treatments becomes increasingly effective for people affected by RDs.

9.6 General conclusion

This thesis has approached the issue of improving the genetic diagnostic yield for RDs by tackling three main reasons for RD lack of diagnosis: genetic heterogeneity, non-coding variants and complex modes of inheritance. With PSAP-genomic-regions, we both addressed the issue of heterogeneity and variant prioritization in the non-coding genome that complicate RD diagnosis. By integrating both PSAP-genomic-regions and PSAP-genes into the Easy-PSAP pipeline, we gave other researchers and clinicians access to our latest developments around the method as well as the possibility of calculating their own PSAP null distributions. The issue of tackling complex modes of inheritance is still an open subject for reflection, although our review on the subject offers a strong understanding of what has currently been proposed in the domain. Further developments for PSAP also include the development of new features for Easy-PSAP, the integration of multi-omics data and the translation to clinical practice.

During the course of my PhD project, I was confronted with the challenge of developing and implementing a statistical tool that offers an innovative solution to the problem of RD diagnosis, but that would also be applicable in real-life setting. Confronting myself with the application of Easy-PSAP to different datasets, both in France and the US, made me realize the importance of portability and user-friendliness of a bioinformatic tool, as well as the importance of constantly maintaining and updating it. In the future, we endeavor to make Easy-PSAP truly accessible for clinicians even without an expertise in bioinformatics.

I benefited immensely during my PhD project of several collaborations, which are all highlighted in this manuscript. First, I learnt a lot from the knowledge on CSVD of Elisabeth Tournier-Lasserve and her team (U1141 Neurodiderot, Paris), around the disease processes and implicated causal variants especially, as well as the expertise in networks and computational methods of Anaïs Baudot and her team (UMR 1251 Marseille Medical Genetics). This collaboration was initiated in the context of the GENETWORK4DIAG project, which aimed at developing novel strategies based on gene networks and computational approaches to identify causal genes in undiagnosed cases of CSVD. I also had the privilege of visiting Pr Conrad's team (Oregon National Primate Research Center, US) who developed the initial PSAP method and are experts on genetic forms of male infertility. Their experience was invaluable for the interpretation of the Malakand families' results and fostered valuable reflections about the potential developments for PSAP in the context of digenism. These sustained and constructive collaborations emphasize the importance and added value of interdisciplinarity in research, which is in line with my background as a MD-PhD student.

Our answers to this complex topic of RD diagnosis are small contributions to the fast-evolving field of RD research. Ultimately, this thesis endeavors to give pertinent tools and insights to clinicians and researchers in their efforts to elucidate the genetic causes of RDs, in order to provide better molecular diagnostics, management and therapeutic options to patients living with these elusive genetic disorders.

Résumé en français

Stratégies statistiques exploitant les données de la population générale pour aider au diagnostic des maladies rares

Par Marie-Sophie OGLOBLINSKY

INSERM UMR1078 – Génétique, Génomique fonctionnelle et Biotechnologies

Université de Bretagne Occidentale

INTRODUCTION

Les maladies rares (MR) constituent un groupe hétérogène de pathologies classiquement caractérisées par une très faible prévalence (moins d'un individu sur 2,000 en Europe). Les MR englobent de 5 000 à plus de 9 000 pathologies et ce nombre ne cesse de croître, de nouvelles pathologies étant décrites chaque année (Ferreira 2019; Haendel *et al.* 2020; Sequeira *et al.* 2021). Alors que les maladies rares sont par définition rares lorsqu'elles sont considérées séparément, elles sont fréquentes dans leur ensemble et touchent environ 350 millions de personnes dans le monde, dont 3 millions en France, soit près d'un citoyen français sur vingt. Les MR sont souvent chroniques, sévères et altèrent significativement la qualité de vie des patients. À ce problème de santé publique s'ajoute le fait que près de 50 % des MR ne sont pas diagnostiquées et que, lorsqu'elles le sont, c'est le plus souvent après des mois, voire des années (Boycott *et al.* 2017). On parle dans ce cas d'une « errance diagnostique ». Le diagnostic est une étape cruciale pour les patients, car il permet de mieux comprendre leur maladie et d'améliorer leur prise en charge (Uhlenbusch, Löwe and Depping 2019). Ce diagnostic si difficile à obtenir est souvent d'ordre génétique ou moléculaire puisque l'on estime qu'environ 80 % des maladies rares sont d'origine génétique (Wright, FitzPatrick and Firth 2018).

À ce jour, la plupart des RD dont la cause génétique est connue sont décrites comme ayant un mode d'hérédité monogénique, selon lequel l'altération d'un seul gène est responsable de la maladie. Les RDs ainsi que les polymorphismes dans la population humaine sont dus à des modifications stables de la séquence d'ADN, appelées mutations ou variants. Lorsque le nucléotide de référence est remplacé par un autre nucléotide, on parle de variant nucléotidique simple (SNV). Les InDel sont des insertions ou des suppressions de quelques nucléotides dans la séquence d'ADN. Les variants que l'on trouve chez moins de 1 % de la population générale sont appelées variants rares. Les variants de l'ADN peuvent être dit codants ou non-codants. En effet, l'information génétique contenue dans les parties codantes des gènes ou exons constitue l'exome d'un individu, tandis que l'ensemble de l'information

génétique d'un individu constitue son génome. Le génome codant ne représente qu'environ 2 % du génome total, le reste du génome est donc appelé « non codant » et contient, entre autres, des éléments impliqués dans la régulation de l'expression des gènes qui peuvent avoir un lien avec les maladies humaines (Dunham *et al.* 2012).

Ces variations de notre code génétique sont essentielles à notre évolution et à notre survie à long terme. Toutefois, un très faible pourcentage de variants génétiques peut également entraîner des maladies comme les MR. Le génome de chaque individu sain contient des milliers de variants, ce qui signifie que la plupart des variants n'ont aucune conséquence biologique détectable pour la cellule ou l'organisme. L'impact réel d'un variant est complexe à prévoir (Zschocke, Byers and Wilkie 2023), bien qu'un grand nombre d'outils de prédiction bio-informatiques et de scores ont été développés depuis lors pour évaluer l'impact délétère et la pathogénicité potentielle des variations génétiques (Eilbeck, Quinlan and Yandell 2017; Garcia, Andrade and Palmero 2022).

Aujourd'hui, nous disposons d'une quantité croissante de données génétiques grâce aux progrès des technologies de séquençage de l'ADN. Le processus d'identification des variants liés à la maladie parmi les milliers ou millions de polymorphismes non pathogènes et d'erreurs de séquençage produits par le séquençage de l'exome entier (WES) et le séquençage du génome entier (WGS) est un véritable défi. La question qui se pose donc maintenant est la suivante : comment analyser ces données pour détecter les maladies rares non diagnostiquées ?

Pour comprendre pourquoi cette question est si difficile, il convient de mentionner quelques faits essentiels concernant les MR. L'architecture génétique des MR est très complexe et peut être différente d'une MR à l'autre. Mes travaux se sont donc concentrés sur trois facteurs importants qui peuvent expliquer les lacunes du diagnostic génétique des maladies rares : l'hétérogénéité génétique, les variants non codants et les modes complexes d'hérédité. En effet, de nombreux paradigmes conventionnels de la génétique, notamment le modèle "un gène, une maladie", ne rendent pas compte de DIversité génétique observée dans les MR. La plupart des MR sont caractérisées par une forte hétérogénéité génétique, où des variants dans différents gènes conduisent à des maladies phénotypiquement indiscernables. L'existence de plusieurs gènes causaux contribuant à un phénotype clinique unique rend le regroupement des patients encore plus difficile, car il n'y a souvent qu'un seul individu porteur d'un variant causal spécifique. En outre, la disponibilité croissante des données de séquençage du génome entier a révélé l'importance des variants non codants dans l'étiologie des MR, ce qui remet en question les approches traditionnelles du diagnostic génétique centrées sur le codage. Les variants non codants peuvent avoir des effets profonds sur la régulation de l'expression des gènes et la fonction des protéines, dévoilant ainsi de nouveaux mécanismes sous-jacents aux maladies rares.

Les modèles complexes d'hérédité viennent s'ajouter à la nature déjà difficile de la génétique des maladies rares. Le principal exemple de transmission génétique complexe exploré dans ce travail est le digénisme, selon lequel l'effet concomitant de variants dans deux gènes distincts sont nécessaires pour développer une maladie (Schäffer 2013).

OBJECTIFS

Le principal défi que j'ai relevé dans le cadre de mon projet de thèse est le suivant : comment améliorer les méthodes d'analyse des informations génétiques d'un patient atteint d'une MR afin de lui proposer un diagnostic génétique ? Mon objectif était de fournir des méthodes statistiques et bio-informatiques qui pourraient facilement être appliquées dans un contexte clinique ou de recherche et qui offriraient des informations précieuses pour le diagnostic des RDs.

ABORDER L'HETEROGENEITE GENETIQUE DANS LE GENOME CODANT ET NON CODANT

Dans cette partie, je discute les stratégies que j'ai mises en œuvre pour aborder le problème de l'hétérogénéité génétique dans les MR et analyser les variants dans le génome non-codant. Comme mentionné précédemment, la forte hétérogénéité génétique dans les MR pose le problème de l'identification du variant causal d'un patient à l'aide de données de séquençage et de méthodes d'analyse standard. Initialement, la méthode PSAP (Population Sampling Probability) (Wilfert *et al.* 2016) avait été développée pour apporter une résolution au problème de l'hétérogénéité génétique dans le génome codant. PSAP utilisait des distributions nulles de scores de pathogénicité CADD (Kircher *et al.* 2014) par gène pour évaluer la probabilité d'observer un génotype donné dans une population saine. Nous proposons PSAP-genomic-regions (Ogloblinsky *et al.* 2024), une extension de la méthode PSAP au génome non codant utilisant comme unités de test des régions prédéfinies reflétant la contrainte fonctionnelle à l'échelle du génome entier, les régions CADD (Bocher *et al.* 2022). Notre méthode est composée de deux stratégies alternatives basées sur la construction des distributions nulles PSAP sur les régions CADD avec deux scores de pathogénicité : le score CADD initial (stratégie PSAP-genomic-regions-CADD) ou l'ACS (stratégie PSAP-genomic-regions-ACS), un score CADD réajusté par régions fonctionnelles construit pour atténuer les scores CADD plus élevés des variants codants.

Nous avons évalué les stratégies de priorisation proposées en utilisant des exomes et des génomes de patients générés artificiellement. Nous avons généré ces exomes et génomes de patients en insérant des SNVs pathogènes codants et non codants de la base de données ClinVar dans 574 exomes de la population générale issus du projet FrEnch Exome (FREX) (Génin *et al.* 2017) et dans 533

génomes entiers issus du projet 1000 Génomes (Byrska-Bishop *et al.* 2022), respectivement, selon les modèles autosomal dominant et récessif (AD et AR). Les variants insérés ont été classés en fonction de leurs p-valeurs de PSAP d'une part, et de leur score de pathogénicité seul d'autre part. Ce protocole d'évaluation nous a permis de comparer nos deux stratégies PSAP-régions génomiques-CADD et PSAP-régions génomiques-ACS à la stratégie initiale PSAP-genes (également appelée PSAP-genes-CADD) et à une priorisation utilisant uniquement le score CADD ou ACS maximal par région CADD.

Sur les génomes de patients générés artificiellement, les deux stratégies PSAP-genomic-regions sont systématiquement plus performantes pour prioriser tous les types de variants pathogènes par rapport aux stratégies utilisant uniquement le score maximal de pathogénicité (CADD ou ACS en fonction de la stratégie). Pour la priorisation des variants pathogènes codants, PSAP-genomic-regions-CADD donne les meilleures performances et parvient à classer 45,5 % et 96 % des variants dans les 10 premiers du génome pour les modèles AD et AR, respectivement. PSAP-genomic-regions-ACS permet la meilleure priorisation des variants pathogènes non codants, en particulier aux variants d'épissage, avec 56,5% et 83,3 % des variants atteignant le top 10 du génome pour les modèles AD et AR, respectivement.

Nous avons également testé notre méthode sur des données d'exome de 6 patients avec des variants connus causant une forme monogénique de maladie des petits vaisseaux cérébraux (Cerebral Small Vessel Disease, CSVD) (Aloui *et al.* 2021) et des données de génome de 9 patients avec des formes familiales d'infertilité masculine (Khan *et al.* 2023). Dans l'ensemble, les stratégies PSAP donnent toujours de meilleurs résultats que le score CADD seul. PSAP-genomic-regions priorise les variants causaux parmi les 100 premiers variants pour chaque individu. PSAP-genomic-regions améliore considérablement le classement des variants causaux chez 4 des 6 individus atteints de CSVD par rapport à PSAP-genes, et conserve un classement similaire pour les 2 individus restants. Sur les données de génome, PSAP-genomic-regions classe les variants candidats à des rangs plus élevés que PSAP-genes. Cela peut s'expliquer par le fait que PSAP-gene ne classe qu'environ 4,000 variants par individu, car il n'analyse que les variants tombant dans les gènes, contre environ 70,000 variants pour PSAP-genomic-regions, qui analyse l'ensemble du génome.

PSAP-genomic-regions est donc un outil de priorisation efficace, qui offre des résultats prometteurs pour le diagnostic de cas non résolus de MR. Pour hiérarchiser les variants non codants, le PSAP-genomic-regions-ACS donne les meilleurs résultats à la fois pour les données WES et WGS. Dans le cas spécifique de la priorisation des variants codants en WGS, l'utilisation des parties codantes des régions CADD comme unités d'analyse (PSAP-coding-genomic-regions) donne de meilleurs

résultats que PSAP-genes. Ainsi, si le variant causal attendu est codant, nous conseillons l'utilisation de PSAP-genomic-regions-CADD pour le WES et de PSAP-coding-genomic-regions-CADD pour le WGS.

Pendant le développement de PSAP-genomic-regions, j'ai proposé plusieurs mises à jour et extensions du pipeline PSAP et j'ai implémenté des scripts efficaces pour effectuer un calcul rapide des distributions nulles PSAP avec un choix flexible de paramètres d'entrée. Afin de mettre à disposition une version actualisée et plus facile à utiliser du pipeline PSAP, j'ai créé Easy-PSAP, un workflow Snakemake facile d'utilisation, flexible et efficace en termes de calcul pour créer et appliquer toutes les distributions nulles PSAP actuellement développées. Comme indiqué précédemment, la méthode PSAP a été mise au point pour résoudre le problème de la priorisation des variants pour un seul patient, en s'appuyant sur les fréquences alléliques des bases de données de population et sur un score de pathogénicité de le variant. Cependant, la mise en œuvre initiale de PSAP comportait des scripts bash et R pour appliquer les distributions nulles de PSAP qui n'étaient pas facilement adaptables à tous les utilisateurs. Les premières distributions nulles de PSAP utilisaient également ExAC comme panel de référence pour les fréquences alléliques et CADD v1.0, et n'ont pas été mises à jour depuis. Les scripts permettant de générer les distributions nulles PSAP n'avaient pas non plus été mis à disposition.

Nous décrivons ici Easy-PSAP, une nouvelle implémentation mise à jour comprenant deux pipelines faciles d'utilisation et adaptables basés sur le principe de PSAP, qui peut évaluer les variants génétiques à l'échelle d'un génome entier en utilisant des informations provenant des dernières bases de données de population et d'annotation. Contrairement au PSAP initial qui était limité à l'exome, Easy-PSAP permet l'analyse des variants dans le génome codant et non codant en intégrant à la fois PSAP-genes et PSAP-genomic-regions dans les paramètres et les distributions nulles disponibles du pipeline. Easy-PSAP comprend à la fois un workflow pour calculer les distributions nulles PSAP et un workflow pour les appliquer aux données des patients. Ces caractéristiques, ainsi que l'accessibilité du pipeline pour les chercheurs et les cliniciens, font d'Easy-PSAP un outil de pointe pour l'analyse des données NGS qui est mis en œuvre pour évoluer au fur et à mesure que de nouveaux cadres et bases de données deviennent disponibles. En particulier, le workflow pour calculer de nouvelles distributions nulles PSAP permet aux chercheurs d'adapter PSAP à leur question de recherche et à leurs exigences.

Enfin, la performance des PSAP-genomic-regions est mise en évidence dans cette partie par une application à des cas réels de RD, dans des familles consanguines touchées par l'infertilité masculine venant de la région Malakand du Pakistan. Nos données comprennent deux jeux de données de WGS distincts pour chaque phase du projet : la première phase sera appelée Malakand phase 1 (pour laquelle des variants candidats avaient déjà été trouvés par [Khan et al. 2021]) et la seconde phase Malakand phase 2. En ce qui concerne les familles de la Malakand phase 1, le variant dans le

gène *TUBA3C* est le seul que nous ayons identifié dans la famille 7, conformément aux découvertes précédentes. Dans les familles 3 et 4, nous avons priorisé les variants déjà candidats dans *SPAG6* et *CCDC9*, respectivement. Pour la famille 3, nous avons identifié un autre variant dans *SYN3*. Cependant, le variant dans le gène SPAG6 semble être un candidat plus fort en raison des nombreuses associations dans la littérature de ce gène avec l'infertilité masculine et d'un rang de PSAP plus favorable. Un autre variant candidat a été identifié dans la famille 4, dans le gène *RSPH6A*. Les deux gènes *RSPH6A* et *CCDC9* semblent être de bons candidats pour expliquer le phénotype d'infertilité masculine, bien que le variant dans *RSPH6A* ait un impact prédit plus sévère.

Notre analyse de l'ensemble des données Malakand phase 2 nous a aussi permis de prioriser un autre variant de *TUBA3C*, dans la même région CADD que le variant de l'analyse des Malakand phase 1. Dans l'ensemble, ce deuxième variant fortement priorisé confirme l'implication potentielle d'altérations de *TUBA3C* dans l'infertilité masculine dans ces deux familles. En outre, nous avons mis en évidence un variant candidat fort pour la famille 15 dans le gène *DYNLRB2*.

ÉVALUATION D'UN MODE D'HEREDITE COMPLEXE : LE MODELE DIGENIQUE

Nous avons décrit l'architecture génétique complexe des MR, y compris la possibilité que certaines MR puissent être caractérisées par une hérédité digénique (Digenic Inheritance, DI) au niveau moléculaire. Bien que la détection de DI puisse aider à diagnostiquer un certain nombre de MR, ce n'est pas une tâche facile et il n'existe à ce jour pas de méthode de référence. La plupart des cas de DI actuellement décrits ont été détectés par des analyses familiales et impliquent souvent des gènes déjà associés à la maladie de manière monogénique. Les bases de données DIDA (DIgenic diseases DAtabase) (Gazzo *et al.* 2016) et sa version mise à jour OLIDA (OLIgenic diseases Database) (Nachtegael *et al.* 2022) répertorient toutes les combinaisons digéniques et oligogéniques connues de la littérature. Depuis 2018, un certain nombre de méthodes bio-informatiques ont donc été développées pour aborder la question de la détection de DI sans s'appuyer autant sur les pédigrées et les connaissances préalables sur la maladie. J'ai réalisé une revue de la littérature et un benchmark des méthodes actuellement publiées pour détecter la DI dans les données de séquençage. Cela a permis d'avoir une vue d'ensemble de la force et des limites des méthodes développées à ce jour.

Tout d'abord, une revue de la littérature nous a permis de classer les méthodes de détection de DI en trois catégories distinctes : les méthodes basées sur les réseaux, les méthodes statistiques et les méthodes d'apprentissage automatique (Machine Learning, ML). La seule méthode dans la catégorie basée sur les réseaux est OligoPVP (Boudellioua *et al.* 2018) qui utilise PVP (Boudellioua *et al.* 2019), un prédicteur monogénique de pathogénicité basé sur la caractérisation phénotypique du

patient, pour évaluer les variants uniquement dans les paires de gènes connectés dans un réseau d'interaction protéine-protéine. Deux méthodes entrent dans la catégorie des méthodes statistiques : la méthode digénique (DM) (Kerner *et al.* 2020) et RareComb (Pounraja and Girirajan 2022). La DM utilise des tests de fardeaux adaptés pour détecter la DI ou un variant modificateur commun d'une maladie monogénique. RareComb utilise l'algorithme Apriori (Agrawal *et al.* 1996) pour dénombrer les combinaisons de variants vus simultanément chez les cas et les témoins. Les deux méthodes impliquent un pré-filtrage des variants afin de limiter le nombre de combinaisons à tester et de ne conserver que les variants rares prédits comme pathogènes. Toutes les autres méthodes de détection de DI peuvent être classées comme des méthodes ML et sont entraînées sur des paires pathogènes de DIDA ou OLIDA et des paires neutres de la population générale. Parmi ces méthodes, VarCoPP (Papadimitriou *et al.* 2019) et son successeur VarCoPP2.0 (Versbraegen *et al.* 2023), qui sont exécutés par la plateforme ORVAL (Renaux *et al.* 2019), prennent en données d'entrée des paires de variants, tandis que DiGePred (Mukherjee *et al.* 2021), DIEP (Yuan *et al.* 2022) et ARBOCK (Renaux *et al.* 2023) sont appliqués à des paires de gènes.

Ces méthodes de ML étant les plus facilement applicables, en particulier dans le cas des MR, et ayant des types d'entrée et de sortie similaires, nous les avons comparées à l'aide d'un benchmark. Nous avons sélectionné des paires de gènes pathogènes connues dans la base de données OLIDA, qui n'ont pas été utilisées pour entraîner les méthodes de ML, et différents scénarios de paires de gènes neutres en utilisant des variants de la base de données FREX. ARBOCK et DIEP ont catégorisé un plus grand nombre de paires OLIDA comme digéniques (plus de 80 %), mais ARBOCK a également classé environ 30 % des paires observées dans la population générale comme pathogènes. En revanche, DiGePred et VarCoPP2.0 ont classé un peu moins de paires OLIDA comme pathogènes (64 % et 49 %, respectivement), tout en maintenant le nombre de faux positifs à un niveau beaucoup plus bas (moins de 1 % et 3 %, respectivement).

Dans l'ensemble, nous avons pu montrer que les méthodes basées sur les réseaux et les méthodes statistiques font de fortes hypothèses sur l'interaction entre les gènes ou le type de variants pour détecter la DI. Les méthodes statistiques nécessitent également des données de cohortes en entrée pour avoir une puissance statistique suffisante. En revanche, les méthodes de ML présentent l'avantages de pouvoir être appliquées à l'échelle individuelle pour détecter l'ID, ce qui est pertinent dans le cas de maladies rares très hétérogènes, bien qu'aucune méthode ne puisse analyser de manière exhaustive toutes les paires potentielles de gènes d'un individu. Notre benchmark des méthodes de ML pour la détection de DI nous a permis de mettre l'accent sur la méthode DiGePred, qui s'est distinguée par un nombre de faux positifs le plus faible, quel que soit le scénario, tout en conservant un nombre conséquent de prédictions réellement positives.

CONCLUSION

Au cours de ma thèse, j'ai abordé la question de l'amélioration du diagnostic génétique des MR en m'attaquant à trois raisons principales de l'absence de diagnostic de certaines MR : l'hétérogénéité génétique, les variants non codants et les modes complexes d'hérédité. Avec PSAP-genomic-regions, nous avons abordé la question de l'hétérogénéité et de la priorisation des variants dans le génome non codant. En intégrant à la fois PSAP-genomic-regions et PSAP-genes dans le pipeline Easy-PSAP, nous avons donné à d'autres chercheurs et cliniciens l'accès à nos derniers développements autour de la méthode ainsi que la possibilité de calculer leurs propres distributions nulles de PSAP. La question du traitement des modes complexes d'hérédité, particulièrement le digénisme ici, reste un sujet de réflexion ouvert, bien que notre revue sur le sujet offre une solide compréhension de ce qui a été actuellement proposé dans le domaine. Les développements futurs autour de la méthode PSAP comprennent également le développement de nouvelles fonctionnalités pour Easy-PSAP, l'intégration de données multi-omiques et l'application en pratique clinique.

References

Abbasi F, Miyata H, Shimada K *et al.* RSPH6A is required for sperm flagellum formation and male fertility in mice. *J Cell Sci* 2018;**131**:jcs221648.

Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics* 2013;**76**:7.20.1-7.20.41.

Agrawal R, Mannila H, Srikant R *et al.* Fast discovery of association rules. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, et al. (eds.). *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press, 1996, 307–28.

Alharbi WS, Rashid M. A review of deep learning applications in human genomics using next-generation sequencing data. *Hum Genomics* 2022;**16**:1–20.

Aloui C, Hervé D, Marenne G *et al.* End-Truncated LAMB1 Causes a Hippocampal Memory Defect and a Leukoencephalopathy. *Annals of Neurology* 2021;**90**:962–75.

Altshuler D, Donnelly P, The International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;**437**:1299–320.

Altshuler DM, Gibbs RA, Peltonen L *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* 2010;**467**:52–8.

Amberger JS, Bocchini CA, Schiettecatte F *et al.* OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* 2015;**43**:D789–98.

Anderson D, Lassmann T. An expanded phenotype centric benchmark of variant prioritisation tools. *Hum Mutat* 2022;**43**:539–46.

Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet* 2018;**59**:253–68.

Antonarakis SE, Chakravarti A, Cohen JC *et al.* Mendelian disorders and multifactorial traits: the big divide or one for all? *Nat Rev Genet* 2010;**11**:380–4.

Ashburner M, Ball CA, Blake JA *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.

Auton A, Abecasis GR, Altshuler DM *et al.* A global reference for human genetic variation. *Nature* 2015;**526**:68–74.

Bajrami E, Spiroski M. Genomic Imprinting. *Open Access Maced J Med Sci* 2016;**4**:181–4.

Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016;**533**:452–4.

Bamshad MJ, Ng SB, Bigham AW *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011;**12**:745–55.

Barbosa P, Savisaar R, Carmo-Fonseca M *et al.* Computational prediction of human deep intronic variation. *GigaScience* 2023;**12**:giad085.

Bocher O, Ludwig TE, Oglobinsky M-S *et al.* Testing for association with rare variants in the coding and non-coding genome: RAVA-FIRST, a new approach based on CADD deleteriousness score. *PLOS Genetics* 2022;**18**:e1009923.

Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008;**40**:695–701.

Boudellioua I, Kulmanov M, Schofield PN *et al.* OligoPVP: Phenotype-driven analysis of individual genomic information to prioritize oligogenic disease variants. *Sci Rep* 2018;**8**:14681.

Boudellioua I, Kulmanov M, Schofield PN *et al.* DeepPVP: phenotype-based prioritization of causative variants using deep learning. *BMC Bioinformatics* 2019;**20**:65.

Boycott KM, Rath A, Chong JX *et al.* International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *The American Journal of Human Genetics* 2017;**100**:695–705.

Butler MG. Imprinting disorders in humans: a review. *Curr Opin Pediatr* 2020;**32**:719–29.

Bycroft C, Freeman C, Petkova D *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9.

Byrska-Bishop M, Evani US, Zhao X *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* 2022;**185**:3426-3440.e19.

Caron B, Luo Y, Rausell A. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. *Genome Biology* 2019;**20**:32.

Carter H, Douville C, Stenson PD *et al.* Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* 2013;**14 Suppl 3**:S3.

Chen S, Francioli LC, Goodrich JK *et al.* A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. 2022:2022.03.20.485034.

Chen S, Francioli LC, Goodrich JK *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* 2024;**625**:92–100.

Chen W, Coombes BJ, Larson NB. Recent advances and challenges of rare variant association analysis in the biobank sequencing era. *Front Genet* 2022;**13**, DOI: 10.3389/fgene.2022.1014947.

Cheng J, Novati G, Pan J *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 2023;**381**:eadg7492.

Choi Y, Sims GE, Murphy S *et al.* Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLOS ONE* 2012;**7**:e46688.

Chong JX, Buckingham KJ, Jhangiani SN *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics* 2015;**97**:199–215.

Chong R, Insigne KD, Yao D *et al.* A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. *Molecular Cell* 2019;**73**:183-194.e8.

Church DM, Schneider VA, Graves T *et al.* Modernizing Reference Genome Assemblies. *PLOS Biology* 2011;**9**:e1001091.

Coetzee SG, Rhie SK, Berman BP *et al.* FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res* 2012;**40**:e139.

Crouch DJM, Bodmer WF. Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proc Natl Acad Sci U S A* 2020;**117**:18924–33.

Davydov EV, Goode DL, Sirota M *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010;**6**:e1001025.

De Baets G, Van Durme J, Reumers J *et al.* SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Research* 2012;**40**:D935–9.

Desmet F-O, Hamroun D, Lalande M *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 2009;**37**:e67.

Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine* 2019;**11**:70.

Ding X, Schimenti JC. Strategies to Identify Genetic Variants Causing Infertility. *Trends in Molecular Medicine* 2021;**27**:792–806.

Dixon JR, Selvaraj S, Yue F *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**:376–80.

Dong S, Zhao N, Spragins E *et al.* Annotating and prioritizing human non-coding variants with RegulomeDB v.2. *Nat Genet* 2023;**55**:724–6.

Dorfman R, Nalpathamkalam T, Taylor C *et al.* Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene? *Clin Genet* 2010;**77**:464–73.

Dunham I, Kundaje A, Aldred SF *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.

Ehrhart F, Willighagen EL, Kutmon M *et al.* A resource to explore the discovery of rare diseases and their causative genes. *Sci Data* 2021;**8**:124.

Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. *Nat Rev Genet* 2017;**18**:599–612.

Elkon R, Agami R. Characterization of noncoding regulatory DNA in the human genome. *Nat Biotechnol* 2017;**35**:732–46.

Feng B-J. PERCH: A Unified Framework for Disease Gene Prioritization. *Hum Mutat* 2017;**38**:243–51.

Ferreira CR. The burden of rare diseases. *American Journal of Medical Genetics Part A* 2019;**179**:885–92.

Forrest ARR, Kawaji H, Rehli M *et al.* A promoter-level mammalian expression atlas. *Nature* 2014;**507**:462–70.

Frazer KA, Ballinger DG, Cox DR *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;**449**:851–61.

Frésard L, Montgomery SB. Diagnosing rare diseases after the exome. *Cold Spring Harb Mol Case Stud* 2018;**4**:a003392.

Frésard L, Smail C, Ferraro NM *et al.* Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med* 2019;**25**:911–9.

Fu Y, Liu Z, Lou S *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology* 2014;**15**:480.

Garcia FA de O, Andrade ES de, Palmero EI. Insights on variant analysis in silico tools for pathogenicity prediction. *Frontiers in Genetics* 2022;**13**.

Gazal S, Sahbatou M, Perdry H *et al.* Inbreeding coefficient estimation with dense SNP data: comparison of strategies and application to HapMap III. *Hum Hered* 2014;**77**:49–62.

Gazzo AM, Daneels D, Cilia E *et al.* DIDA: A curated and annotated digenic diseases database. *Nucleic Acids Research* 2016;**44**:D900–7.

Génin E. Missing heritability of complex diseases: case solved? *Hum Genet* 2020;**139**:103–13.

Génin E, Feingold J, Clerget-Darpoux F. Identifying modifier genes of monogenic disease: strategies and difficulties. *Hum Genet* 2008;**124**:357–68.

Génin E, Redon R, Deleuze J *et al.* The French Exome (FREX) Project: A Population-based panel of exomes to help filter out common local variants. *Genetic Epidemiology* 2017;**41**:691–691.

Gerasimavicius L, Livesey BJ, Marsh JA. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat Commun* 2022;**13**:3895.

Giacopuzzi E, Popitsch N, Taylor JC. GREEN-DB: a framework for the annotation and prioritization of non-coding regulatory variants from whole-genome sequencing data. *Nucleic Acids Research* 2022;**50**:2522–35.

Giani AM, Gallo GR, Gianfranceschi L *et al.* Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal* 2020;**18**:9–19.

Gibbs RA, Belmont JW, Hardenbol P *et al.* The International HapMap Project. *Nature* 2003;**426**:789–96.

GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;**369**:1318–30.

Gulko B, Hubisz MJ, Gronau I *et al.* Probabilities of Fitness Consequences for Point Mutations Across the Human Genome. *Nat Genet* 2015;**47**:276–83.

Gussow AB, Copeland BR, Dhindsa RS *et al.* Orion: Detecting regions of the human non-coding genome that are intolerant to variation using population genetics. *PLOS ONE* 2017;**12**:e0181604.

Haendel M, Vasilevsky N, Unni D *et al.* How many rare diseases are there? *Nat Rev Drug Discov* 2020;**19**:77–8.

Havrilla JM, Pedersen BS, Layer RM *et al.* A map of constrained coding regions in the human genome. *Nat Genet* 2019;**51**:88–95.

He S, Gillies JP, Zang JL *et al.* Distinct dynein complexes defined by DYNLRB1 and DYNLRB2 regulate mitotic and male meiotic spindle bipolarity. *Nat Commun* 2023;**14**:1715.

Hindorff LA, Sethupathy P, Junkins HA *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* 2009;**106**:9362–7.

Houston BJ, Riera-Escamilla A, Wyrwoll MJ *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* 2022;**28**:15–29.

Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* 2017;**49**:618–24.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004;**431**:931–45.

Ioannidis NM, Rothstein JH, Pejaver V *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics* 2016;**99**:877–85.

Ionita-Laza I, Makarov V, ARRA Autism Sequencing Consortium *et al.* Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. *Am J Hum Genet* 2012;**90**:1002–13.

Ionita-Laza I, McCallum K, Xu B *et al.* A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 2016;**48**:214–20.

Itan Y, Shang L, Boisson B *et al.* The human gene damage index as a gene-level approach to prioritizing exome variants. *Proceedings of the National Academy of Sciences* 2015;**112**:13615–20.

Jagadeesh KA, Wenger AM, Berger MJ *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* 2016;**48**:1581–6.

Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 2019;**176**:535-548.e24.

Jäger M, Schubach M, Zemojtel T *et al.* Alternate-locus aware variant calling in whole genome sequencing. *Genome Medicine* 2016;**8**:130.

James RA, Campbell IM, Chen ES *et al.* A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. *Genome Medicine* 2016;**8**:13.

Javed A, Agrawal S, Ng PC. Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat Methods* 2014;**11**:935–7.

Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res* 2014;**42**:13534–44.

Johnson RC, Nelson GW, Troyer JL *et al.* Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* 2010;**11**:724.

Joutel A, Corpechot C, Ducros A *et al.* Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. *Nature* 1996;**383**:707–10.

Joutel A, Haddad I, Ratelade J *et al.* Perturbations of the cerebrovascular matrisome: A convergent mechanism in small vessel disease of the brain? *J Cereb Blood Flow Metab* 2016;**36**:143–57.

Jumeau F, Chalmel F, Fernandez-Gomez F-J *et al.* Defining the human sperm microtubulome: an integrated genomics approach. *Biology of Reproduction* 2017;**96**:93–106.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.

Kajiwara K, Berson EL, Dryja TP. Digenic Retinitis Pigmentosa Due to Mutations at the Unlinked Peripherin/RDS and ROM1 Loci. *Science* 1994;**264**:1604–8.

Karczewski KJ, Francioli LC, Tiao G *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;**581**:434–43.

Katsanis N, Ansley SJ, Badano JL *et al.* Triallelic Inheritance in Bardet-Biedl Syndrome, a Mendelian Recessive Disorder. *Science* 2001;**293**:2256–9.

Kennedy B, Kronenberg Z, Hu H *et al.* Using VAAST to Identify Disease-Associated Variants in Next-Generation Sequencing Data. *Curr Protoc Hum Genet* 2014;**81**:6.14.1-6.14.25.

Kerner G, Bouaziz M, Cobat A *et al.* A genome-wide case-only test for the detection of digenic inheritance in human exomes. *PNAS* 2020;**117**:19367–75.

Khan MR, Akbari A, Nicholas TJ *et al.* Genome sequencing of Pakistani families with male infertility identifies deleterious genotypes in SPAG6, CCDC9, TKTL1, TUBA3C, and M1AP. *Andrology* 2023, DOI: 10.1111/andr.13570.

Kircher M, Witten DM, Jain P *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**:310–5.

Kleinert P, Kircher M. A framework to score the effects of structural variants in health and disease. *Genome Res* 2022a;**32**:766–77.

Kleinert P, Kircher M. A framework to score the effects of structural variants in health and disease. *Genome Res* 2022b:gr.275995.121.

Köhler S, Vasilevsky NA, Engelstad M *et al.* The Human Phenotype Ontology in 2017. *Nucleic Acids Res* 2017;**45**:D865–76.

Kolovos P, Knoch TA, Grosveld FG *et al.* Enhancers and silencers: an integrated and simple model for their function. *Epigenetics & Chromatin* 2012;**5**:1.

Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012;**28**:2520–2.

Kousi M, Katsanis N. Genetic Modifiers and Oligogenic Inheritance. *Cold Spring Harb Perspect Med* 2015;**5**:a017145.

Krausz C, Riera-Escamilla A. Genetics of male infertility. *Nat Rev Urol* 2018;**15**:369–84.

Landrum MJ, Lee JM, Benson M *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;**46**:D1062–7.

Lee S, Abecasis GR, Boehnke M *et al.* Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am J Hum Genet* 2014;**95**:5–23.

Lee S, Emond MJ, Bamshad MJ *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012;**91**:224–37.

Lek M, Karczewski KJ, Minikel EV *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;**536**:285–91.

Li B, Krishnan VG, Mort ME *et al.* Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 2009;**25**:2744–50.

Li J, Zhao T, Zhang Y *et al.* Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res* 2018;**46**:7793–804.

Li Q, Zhao K, Bustamante CD *et al.* Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Genetics in Medicine* 2019;**21**:2126.

Liu DJ, Leal SM. A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLOS Genetics* 2010;**6**:e1001156.

Liu Y, Yeung WSB, Chiu PCN *et al.* Computational approaches for predicting variant impact: An overview from resources, principles to applications. *Front Genet* 2022;**13**, DOI: 10.3389/fgene.2022.981005.

Lu Q, Hu Y, Sun J *et al.* A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep* 2015;**5**:10576.

Lupski JR. Digenic inheritance and Mendelian disease. *Nat Genet* 2012;**44**:1291–2.

Madsen BE, Browning SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLOS Genetics* 2009;**5**:e1000384.

Mahyari E, Guo J, Lima AC *et al.* Comparative single-cell analysis of biopsies clarifies pathogenic mechanisms in Klinefelter syndrome. *Am J Hum Genet* 2021;**108**:1924–45.

Mahyari E, Vigh-Conrad KA, Daube C *et al.* The Human Infertility Single-cell Testis Atlas (HISTA): An interactive molecular scRNA-Seq reference of the human testis. 2023:2023.09.23.558896.

Makrythanasis P, Antonarakis S. Pathogenic variants in non-protein-coding sequences. *Clinical Genetics* 2013;**84**:422–8.

Manolio TA, Collins FS, Cox NJ *et al.* Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.

Marenne G, Ludwig TE, Bocher O *et al.* RAVAQ: An integrative pipeline from quality control to region-based rare variant association analysis. *Genetic Epidemiology* 2022;**46**:256–65.

Marini S, Anderson CD, Rosand J. Genetics of Cerebral Small Vessel Disease. *Stroke* 2020;**51**:12–20.

Martinot E, Boerboom D. Slit/Robo signaling regulates Leydig cell steroidogenesis. *Cell Commun Signal* 2021;**19**:8.

Marwaha S, Knowles JW, Ashley EA. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Medicine* 2022;**14**:23.

Mather CA, Mooney SD, Salipante SJ *et al.* CADD score has limited clinical validity for the identification of pathogenic variants in non-coding regions in a hereditary cancer panel. *Genet Med* 2016;**18**:1269–75.

McCarthy DJ, Humburg P, Kanapin A *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine* 2014;**6**:26.

McClellan J, King M-C. Genetic Heterogeneity in Human Disease. *Cell* 2010;**141**:210–7.

McKenna A, Hanna M, Banks E *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303.

McLaren W, Gil L, Hunt SE *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* 2016;**17**:122.

Mendel G. EXPERIMENTS IN PLANT HYBRIDIZATION (1865).

Mesirov JP. COMPUTER SCIENCE. Accessible Reproducible Research. *Science* 2010;**327**:10.1126/science.1179653.

Messaoud O, Dutta AK, Cornejo-Olivas MR *et al.* Editorial: Monogenic vs. Oligogenic Reclassification. *Front Genet* 2021;**12**:821591.

Mi H, Muruganujan A, Ebert D *et al.* PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* 2019;**47**:D419–26.

Mistry J, Chuguransky S, Williams L *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* 2021;**49**:D412–9.

Mölder F, Jablonski KP, Letcher B *et al.* Sustainable data analysis with Snakemake. *F1000Res* 2021;**10**:33.

Moore JE, Purcaro MJ, Pratt HE *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 2020;**583**:699–710.

Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. *J R Soc Med* 2011;**104**:510–20.

Moyon L, Berthelot C, Louis A *et al.* Classification of non-coding variants with high pathogenic impact. *PLOS Genetics* 2022;**18**:e1010191.

Mukherjee S, Cogan JD, Newman JH *et al.* Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network. *The American Journal of Human Genetics* 2021;**108**:1946–63.

Munafò MR, Nosek BA, Bishop DVM *et al.* A manifesto for reproducible science. *Nat Hum Behav* 2017;**1**:0021.

Mustapha M, Nassir CMNCM, Aminuddin N *et al.* Cerebral Small Vessel Disease (CSVD) – Lessons From the Animal Models. *Front Physiol* 2019;**10**:1317.

Nachtegael C, Gravel B, Dillen A *et al.* Scaling up oligogenic diseases research with OLIDA: the Oligogenic Diseases Database. *Database* 2022;**2022**:baac023.

Nagirnaja L, Lopes AM, Charng W-L *et al.* Diverse monogenic subforms of human spermatogenic failure. *Nat Commun* 2022;**13**:7953.

Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;**31**:3812–4.

Nguengang Wakap S, Lambert DM, Olry A *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet* 2020;**28**:165–73.

Nicora G, Zucca S, Limongelli I *et al.* A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. *Sci Rep* 2022;**12**:2517.

Nogales E, Downing KH, Amos LA *et al.* Tubulin and FtsZ form a distinct family of GTPases. *Nat Struct Mol Biol* 1998;**5**:451–8.

Ogloblinsky M-SC, Bocher O, Aloui C *et al.* PSAP-genomic-regions: a method leveraging population data to prioritize coding and non-coding variants in whole genome sequencing for rare disease diagnosis. 2024:2024.02.13.580050.

Ong C-T, Corces VG. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* 2014;**15**:234–46.

Pagnamenta AT, Camps C, Giacopuzzi E *et al.* Structural and non-coding variants increase the diagnostic yield of clinical whole genome sequencing for rare diseases. *Genome Medicine* 2023;**15**:94.

Panoutsopoulou K, Wheeler E. Key Concepts in Genetic Epidemiology. In: Evangelou E (ed.). *Genetic Epidemiology: Methods and Protocols*. New York, NY: Springer, 2018, 7–24.

Pantoni L. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. *The Lancet Neurology* 2010;**9**:689–701.

Papadimitriou S, Gazzo A, Versbraegen N *et al.* Predicting disease-causing variant combinations. *PNAS* 2019;**116**:11878–87.

Pollard KS, Hubisz MJ, Rosenbloom KR *et al.* Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 2010;**20**:110–21.

Posey JE. Genome sequencing and implications for rare disorders. *Orphanet Journal of Rare Diseases* 2019;**14**:153.

Pounraja VK, Girirajan S. A general framework for identifying oligogenic combinations of rare variants in complex disorders. *Genome Res* 2022;**32**:904–15.

Price AL, Kryukov GV, de Bakker PIW *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010;**86**:832–8.

Prokisch H. Molecular diagnostics of Mendelian disorders via combined DNA and RNA sequencing. *Medizinische Genetik* 2019;**31**:191–7.

Pulst SM. Genetic Linkage Analysis. *Archives of Neurology* 1999;**56**:667–72.

Qian D, Li S, Tian Y *et al.* A Bayesian framework for efficient and accurate variant prediction. *PLoS One* 2018;**13**:e0203553.

Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015;**31**:761–3.

Rahit KMTH, Tarailo-Graovac M. Genetic Modifiers and Rare Mendelian Disease. *Genes* 2020;**11**:239.

Rannikmäe K, Henshall DE, Thrippleton S *et al.* Beyond the Brain. *Stroke* 2020;**51**:3007–17.

Reese MG, Eeckman FH, Kulp D *et al.* Improved splice site detection in Genie. *J Comput Biol* 1997;**4**:311–23.

Renaux A, Papadimitriou S, Versbraegen N *et al.* ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Research* 2019;**47**:W93–8.

Renaux A, Terwagne C, Cochez M *et al.* A knowledge graph approach to predict and interpret disease-causing gene interactions. *BMC Bioinformatics* 2023;**24**:324.

Rentzsch P, Schubach M, Shendure J *et al.* CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Medicine* 2021;**13**:31.

Rentzsch P, Witten D, Cooper GM *et al.* CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* 2019;**47**:D886–94.

Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011;**39**:e118.

Richards S, Aziz N, Bale S *et al.* Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;**17**:405–24.

Ritchie GRS, Dunham I, Zeggini E *et al.* Functional annotation of noncoding sequence variants. *Nat Methods* 2014;**11**:294–6.

Robette N, Génin E, Clerget-Darpoux F. Heritability: What's the point? What is it not for? A human genetics perspective. *Genetica* 2022;**150**:199–208.

Robinson PN, Köhler S, Bauer S *et al.* The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *Am J Hum Genet* 2008;**83**:610–5.

Robinson PN, Köhler S, Oellrich A *et al.* Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res* 2014;**24**:340–8.

Robinson PN, Ravanmehr V, Jacobsen JOB *et al.* Interpretable Clinical Genomics with a Likelihood Ratio Paradigm. *Am J Hum Genet* 2020;**107**:403–17.

Rojano E, Seoane P, Ranea JAG *et al.* Regulatory variants: from detection to predicting impact. *Briefings in Bioinformatics* 2019;**20**:1639–54.

Ruscheinski A, Reimler AL, Ewald R *et al.* VPMBench: a test bench for variant prioritization methods. *BMC Bioinformatics* 2021;**22**:543.

Sapiro R, Kostetskii I, Olds-Clarke P *et al.* Male Infertility, Impaired Sperm Motility, and Hydrocephalus in Mice Deficient in Sperm-Associated Antigen 6. *Molecular and Cellular Biology* 2002;**22**:6298–305.

Schäffer AA. Digenic inheritance in medical genetics. *J Med Genet* 2013;**50**:641–52.

Schaid DJ, Sinnwell JP, McDonnell SK *et al.* Detecting genomic clustering of risk variants from sequence data: cases versus controls. *Hum Genet* 2013;**132**:1301–9.

Schneider VA, Graves-Lindsay T, Howe K *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* 2017;**27**:849–64.

Schubach M, Maass T, Nazaretyan L *et al.* CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Research* 2024;**52**:D1143–54.

Schwarz JM, Rödelsperger C, Schuelke M *et al.* MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;**7**:575–6.

Sequeira AR, Mentzakis E, Archangelidi O *et al.* The economic and health impact of rare diseases: A meta-analysis. *Health Policy and Technology* 2021;**10**:32–44.

Sha Y, Xu Y, Wei X *et al.* CCDC9 is identified as a novel candidate gene of severe asthenozoospermia. *Systems Biology in Reproductive Medicine* 2019;**65**:465–73.

Sharma A, Lysenko A, Jia S *et al.* Advances in AI and machine learning for predictive medicine. *J Hum Genet* 2024:1–11.

Shihab HA, Gough J, Cooper DN *et al.* Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation* 2013;**34**:57–65.

Shihab HA, Rogers MF, Gough J *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 2015;**31**:1536–43.

Siepel A, Bejerano G, Pedersen JS *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;**15**:1034–50.

Sifrim A, Popovic D, Tranchevent L-C *et al.* eXtasy: variant prioritization by genomic data fusion. *Nat Methods* 2013;**10**:1083–4.

Singleton MV, Guthery SL, Voelkerding KV *et al.* Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am J Hum Genet* 2014;**94**:599–610.

Sironen A, Shoemark A, Patel M *et al.* Sperm defects in primary ciliary dyskinesia and related causes of male infertility. *Cell Mol Life Sci* 2020;**77**:2029–48.

Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 2008;**9**:477–85.

Smedley D, Jacobsen JOB, Jager M *et al.* Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc* 2015;**10**:2004–15.

Smedley D, Schubach M, Jacobsen JOB *et al.* A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *Am J Hum Genet* 2016;**99**:595–606.

Stark Z, Dolman L, Manolio TA *et al.* Integrating Genomics into Healthcare: A Global Responsibility. *Am J Hum Genet* 2019;**104**:13–20.

Steinhaus R, Proft S, Schuelke M *et al.* MutationTaster2021. *Nucleic Acids Res* 2021;**49**:W446–51.

Stenson PD, Mort M, Ball EV *et al.* The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum Genet* 2020;**139**:1197–207.

Stenton SL, Kremer LS, Kopajtich R *et al.* The diagnosis of inborn errors of metabolism by an integrative "multi-omics" approach: A perspective encompassing genomics, transcriptomics, and proteomics. *Journal of Inherited Metabolic Disease* 2020;**43**:25–35.

Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* 2005;**15**:978–86.

Sundaram L, Gao H, Padigepati SR *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* 2018;**50**:1161–70.

Tavtigian SV, Byrnes GB, Goldgar DE *et al.* Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. *Human Mutation* 2008;**29**:1342–54.

Thompson CS, Hakim AM. Living Beyond Our Physiological Means. *Stroke* 2009;**40**:e322–30.

Tosco-Herrera E, Muñoz-Barrera A, Jáspez D *et al.* Evaluation of a whole-exome sequencing pipeline and benchmarking of causal germline variant prioritizers. *Human Mutation* 2022;**43**:2010–20.

Uffelmann E, Huang QQ, Munung NS *et al.* Genome-wide association studies. *Nat Rev Methods Primers* 2021;**1**:1–21.

Uhlenbusch N, Löwe B, Depping MK. Perceived burden in dealing with different rare diseases: a qualitative focus group study. *BMJ Open* 2019;**9**:e033353.

Umlai U-KI, Bangarusamy DK, Estivill X *et al.* Genome sequencing data analysis for rare disease gene discovery. *Briefings in Bioinformatics* 2022;**23**:bbab363.

Versbraegen N, Gravel B, Nachtegael C *et al.* Faster and more accurate pathogenic combination predictions with VarCoPP2.0. *BMC Bioinformatics* 2023;**24**:179.

Vissers LELM, Gilissen C, Veltman JA. Genetic studies in intellectual disability and related disorders. *Nat Rev Genet* 2016;**17**:9–18.

Vitsios D, Dhindsa RS, Middleton L *et al.* Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. *Nat Commun* 2021;**12**:1504.

Wall JD, Sathirapongsasuti JF, Gupta R *et al.* South Asian medical cohorts reveal strong founder effects and high rates of homozygosity. *Nat Commun* 2023;**14**:3377.

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* 2010;**38**:e164.

Wang Z, Zhao G, Li B *et al.* Performance Comparison of Computational Methods for the Prediction of the Function and Pathogenicity of Non-coding Variants. *Genomics, Proteomics & Bioinformatics* 2023;**21**:649–61.

Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res* 2016;**44**:D877–81.

Wei W, Chinnery PF. Inheritance of mitochondrial DNA in humans: implications for rare and common diseases. *J Intern Med* 2020;**287**:634–44.

Wells A, Heckerman D, Torkamani A *et al.* Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nature Communications* 2019;**10**, DOI: 10.1038/s41467-019-13212-3.

Whittaker E, Thrippleton S, Chong LYW *et al.* Systematic Review of Cerebral Phenotypes Associated With Monogenic Cerebral Small-Vessel Disease. *Journal of the American Heart Association* 2022;**11**:e025629.

Wigby KM, Brockman D, Costain G *et al.* Evidence review and considerations for use of first line genome sequencing to diagnose rare genetic disorders. *npj Genom Med* 2024;**9**:1–11.

Wilfert AB, Chao KR, Kaushal M *et al.* Genomewide significance testing of variation from single case exomes. *Nat Genet* 2016;**48**:1455–61.

Wojcik MH, Reuter CM, Marwaha S *et al.* Beyond the exome: What's next in diagnostic testing for Mendelian conditions. *The American Journal of Human Genetics* 2023;**110**:1229–48.

Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet* 2018;**19**:253–68.

Wu G, Zhi D. Pathway-based approaches for sequencing-based genome-wide association studies. *Genet Epidemiol* 2013;**37**:478–94.

Wu H, Wang J, Cheng H *et al.* Patients with severe asthenoteratospermia carrying SPAG6 or RSPH3 mutations have a positive pregnancy outcome following intracytoplasmic sperm injection. *J Assist Reprod Genet* 2020;**37**:829–40.

Xu C, Tang D, Shao Z *et al.* Homozygous SPAG6 variants can induce nonsyndromic asthenoteratozoospermia with severe MMAF. *Reproductive Biology and Endocrinology* 2022;**20**:41.

Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods* 2015;**12**:841–3.

Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 2004;**11**:377–94.

Yuan Y, Zhang L, Long Q *et al.* An accurate prediction model of digenic interaction for estimating pathogenic gene pairs of human diseases. *Comput Struct Biotechnol J* 2022;**20**:3639–52.

Zemojtel T, Köhler S, Mackenroth L *et al.* Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med* 2014;**6**:252ra123.

Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods* 2015;**12**:931–4.

Ziegler A. Basic Mechanisms of Monogenic Inheritance. *Epilepsia* 1999;**40**:4–8.

Zschocke J, Byers PH, Wilkie AOM. Mendelian inheritance revisited: dominance and recessiveness in medical genetics. *Nat Rev Genet* 2023;**24**:442–63.

Zuccarello D, Ferlin A, Cazzadore C *et al.* Mutations in dynein genes in patients affected by isolated non-syndromic asthenozoospermia. *Human Reproduction* 2008;**23**:1957–62.

Appendix I

Supplementary Materials

PSAP-genomic-regions: a method leveraging population data to prioritize coding and non-coding variants in whole genome sequencing for rare disease diagnosis

Supplementary Materials

PSAP-genomic-regions: a method leveraging population data to prioritize coding and non-coding variants in whole genome sequencing for rare disease diagnosis

Method to generate PSAP null distributions

The PSAP null distribution for a genomic region was calculated through an analytical method in 5 steps based on allele frequencies from a reference panel and pathogenicity scores of all variants in the region:

Step 1: Annotate each variant *x* with its allele frequency *p* in the reference panel (*n* variants in total have an allele frequency for the region). A "variant" is defined here as a possible alternative allele at a given genomic position, thus several variants can have the same position. Variants can be excluded from the calculation based on the coverage in the reference panel. Variants that are not present in the panel are assigned a frequency of 0 and do not contribute to the calculation of null distribution.

Step 2: Calculate the probabilities of an heterozygote genotype and of an homozygote genotype respectively at a variant x in the reference database

$$P(x_{het}) = 2p(1-p)$$

$$P(x_{hom}) = p^2$$

Step 3: Group variants with the same pathogenicity score (*y* is a unique pathogenicity score observed in a region, with *z* variants having this particular pathogenicity score) and compute the probability of at least one heterozygote genotype, and respectively one homozygote genotype, across all the *z* variants with this *y* pathogenicity score in the reference database

$$P\left(y_{at\_least\_one\_het}\right) = 1 - \prod_{x=1}^{z} \left(1 - P\left(x_{het}\right)\right)$$

$$P\left(y_{at\_least\_one\_hom}\right) = 1 - \prod_{x=1}^{z} \left(1 - P\left(x_{hom}\right)\right)$$

Step 4: Order pathogenicity scores (*y*) by descending order (note *i* the rank of the pathogenicity score in the region, *i*=1 being the higher possible pathogenicity score in the region, y_1) and calculate the probability for pathogenicity score y_i to be the maximum pathogenicity for the AD and the AR model respectively, i.e. across the heterozygous and across the homozygous genotypes respectively

$$P\left(y_{i;no\_het}\right) = 1 - P\left(y_{i;atleast\_one\_het}\right)$$

$$P\left(y_{i;no\_hom}\right) = 1 - P\left(y_{i;atleast\_one\_hom}\right)$$

$$P\left(y_{i;CADDmax\_het}\right) = P\left(y_{1;at\_least\_one\_het}\right), if\ if\ i = 1$$
$$= P\left(y_{1;at\_least\_one\_het}\right) \times \prod_{k=1}^{i-1}\left(P\left(y_{k;no\_het}\right)\right), otherwise$$

$$P\left(y_{i;CADDmax\_hom}\right) = P\left(y_{1;at\_least\_one\_hom}\right), if\ i = 1$$
$$= P\left(y_{i;at\_least\_one\_hom}\right) \times \prod_{k=1}^{i-1}\left(P\left(y_{k;no\_hom}\right)\right), otherwise$$

Step 5: 1,400 bins of pathogenicity scores were considered ranging from 0 to 70 by steps of 0.5, considering the range of the CADD score currently used to calibrate PSAP null distributions. The cumulative probability of observing a pathogenicity score as high or higher than this bin is computed for each bin, which creates the PSAP null distributions for the AD and AR models respectively.

Supp. Table S1. Currently available PSAP null distributions
(at https://lysine.univ-brest.fr/~msogloblinsky/share/data/)

Build	Pathogenicity score	Unit of testing	Name of PSAP null distributions in Easy-PSAP for AD and AR models (* is replaced by "het" for AD and "hom" for AR model)
hg19	CADD	Gene	latest_gnomadgen_string_ensembl_cadd1.6_af_nosing_lookup_*.txt.gz
		CADD region	latest_gnomadgen_string_caddregions_cadd1.6_af_nosing_lookup_*.txt.gz
		Coding CADD region	latest_gnomadgen_string_coding_caddregions_cadd1.6_af_nosing_lookup_*.txt.gz
	ACS	CADD region	gnomadgen_string_cadd1.6_PHREDbyfunctional_af_nosing_lookup_caddregion_*.txt.gz
hg38	CADD	Gene	latest_gnomadgenv3_string_cadd1.6_af_nosing_hg38_caddregionsliftover_lookup_coding_gene_*.txt.gz
		CADD region	latest_gnomadgenv3_string_cadd1.6_af_nosing_hg38_caddregionsliftover_lookup_cadd_region_*.txt.gz
		Coding CADD region	latest_gnomadgenv3_string_cadd1.6_af_nosing_hg38_caddregionsliftover_lookup_coding_cadd_region_*.txt.gz
	ACS	CADD region	latest_gnomadgenv3_string_cadd1.6_PHREDbyfunctional_af_nosing_hg38_caddregionsliftover_lookup_cadd_region_*.txt.gz

Supp. Table S2. Strategies applied to construct and test PSAP null distributions

Type of variants		Sequence for insertion	PSAP strategy	Number of ClinVar variants
Coding variants		• Exomes FREX • NFE genomes	• PSAP-genes-CADD • PSAP-genomic-regions-CADD • PSAP-genomic-regions-ACS	• AD model: 4,965 variants • AR model: 2,680 variants
Non-coding variants	Covered in FREX	• Exomes FREX •	• PSAP-genomic-regions-CADD • PSAP-genomic-regions-ACS	• AD model: 48 variants • AR model: 64 variants
	All	• NFE genomes		• AD model: 175 variants • AR model: 102 variants

Supp. Table S3. Number and percentage of non-coding ClinVar variants in the top 10 of NFE genomes with PSAP-genomic-regions-CADD and PSAP-genomic-regions-ACS, by category of VEP consequence

(A) Autosomal Dominant model

Type	VEP consequence	N variants	% top 10 - PSAP-genomic-regions-ACS	% top 10 - PSAP-genomic-regions-CADD
Splicing	splice acceptor variant	61	91.8	52.5
	splice donor 5th base variant	10	90	0
	splice donor region variant	10	50	0
	splice region variant	9	44.4	0
	splice donor variant	16	43.8	25
	splice polypyrimidine tract variant	9	33.3	11.1
Other	5 prime UTR variant	2	100	100
	downstream gene variant	14	42.9	14.3
	intron variant	10	30	10
	upstream gene variant	31	12.9	3.2
	start lost	3	0	0

(B) Autosomal Recessive Model

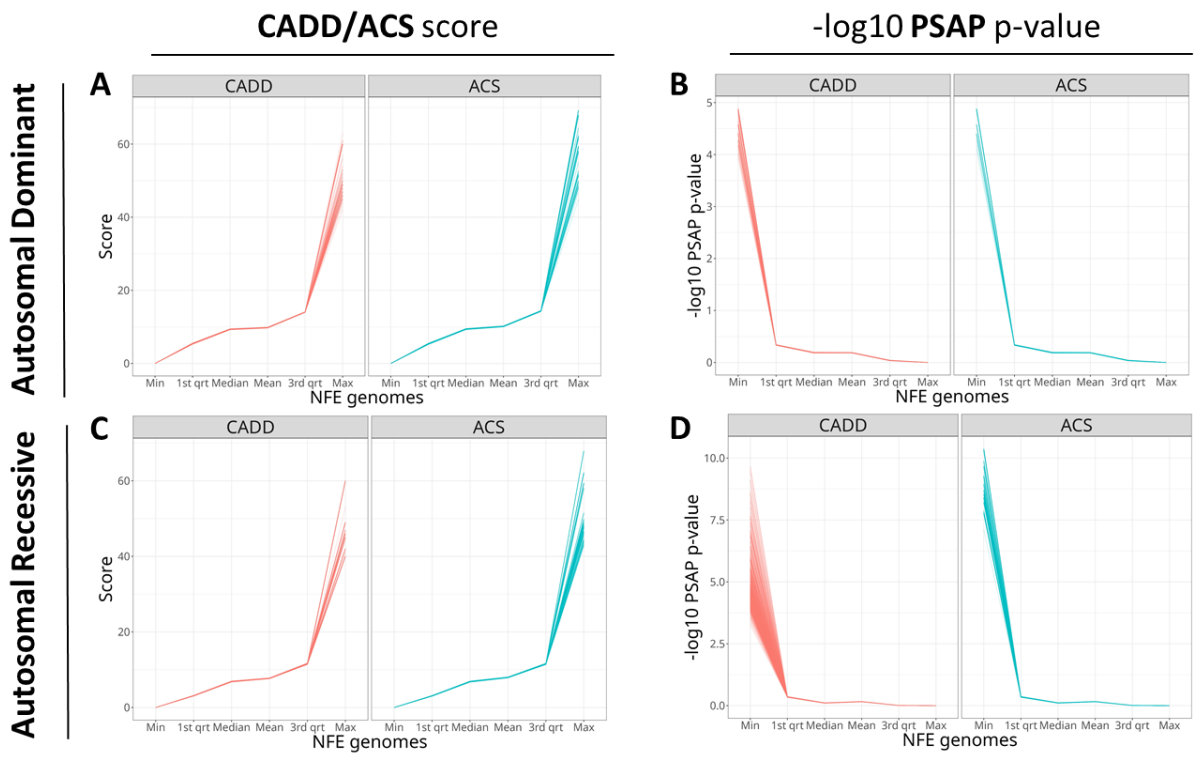
Type	VEP consequence	N variants	% top 10 - PSAP-genomic-regions-ACS	% top 10 - PSAP-genomic-regions-CADD
Splicing	splice acceptor variant	27	100	100
	splice donor 5th base variant	15	100	100
	splice donor region variant	10	100	90
	splice region variant	7	100	85.7
	splice donor variant	12	83.3	66.7
	splice polypyrimidine tract variant	1	0	100
Other	5 prime UTR variant	1	100	100
	downstream gene variant	8	62.5	50
	intron variant	12	41.7	33.3
	upstream gene variant	2	0	50
	start lost	1	0	0

Supp. Table S4: tops_noncoding_var_noncoding_Clinvar_allmodels_hg38_TGP_RAVAQ (.xlsx file)

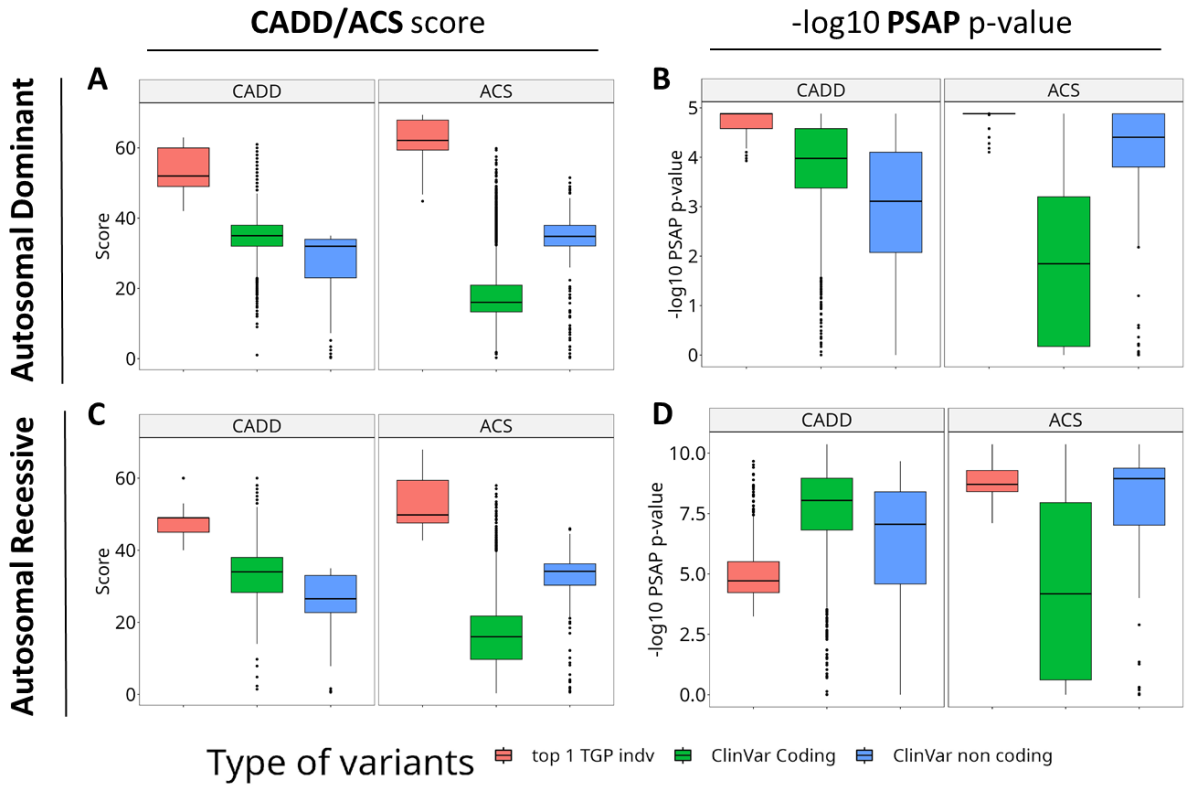
Supp. Table S5: tops_new_genomiser_1kg_QC_RAVAQ_hg38_TGP_allmodels (.xlsx file)

Supp. Table S6: Ranks of 6 known CSVD variants and 3 male infertility candidate variants with PSAP-genes-CADD and PSAP-genomic-regions-CADD (1 row per individual). Each CSVD variant was observed in a different individual. Each male infertility variant was observed in a different family consisting of three members each.

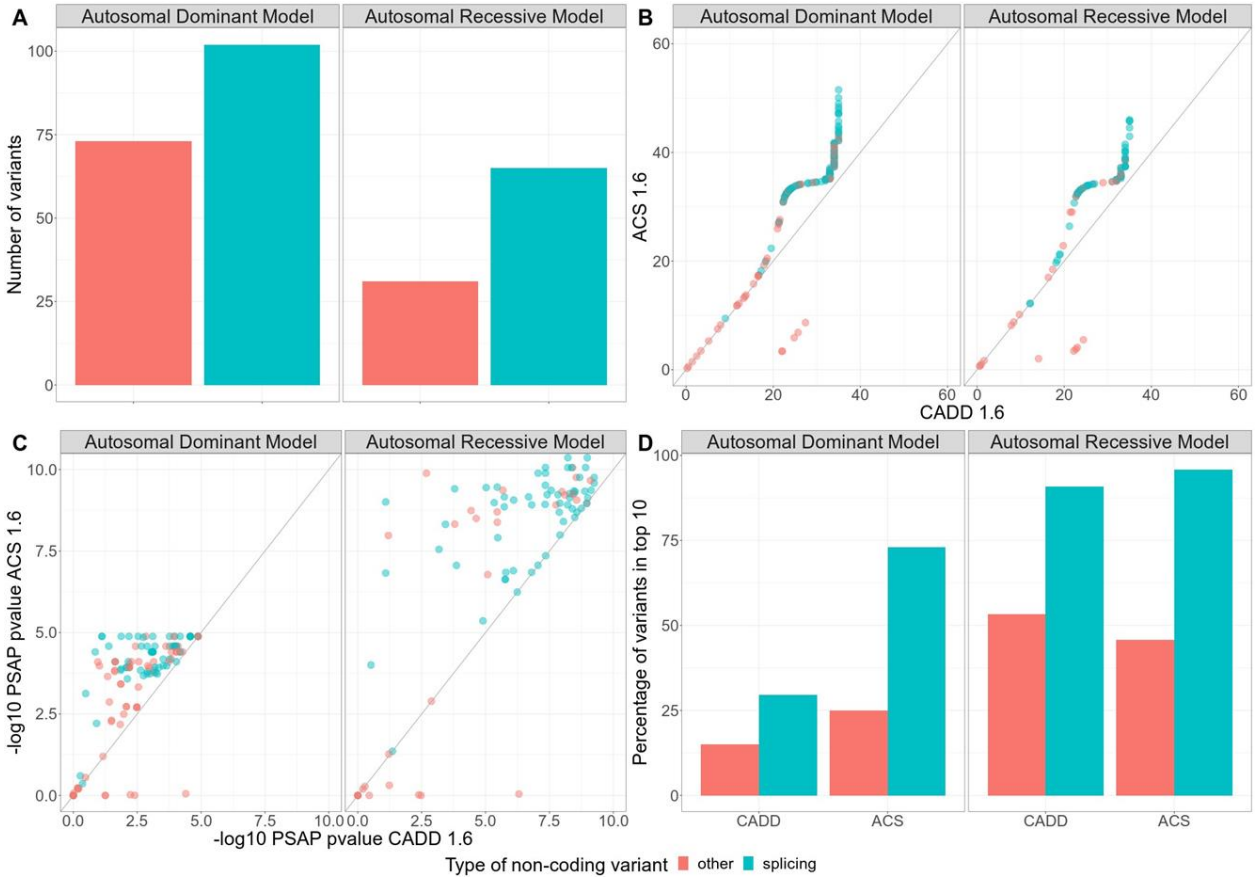
Disease	Gene	CADD region	Chr:Pos Ref/Alt	Rank PSAP – genes-CADD	Rank maximal CADD score on genes	Total number variants gene strategies	Rank PSAP-genomic-regions-CADD	Rank maximal CADD score on genomic regions	Rank PSAP-genomic-regions-ACS	Rank maximal ACS score on genomic regions	Total number variants genomic regions strategies
CSVD	HTRA1	R088914	10:124266285 T/G	69	117	6,639	38	120	4,054	3,784	16,243
CSVD	HTRA1	R088914	10:124266281 C/A	193	330	6,152	69	330	8,184	7,224	14,838
CSVD	HTRA1	R088914	10:124266885 G/A	4	32	6,435	7	32	3	10	15,807
CSVD	COL4A2	R105356	13:111132702 G/T	110	320	6,574	39	315	12,345	5,922	15,757
CSVD	NOTCH3	R127589	19:15303053 G/A	24	128	6,704	60	123	5,162	4,452	15,922
CSVD	NOTCH3	R127589	19:15303260 G/A	40	134	6,605	82	130	308	1,031	15,863
Male infertility	SPAG6	R083368	10:22389235 C>T	12	337	3,888	47	759	17,241	54,615	70,718
				12	310	3,823	44	736	16,446	53,849	70,241
				7	290	3,784	43	717	15,983	53,435	69,965
				8	133	3,904	21	135	34,392	19,539	70,803
Male infertility	TUBA3C	R100895	13:19177247 C>T	7	120	3,789	17	123	32,844	18,496	69,787
				13	122	4,087	37	128	34,538	19,522	70,803
Male infertility	CCDC9	R128563	19:47260609 C>T	12	110	3,959	51	114	7,922	7,202	70,917
				10	127	3,843	45	132	7,531	7,005	70,638
				15	115	3,880	88	121	8,383	7,371	71,357



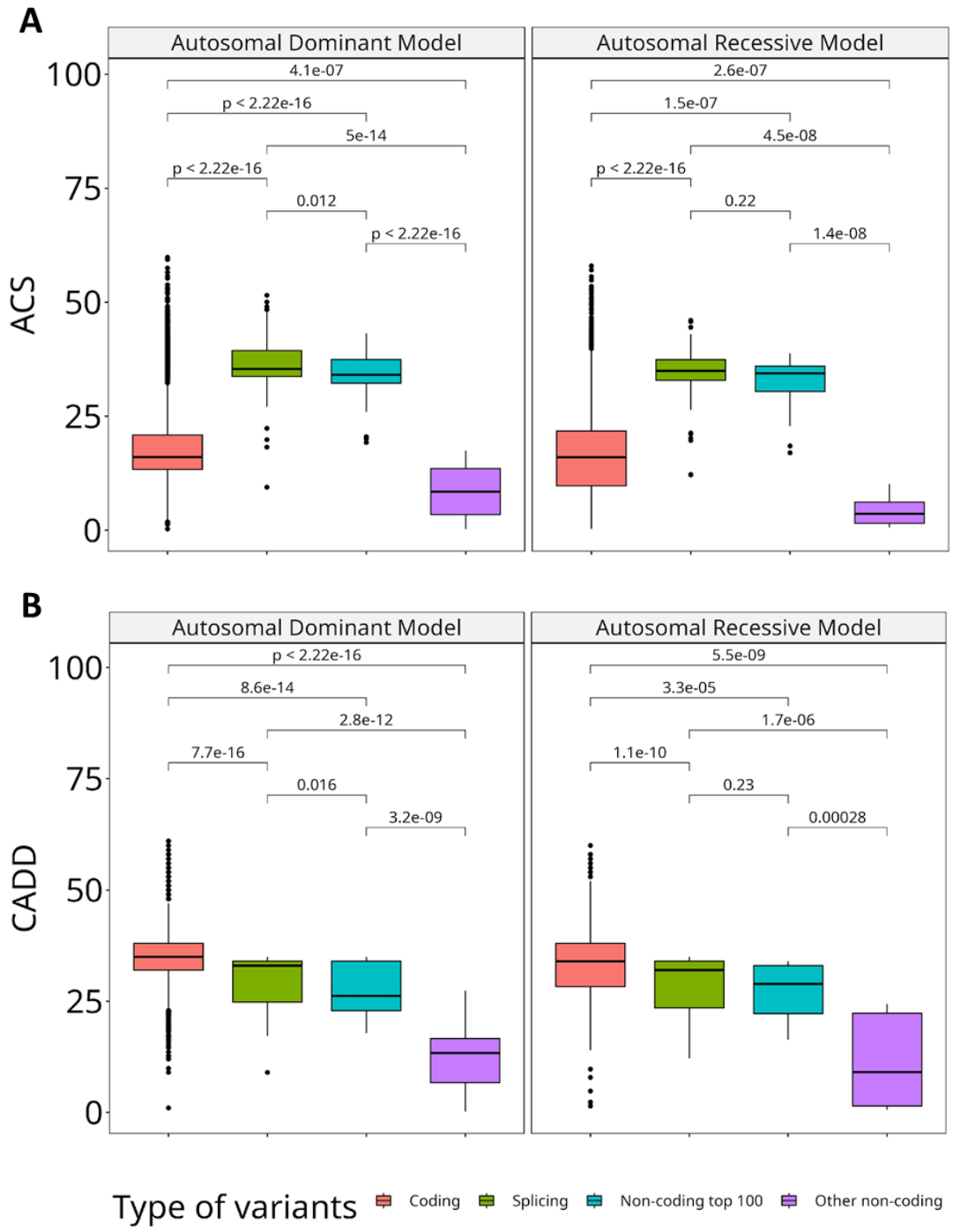
Supp. Fig.S1. Summary statistics of pathogenicity scores and PSAP p-values (scale –log 10) for NFE individuals (one line by individual)



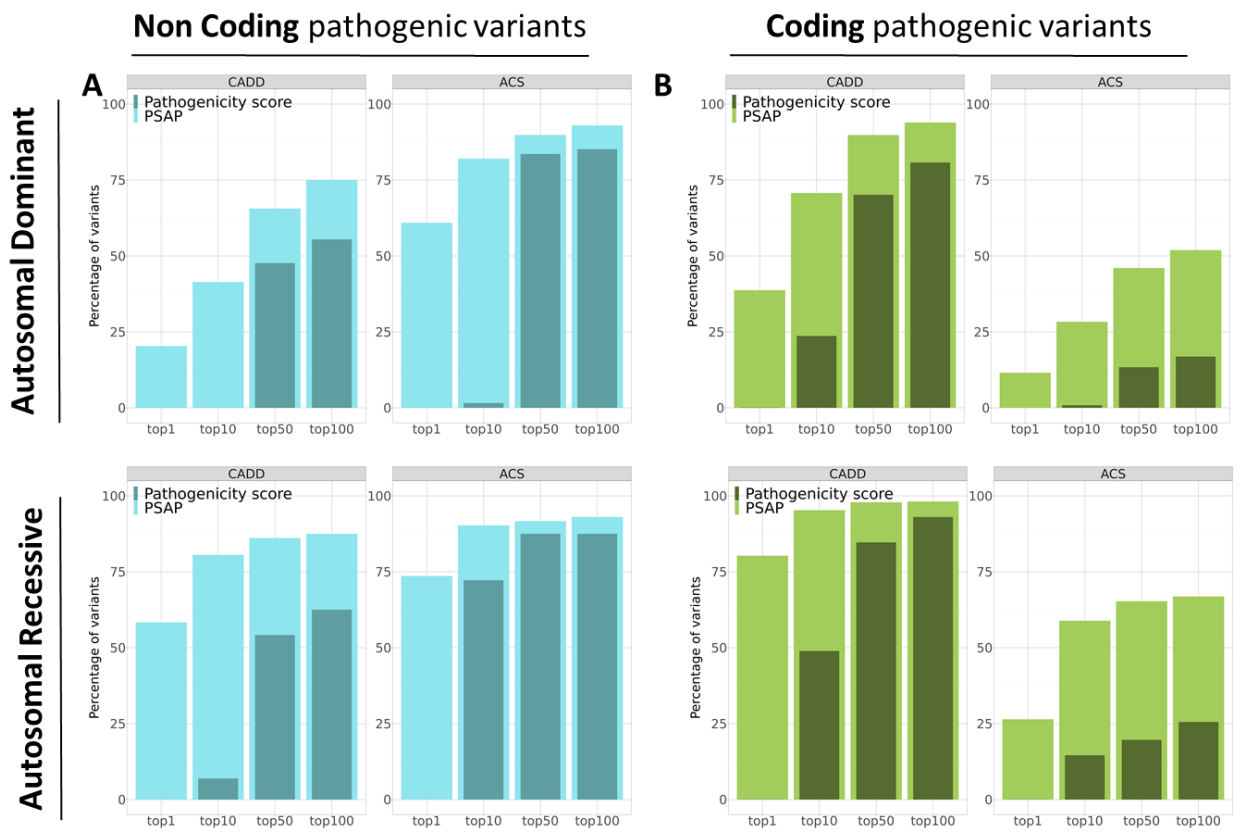
Supp. Fig.S2. Pathogenicity scores and PSAP p-values (scale –log 10) distributions for NFE individuals (maximal value for each genome), coding and non-coding ClinVar variants



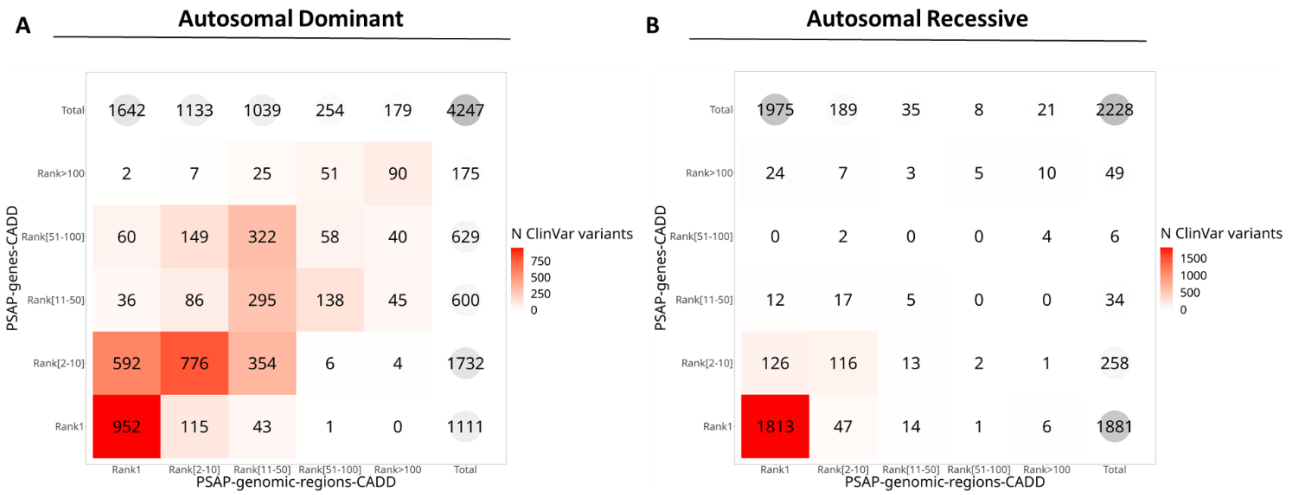
Supp. Fig.S3. Prioritization of splice variants versus other non-coding variants with PSAP on CADD regions with CADD or ACS. P-values at 0 were replaced by a p-value of 10-12, which is lower than all the other non-zero p-values, for visualization purposes.



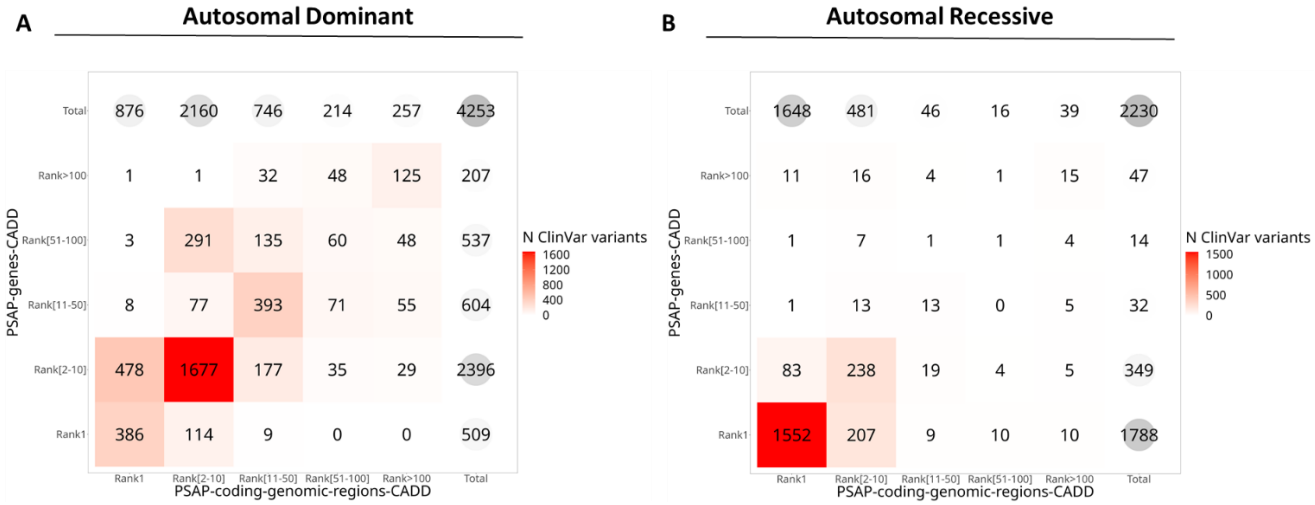
Supp. Fig.S4. (A) Distribution of CADD scores for ClinVar variants, by type of variant and mode of inheritance
(B) Distribution of ACS scores for ClinVar variants, by type of variant and mode of inheritance
Coding: N=4,253 variants AD model and 2,245 variants AR model, Splicing: 102 variants AD model and 65 variants AR model, Non-coding top 100: 49 variants AD model and 19 variants AR model, Other non-coding: 24 variants AD model and 12 variants AR model



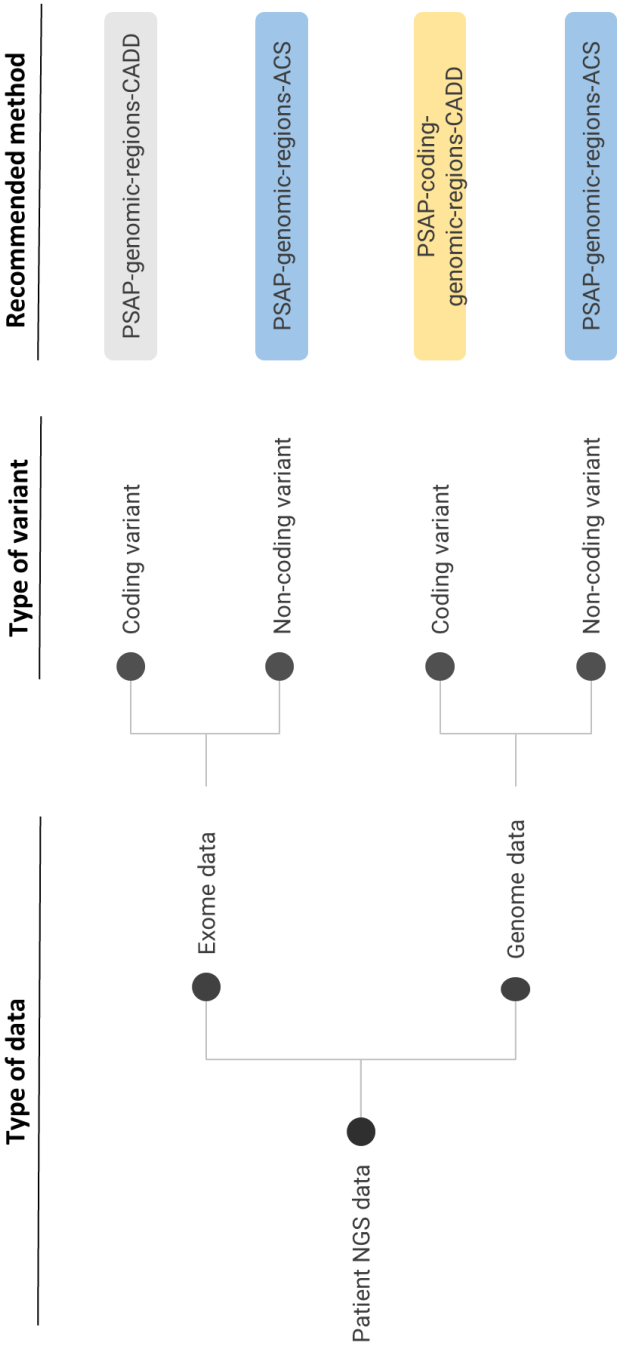
Supp. Fig.S5. Comparison of the PSAP-genomic-regions strategy versus a pathogenicity score alone in artificially-simulated disease exomes: percentage of pathogenic non-coding and coding ClinVar variants reaching the top N of variants in at least 90% of FREX individuals, with PSAP-genomic-regions (darker shade of blue or green) or the pathogenicity score alone (lighter shade of blue or green), CADD or ACS N = 48 non-coding AD variants and N = 64 non-coding AR variants (B) N = 4,965 coding AD variants and N = 2,680 coding AR variants



Supp. Fig.S6. Comparison of PSAP-genomic-regions-CADD and PSAP-genes-CADD for in artificially-simulated disease exomes: number of coding pathogenic ClinVar variants reaching the top N of variants in at least 90% of FREX individuals for each strategy



Supp. Fig.S7. Comparison of PSAP-coding-genomic-regions-CADD and PSAP-genes-CADD strategies for in artificially-simulated disease genomes: number of coding pathogenic ClinVar variants reaching the top N of variants in at least 90% of NFE individuals for each strategy



Supp. Fig.S8. Flowchart to choose the PSAP method of analysis depending on type of data and variants analyzed

Appendix II

Supplementary Materials

Easy-PSAP: an integrated workflow to prioritize pathogenic variants in sequence data from a single individual

Supplementary Materials

Easy-PSAP: an integrated workflow to prioritize pathogenic variants in sequence data from a single individual

Curation of a high-confidence list of ClinVar pathogenic variants with a known mode of inheritance

The pathogenic ClinVar (1) variants were downloaded from the NCBI website (https://www.ncbi.nlm.nih.gov/clinvar/, accessed on the 3rd of June 2022). Some of these ClinVar variants had an annotated mode of inheritance (moi) : "moi autosomal recessive" or "moi autosomal dominant".

From ClinVar, there were 12,776 variants annotated as autosomal dominant (AD) and 12,776 variants annotated as autosomal recessive (AR). Variants were filtered to keep only autosomal pathogenic SNVs having as review status either “reviewed by expert panel” or “criteria provided, multiple submitters, no conflicts”, which are the two best review status in ClinVar. There were 1,518 AD and 1,118 AR variants meeting these criteria.

For variants which did not have an annotated mode of inheritance, we used a curated version of the database OMIM, hOMIM (2) to retrieve a mode of inheritance, and kept variants that were always associated with an AD or AR mode of inheritance in hOMIM. The same filtering was applied, which left 3,641 additional variants for the AD and 1,706 for the AR model. In total, we had a set of 5,159 variants for the AD model and 2,824 variants for the AR model.

Among these ClinVar variants, 4,593 and 2,430 variants were coding SNVs respectively for the AD and AR models.

Evaluating the performance of PSAP: simulation of disease exomes using ClinVar variants

Easy-PSAP was evaluated through simulations by checking the prioritization of known pathogenic variants in the mutational background of a healthy individual. The initial and updated PSAP null distributions were compared in term of their performances to prioritize known pathogenic variants.

Disease exomes were simulated by introducing pathogenic ClinVar (1) variants within 574 exomes from healthy individuals of the French Exome Project (3). The ClinVar variants were annotated using the VEP software (4) and their PSAP p-values were computed for the set of null distributions under evaluation.

The same null distributions were applied to the exomes of FREX individuals using Easy-PSAP. The ClinVar variants were inserted one by one in each exome as followed: if the individual carries a variant in the region of the ClinVar variant, then the individual variant and its p-value were replaced by the ClinVar variant’s; else the ClinVar variant was added for the region. Each individual’s genes were then ranked by ascending PSAP p-value and the rank of the ClinVar variant were retrieved. The process was repeated for each ClinVar variant.

References

1. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4;46(Database issue):D1062–7.

2. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, et al. Natural selection on genes that underlie human disease susceptibility. Curr Biol. 2008 Jun 24;18(12):883–9.

3. Génin E, Redon R, Deleuze J, Campion D, Lambert J, Dartigues J, et al. The French Exome (FREX) Project: A Population-based panel of exomes to help filter out common local variants. Genetic Epidemiology. 2017;41(7):691–691.

4. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016 Jun 6;17(1):122.

Appendix III

Supplementary Materials

Easy-PSAP user guide

Supplementary Materials

Easy-PSAP: an integrated workflow to prioritize pathogenic variants in sequence data from a single individual

Introduction

Easy-PSAP is a **Snakemake** **workflow** [*Köster et al, 2012*] which allows the evaluation of genetic variants at the scale of a whole exome or genome. It is composed of two pipelines based on the Population Sampling Method (PSAP) method [*Wilfert et al, 2016*]. PSAP uses allele frequencies from large population databases to construct gene-based null distributions of pathogenicity scores [*Kircher et al, 2014*] and ultimately gives a p-value by gene for each individual, which summarizes how unlikely it is to observe a variant with such pathogenicity score in the general population in this gene.

The first workflow **snakemake_makedistrib_PSAP** allows the custom calculation of PSAP null distributions from allele frequencies data and a pathogenicity score (CADD score or other score). The second workflow **snakemake_apply_PSAP** is a new implementation of the initial PSAP pipeline which applies these null distributions to a vcf file of a patient or multiple patients and/or controls.

Easy-PSAP aims at making the PSAP method accessible and user-friendly for both clinicians and researchers. A set of PSAP null distributions with the latest database information in hg19 are readily available in the **/snakemake_apply_PSAP** directory, with global allele frequencies from the gnomAD V2 genome database [Karczewski et al 2020] and CADD v1.6 [Rentzsch et al 2021] as the pathogenicity score of variants. PSAP null distributions in hg38 are now available as well, calibrated using allele frequencies from the gnomAD V3 genome database [Chen et al. 2024] and CADD v1.6 in hg38. The option of using the hg19 or hg38 assembly of the human genome is offered to users of the pipeline, both for creating PSAP null distributions and applying them to their data.

Easy-PSAP is currently implemented with genes as units of testing, and the option of using CADD regions is also available which allows the analysis of the whole genome and not just its coding parts as described in [Ogloblinsky et al. 2024].

Requirements

Set up the conda environment

Easy-PSAP requires the conda package manager to function. If it is not installed already, please see bioconda installation instructions to set it up. Snakemake can then be installed through conda, as described in the Snakemake installation instructions.

Download Easy-PSAP repository from GitHub

Git clone command can be used to create a local copy of Easy-PSAP:

```
git clone https://github.com/msogloblinsky/Easy-PSAP.git
```

Download large files for to run snakemake_apply_PSAP

Files too large to be hosted on GitHub can be dowloaded via this link. They are separated in two folders `example` and `data` that can directly be added to the `/snakemake_apply_PSAP/` folder to obtain the following configuration:

```
├── snakemake_apply_PSAP
│   ├── example
│   ├── data
│   ├── config
│   ├── slurm
│   ├── optional
│   ├── src
│   ├── envs
│   ├── README.md
└── └── Snakefile
```

Specific instructions to run each of the two workflows can be found in their dedicated README.md file, in each of their folders.

References

- Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28, 2520–2522 (2012)
- Wilfert, A. B. et al. Genomewide significance testing of variation from single case exomes. Nat Genet 48, 1455–1461 (2016)
- Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46, 310–315 (2014)
- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020)
- Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Medicine 13, 31 (2021)
- Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature 625, 92–100 (2024).
- Ogloblinsky, M.-S. C. et al. PSAP-genomic-regions: a method leveraging population data to prioritize coding and non-coding variants in whole genome sequencing for rare disease diagnosis. 2024.02.13.580050 Preprint at https://doi.org/10.1101/2024.02.13.580050 (2024).

Snakemake workflow to make PSAP null distributions

Introduction

This first workflow **snakemake_makedistrib_PSAP** allows the easy calculation of PSAP null distributions according to allele frequencies from a vcf file and a pathogenicity score (here the CADD score). The input data is highly customizable and the use of Snakemake makes it possible to run the pipeline multiple times with different input parameters without running again the steps creating files that have already been created.

Usage

Once the conda environment with Snakemake installed is setup and activated, the user can modify the configuration file according to the desired input files and parameters, and run the pipeline.

1. Configure the pipeline

config.yaml file

snakemake_makedistrib_PSAP/config/config.yaml contains all pipeline parameters which are tuned by the user with example parameters for the GRCh37 assembly. **snakemake_makedistrib_PSAP/config/config.hg38.yaml** contains example parameters to run the pipeline for data in hg38. Available parameters are as follows:

Parameter	Description
snakemake_directory	Path to `/snakemake_makedistrib_PSAP` directory on the user's machine
genes	bed file for the coordinates of genes' coding regions. The file for GRCh37 is provided with the pipeline and a new one for GRCh38 can be created using BioMart
cadd_regions	bed file for the coordinates of CADD regions. The file for GRCh37 is provided with the pipeline or can be downloaded here
coding_cadd_regions	bed file for the coordinates of coding CADD regions. Created from the intersection of genes and CADD regions bed files
score_file	Score file used to calibrated PSAP null distributions. Format of a CADD file currently supported. File needs to be indexed
score_prefix	Prefix describing the score file used
score_max	Maximal possible value of the pathogenicity score (70 for CADD) used to calibrate null distributions
allele_frequencies	Allele frequencies file used to calibrated PSAP null distributions, in the format vcf.gz

Parameter	Description
coverage	bed file of regions well-covered regions in allele frequency database
af_prefix	Prefix describing the allele frequencies file used
outdir	Output directory
outfile	Prefix for output files, will be the name of the PSAP null distributions file
units	Unit of testing to construct PSAP null distributions, can take the values “gene”, “cadd_region” or “coding_cadd_region”
assembly	Both “GRCh37” and “GRCh38” options are currently supported
compound_heterozygote_model	TRUE or FALSE, if TRUE will calculate PSAP null distributions for the compound heterozygote (CHET) model
hemizygote_model	TRUE or FALSE, if TRUE will calculate PSAP null distributions for the hemizygote model (HEM PSAP null distributions precalculated for genes only)

WARNING: Users need to check the format of their input files. For the GRCh37 assembly, no “chr” prefix is expected before the name of the chromosome in the allele frequency file and in the score file. For the GRCh38 assembly, a “chr” prefix is expected before the name of the chromosome in the allele frequency file but not in the score file. These specifications come from the format of gnomAD files and CADD files. If not the case, the user needs to format their vcf file accordingly, as the pipeline expects this format from input files.

For the hemizygote model: units of testing need to be defined on chrX. This option will also calculate AD/AR/CHET models for females. The names of female/male allele frequencies columns in the allele frequencies file can be changed on line 97 of the Snakefile (currently “AF_Female” and “AF_Male”)

It can also be noted that the allele frequency input file can be split by chromosome, which needs to be specified accordingly in the config.yaml file (see **snakemake_makedistrib_PSAP/config/config.hg38.yaml** for an example).

For the coverage file, the command line used to get the bed file provided for gnomAD genome V2 (good coverage = 90% of individuals at dp10) was the following. A similar process was used for gnomAD V3 coverage file.

```
zcat gnomad.genomes.r2.0.1.coverage.txt.gz | tail -n+2 | awk '{print $1"\t"($2-1)"\t"$2"\t"$7}' > gnomad.genome.r2.0.1.dp10.bed
```

The workflow can also be run for all three units at the same time by replacing `units = config["units"]` in the Snakefile by: `units = ["gene","coding_caddregion","caddregion"]`

> TIP: different analyses can be run using just one cloned repository. This is achieved by changing the outdir and outfile in the configuration file. Also different parameters values can be used in the different analyses.

2. Create the Conda environments

The Snakemake relies on conda environements which contain all the necessary dependencies to run the pipeline. The conda environements need to be installed first, which can take some time. This step is easily done using the following command line:

```
snakemake --use-conda --conda-create-envs-only --conda-frontend conda -j 1
```

3. Run the pipeline.

Once the pipeline is configured and conda environments are created, the user just needs to run Snakemake pipeline to make PSAP null distributions:

```
snakemake --use-conda --conda-frontend conda -j 22
```

The mandatory arguments are: * **–use-conda**: to use the conda environments created at the previous step. The **–conda-frontend conda** specifies the use of conda to run the environments, instead of mamba. * **-j**: number of threads/jobs provided to snakemake. The pipeline splits each step by chromosome so running it on at least 22 threads is recommended.

An optional argument **–dry-run** can be used to make a dry run of the pipeline, check if there are no warnings and see all the files that will be created. With the additionnal argument **–configfile=/path_to_configfile/name_config.yaml**, the user can use a configuration file in a different location instead of modifying the default config.yaml file in the snakemake directory.

> TIP: If the execution of a step of the pipeline fails or if you want to see how it progresses, you can check the `{outdir}/log` directory where a log file is created for each rule of the Snakemake by chromosome if applicable. A `{outdir}/benchmark` directory is also created automatically to monitor how much memory and cpu is used by each step of the pipeline.

Pipeline steps

The steps of the pipeline are not meant to be used separately and will be run by Snakemake in the most efficient way depending on input and output files.

1. split_multiallelic_panel

The vcf file with allele frequencies from the reference database is split by chromosome and multi-allelic variants are split. Output located in `{outdir}/reference_panel`.

2. make_allele_frequencies_file

The allele frequency files created at the previous step are processed by default to keep only PASS SNVs, in well-covered regions of the database for the calculation of PSAP null distributions. If other filtering needs to be applied, or all filtering removed (no column `FILTER` in the input vcf), the following line can be modified in the Snakefile:

```
bcftools view -i 'FILTER=="PASS"' --regions-file {params.coverage} --types
 snps {input.file_allele_frequencies_split}
```

The vcf input file needs to have the columns `CHROM`, `POS`, `REF`, `ALT`, `CHROM`, `AF`, `AC` which will be kept in the output files. If any column has a different name or if another column needs to be used for allele frequencies, the following line in the Snakefile can be modified:

```
bcftools query -f'%CHROM\t%POS\t%REF\t%ALT\t%AF\t%AC\n'
```

Output located in `{outdir}/reference_panel`.

3. make_notpassvariants_file

Variants that do not pass the filtering descrbed in the previous step will have to be removed from the calculation of PSAP null distributions downstream of the pipeline. Changes can be made to the default filtering as needed. In this step, a file by chromosome is created with the coordinates of these variants. Output located in `{outdir}/reference_panel`.

4. list_units_wellcovered_panel

The bed files for genes, CADD regions and coding CADD regions are intersected with the bed file of good coverage in the allele frequencies database using bedmap. A list of genes, CADD regions and coding CADD regions well-covered in at least 50% of their length in the database is computed. Final PSAP null distributions will only include this list of units. Output located in `{outdir}/reference_panel`.

5. make_input_table_forPSAP

In this step, information used to calculate PSAP null distributions is gathered in tables by chromosome using bedtools. These tables include the chromosome, start, end, reference allele, alternative allele, raw CADD score, PHRED CADD score, corresponding CADD region and gene if there is one of all possible SNVs of the genome in the well-covered regions by database. Output located in `{outdir}/score_tables`.

6. calculate_null_distributions_PSAP

This step carries out the main function of the pipeline, which is to calculate PSAP null distributions. Overall, it evaluates the probability of seeing a heterozygote or homozygote variant with such a maximal CADD score as high or higher in the chosen unit of testing (gene, CADD regions or coding CADD region) in the reference database. Variants from the step `make_notpassvariants_file` are removed from the analysis. The PHRED CADD score is used (column 7). One file by chromosome and by model is created, each line corresponds to a unit of testing and columns to PSAP probabilities. Output located in `{outdir}/temp_lookup_bychr`.

7. make_final_null_distributions_PSAP

The null distributions by chromosome are gathered in a unique table for the autosomal dominant (het) model and autosomal recessive (hom model), for the units of testing well-covered in at least 50% of their length in the database. The final two output files created are `{outfile}_het.txt.gz` and `{outfile}_hom.txt.gz`. Output located in `{outdir}/final_lookuptables_PSAP`.

Configuration of computation resources

This Snakemake pipeline is portable to cluster engines, see the Snakemake cluster documentation for more detailed information. Taking the SLURM scheduler as an example, the pipeline can be run without any changes using the following command line:

```
snakemake --use-conda --slurm --default-resources slurm_account=<your SLUR
M account> slurm_partition=<your SLURM partition>
```

For a custom use of ressources and input parameters, a **profile** can be used. When the argument `--profile` is added, snakemake looks for a directory with the name of the given profile (here **slurm**) containing a `config.yaml` file. This file contains the parameter `cluster` which tells snakemake how to submit jobs to the cluster. In the case of slurm, the `sbatch` command is used with its arguments. Other arguments include `jobs` which specifies the maximum number of jobs submitted at the same time, `default-resources` requested for each job and `resources`, which define the resource limits. Where to save SLURM logs and what to call them is also specified. Note that this folder must already exist. All of these arguments can be modified in the file `slurm/config.yaml` by the user depending on the desired ressources to allocate to the pipeline. To run the pipeline according to a specific profile, the following command line is used:

```
snakemake --use-conda --profile slurm
```

Snakemake can also use generic cluster support like `qsub`, by giving an other argument to `--cluster` combined with ressources. The profile file can also be altered accordingly depending on user needs. The most simple command line in that setting is:

```
snakemake --use-conda --cluster qsub --jobs 22
```

Test data

This pipeline has been tested using allele frequencies from the genomes of the gnomAD database v2 and CADD v1.6 both in GRCh37, with the default parameters of the pipeline. Bed files for genes and CADD regions are available with the pipeline as well. Both of these files are in open access, and can be used to test the pipeline and its functionalities.

Snakemake workflow to apply PSAP null distributions

Introduction

This second workflow **snakemake_apply_PSAP** applies these null distributions to a vcf file of a patient or multiple patients or controls. The input data is highly customizable and the use of Snakemake makes it possible to run the pipeline multiple times with different input parameters without running again the steps creating files that have already been created.

Usage

Once the conda environmenet with Snakemake installed is setup and activated, the user can modify the configuration file according to the desired input files and parameters, and run the pipeline. The VEP software is used for the vcf file annotation, without the need to install VEP on the machine thanks to the conda package manager. The necessary files for annotation are (cf next paragraph): * VEP cache and FASTA (for the version 107 of VEP, which is the one currently used in the pipeline) * CADD score files for SNVs and InDels

1. Configure the pipeline

config.yaml file

snakemake_apply_PSAP/config/config.yaml contains all pipeline parameters which are tuned by the user with example parameters for the GRCh37 assembly. **snakemake_apply_PSAP/config/config.hg38.yaml** contains example parameters to run the pipeline for data in GRCh38. Available parameters are as follows:

Parameter	Description
snakemake_directory	Path to `/snakemake_apply_PSAP` directory on the user's machine
vcf	vcf file with individual data to score with PSAP, needs to be in format GRCh37 (no "chr" prefix for chromosomes) and bgzipped
ped	Corresponding PED file for the vcf file, tab-delimited, important columns are 2nd column = individual IDs, 5th column = sex (1=male, 2=female, other=unknown) and 6th column = status (1=unaffected, 2=affected)
coverage	bed file of regions well-covered regions in allele frequency database
lookup_namefile	Path and prefix of the lookup table for PSAP null distributions (lead to two files ending by "_het.txt.gz" and "_hom.txt.gz")
variants_exclude	List of variants to exclude from the vcf file, that were excluded from the calculation of

Parameter	Description
	null distributions. The file for gnomAD V2 is provided with the pipeline
cadd_path	Directory where the CADD files for annotation are located
score_max	Maximal possible value of the pathogenicity score (70 for CADD) used to calibrate null distributions
genes	bed file for the coordinates of genes coding regions. The file for GRCh37 is provided with the pipeline and a new one can be created using BioMart
cadd_regions	bed file for the coordinates of CADD regions. The file for GRCh37 is provided with the pipeline or can be downloaded here
vep_cache	Directory of the VEP cache. Instructions on how to download the cache can be found on the Ensembl website with the necessary files located here for VEP 107
vep_cache_merged	"TRUE" if VEP cache is merged, otherwise "FALSE"
vep_fasta	VEP FASTA file with its path which can be downloaded here for VEP 107
outdir	Output directory
outfile	Prefix for output files, will be the name of the PSAP null distributions file
unit	Unit of testing to construct PSAP null distributions, can take the values "gene", "cadd_region" or "coding_cadd_region"
cadd_version	Version of CADD to be used for annotation
assembly	Both "GRCh37" and "GRCh38" options are currently supported
compound_heterozygote_model	TRUE or FALSE, if TRUE will calculate PSAP p-values for the compound heterozygote (CHET) model
hemizygote_model	TRUE or FALSE, if TRUE will calculate PSAP p-values for the hemizygote model for males depending on the sex indicated in the ped file (HEM PSAP null distributions pre-calculated for genes only)
indel_file File with InDels annotated by CADD website or "NA". Column names of	

Parameter	Description
InDel file must be “#Chrom Pos Ref Alt RawScore PHRED” (format of CADD 1.6 InDel file)	

For the hemizygote model: this option will calculate hemizygote model for males and AD/AR/CHET models for females on chrX.

TIP: the user running the pipeline needs to have permission to write in the folder where the input vcf file is located. If not, the vcf file needs to already be indexed as the first step of the pipeline involves indexing the input vcf file. Suggested command for indexing a vcf file, using bcftools :

```
bcftools index {vcf}
```

WARNING: The input vcf file for the PSAP pipeline must only contain biallelic variants (multiallelic variants need to be split in different lines). This process can be done using a tool like VCFprocessor or during a Quality Control of the vcf file using a tool like the R package RAVAQ. We provide a script optional/qc_vcf.R to perform the recommended QC anf formatting for the vcf file. The script can be run using the command:

```
Rscript qc_vcf.R {vcf} {ped} {outfile} {outdir}
```

Users also need to check the format of the vcf file. For the GRCh37 assembly, no “chr” prefix is expected before the name of the chromosome. For the GRCh38 assembly, a “chr” prefix is expected before the name of the chromosome. If not the case, the user needs to format their vcf file accordingly, as the pipeline expects this format for annotation purposes.

For the coverage file, the command line used to get the bed file provided for gnomAD genome V2 (good coverage = 90% of individuals at dp10) was:

```
zcat gnomad.genomes.r2.0.1.coverage.txt.gz | tail -n+2 | awk '{print $1"\t"($2-1)"\t"$2"\t"$7}' > gnomad.genome.r2.0.1.dp10.bed
```

The chosen lookup table for PSAP null distributions has to match the “unit” and “cadd_version” specified in the configuration file. Already calculated lookup tables include : *latest_gnomadgen_string_ensembl_cadd1.6_af_nosing_lookup : genes* latest_gnomadgen_string_coding_caddregions_cadd1.6_af_nosing_lookup : coding CADD regions *latest_gnomadgen_string_caddregions_cadd1.6_af_nosing_lookup : CADD regions Available PSAP null distributions have used gnomAD V2 genome for allele frequencies calibration (AF specifically) and CADD v1.6 for the pathogenicity score, all in GRCh37.

For the CADD files, the path and version need to lead to the following files:

```
{cadd_path}/CADD_v{cadd_version}/whole_genome_SNVs.tsv.gz
{cadd_path}/CADD_v{cadd_version}/InDels.tsv.gz
```

The formats of the CADD files are currently supported. Files need to be indexed.

The file with InDels anotated by CADD is optional. If no file is provided, InDels will not be scored with PSAP and only SNVs will be kept for the analysis. The script optional/make_indel_file_forcadd.sh can be used to generate input file with InDels to

upload to the CADD website in the correct format. The script can be run using the following command (bcftools and tabix need to be installed in the environement for this command to work):

```
make_indel_file_forcadd.sh {vcf} {outfile} {outdir}
```

> TIP: different analyses can be run using the same repository. This is achieved by changing the outdir and outfile in the configuration file. Also different parameters values can be used in the different analyses.

2. Create the Conda environments

The Snakemake relies on conda environements which contain all the necessary dependencies to run the pipeline. The conda environements need to be installed first, which can take some time. This step is easily done using the following command line:

```
snakemake --use-conda --conda-create-envs-only --conda-frontend conda
```

3. Run the pipeline.

Once the pipeline is configured and conda environments are created, the user just needs to run the Snakemake pipeline to score the vcf file with PSAP:

```
snakemake --use-conda --conda-frontend conda -j 20
```

The mandatory arguments are: * **–use-conda**: to use the conda environments created at the previous step. The **–conda-frontend conda** specifies the use of conda to run the environments, instead of mamba. * **-j**: number of threads/jobs provided to snakemake. The VEP annotation uses 20 threads if provided (not more), and the number of threads will condition the number of individuals analyzed at the same time.

An optional argument **–dry-run** can be used to make a dry run of the pipeline, check if there are no warnings and see all the files that will be created. With the additionnal argument **–configfile=/path_to_configfile/name_config.yaml**, the user can use a configuration file in a different location instead of modifying the default config.yaml file in the snakemake directory.

> TIP: If the execution of a step of the pipeline fails or if you want to see how it progresses, you can check the `{outdir}/log` directory where a log file is created for each rule of the Snakemake by individual if applicable. A `{outdir}/benchmark` directory is also created automatically to monitor how much memory and cpu is used by each step of the pipeline.

Pipeline steps

The steps of the pipeline are not meant to be used separately and will be run by Snakemake in the most efficient way depending on input and output files.

1. filter_regions

The vcf file is filtered to exclude regions not well-covered in the database used to construct PSAP null distribution. This step ensures the reliability of PSAP results, so that variants analyzed in the vcf were also present in the construction of PSAP null distributions. Output located in `{outdir}`.

2. write_column_names

Column names from the vcf file are written in a separate file. This file will then be used during PSAP annotation to check fields present in input files. Output located in `{outdir}`.

3. vep_annotation

During this step, the vcf file is annotated by the VEP software. Multiple annotations critical for downstream PSAP analysis are added, among which: the corresponding gene and CADD region, if applicable, and the CADD score of variants. VEP cache and FASTA file need to be downloaded prior to the analysis, but the software itself is run through the conda environment. Annotation is slightly different depending on the unit of testing. Output located in `{outdir/annotated}`.

4. filter_variants

Variants that were filtered out in the database used to calculate PSAP null distributions are filtered out from the VEP output file in this step. Again, this ensures the reliability of PSAP results. Output located in `{outdir/annotated}`.

5. apply_PSAP_calculations

This step carries out the main function of the pipepline, which is the calculation of PSAP p-values for variants in the vcf input file. These calculations are run by individual. Preprocessing steps harmonize the data between the vcf file and the VEP annotated file using the gaston R package. If an InDel CADD score file is provided, InDels are included in the analysis. Otherwise only autosomal SNVs are kept. Then, the variant with the maximal CADD score by unit of testing (gene, CADD region or coding CADD region) is scored with the PSAP null distribtion for the autosomal dominant (heterozygote variants) or recessive model (homozygote variants). Additional models will be run if indicated in the configuration file, and can include the compound heterozygote and/or hemizygote models. Output located in `{outdir/annotated}`.

6. make_report_file

If affected individuals are present in the ped file (value 2 in the 6th column), this step makes a report file that merges the individual PSAP records. If a variant is present in multiple individuals, it will be a single record in the report file. If there are control individuals in the vcf (value 1 in the 6th column of the ped file), a validation step is carried out. If the variant of the affected individual(s) is not seen in controls, the validation column has an “no_controls” value, otherwise it has the value “violation”. Output located in `{outdir}/annotated}`.

7. compress_output_files

Output files from PSAP are bgzipped to minimize the space taken by PSAP results, especially if a large number of individuals was analyzed. Report file is not compressed. Output located in `{outdir/annotated}`.

Configuration of computation resources

This Snakemake pipeline is portable to cluster engines, see the Snakemake cluster documentation for more detailed information. Taking the SLURM scheduler as an example, the pipeline can be run without any changes using the following command line:

```
snakemake --use-conda --slurm --default-resources slurm_account=<your SLUR
M account> slurm_partition=<your SLURM partition>
```

For a custom use of ressources and input parameters, a **profile** can be used. When the argument `--profile` is added, snakemake looks for a directory with the name of the given profile (here **slurm**) containing a `config.yaml` file. This file contains the parameter `cluster` which tells snakemake how to submit jobs to the cluster. In the case of slurm, the `sbatch` command is used with its arguments. Other arguments include `jobs` which specifies the maximum number of jobs submitted at the same time, `default-resources` requested for each job and `resources`, which define the resource limits. Where to save SLURM logs and what to call them is also specified. Note that this folder must already exist. All of these arguments can be modified in the file `slurm/config.yaml` by the user depending on the desired ressources to allocate to the pipeline. To run the pipeline according to a specific profile, the following command line is used:

```
snakemake --use-conda --profile slurm
```

Snakemake can also use generic cluster support like `qsub`, by giving an other argument to `--cluster` combined with ressources. The profile file can also be altered accordingly depending on user needs. The most simple command line in that setting is:

```
snakemake --use-conda --cluster qsub --jobs 20
```

Test data

A test vcf file, ped file and InDel file are available in the `/example` folder. They correspond to the names of the files in the configuration file (path need to be altered). They can be used as an example to format the user's files or to test run the pipeline. The vcf file has been generated by sampling 10 individuals of European descent from the 1000 Genomes Project, restricting to regions covered by the exome, and inserting known best-reviewed coding pathogenic variants (one in each individual exome) from the ClinVar database. The list of inserted variants and their genotypes can be found in the file `added_pathogenic_mutations_AD_AR_1kG.txt`. Running PSAP on these synthetic exomes shows how the method performs in a real-life setting.

Appendix IV

Supplementary Materials

Computational methods to detect digenism in sequencing data: a comprehensive review and benchmark

Supplementary Materials

Computational methods to detect digenism in sequencing data: a comprehensive review and benchmark

Marie-Sophie C. Ogloblinsky[1,*], The FrEx Consortium, Donald F. Conrad[2], Emmanuelle Génin[1,3,¶], Gaëlle Marenne[1,*,¶]

[1]Univ Brest, Inserm, EFS, UMR 1078, GGB, Brest, France

[2]Division of Genetics, Oregon National Primate Research Center, Oregon Health & Science University, Portland, Oregon, United States of America

[3]Centre Hospitalier Régional Universitaire de Brest, Brest, France

*Corresponding authors:

Email: marie-sophie.ogloblinsky@inserm.fr (M-S.O.); gaelle.marenne@inserm.fr (G.M.)

¶These authors contributed equally to this work

Feature	Level	VarCoPP 2.0	DIEP	DiGePred	VarCoPP
• Flexibility • Hydrophobicity	Variant level				×
• CADD score		×			×
• Recessiveness	Gene level		×		×
• Haploinsufficiency		×	×	×	×
• Evolutionary age • Essentiality • Number of pathways • Number of phenotypes • Network neighbors • Number co-expressed				×	
• Intolerance to loss-of-function mutations			×	×	
• Selection pressure (dN/dS ratio)		×		×	
• Inheritance Specific Pathogenicity Predictor		×			
• PPI distance	Gene pair level	×	×	×	×
• Phenotype similarity			×	×	
• Coexpression		×	×	×	
• Pathway similarity • Pathway distance • Literature distance • Gene-focused network and functional score				×	
• Knowledge Graph distance		×			
• Biological distance • Semantic similarity of gene GO annotations • Weight of the gene function relationship • Protein functional interaction effects • Tissue co-expression			×		

Supplementary Table S1: Feature of the RF methods to detect DI

Scenario	Positive pairs	Negative pairs
True digenic pairs held out from ARBOCK vs combination of genes from these true digenic pairs	• ARBOCK_held_out (N = 14)	• ARBOCK_shuffled (N=364)
All true OLIDA digenic pairs vs neutral pairs from FREX	• OLIDA_total (N = 69)	• random100_FREX (N = 32,953) • top100PSAP_FREX (N = 18,348)
ClinVar and FREX variants pairs	-	• random100_FREX_clinvar100_AD (N = 70) • top100PSAP_FREX_clinvar100_AD (N = 69) • random100_FREX_clinvar100_AR (N = 74) • top100PSAP_FREX_clinvar100_AR (N = 71)

Supplementary Table S2: Benchmarking scenarios for ML methods to detect DI

	ARBOCK_shuffled – negative pairs		**ARBOCK_held_out – positive pairs**	
	True negative	False positive	True positive	False negative
DiGePred	356	8	4	10
DIEP	319	45	8	6
VarCoPP2.0	306	58	2	12
ARBOCK.excl.pheno	128	236	11	3
ARBOCK.incl.pheno	155	209	9	5

Supplementary Table S3: Categorization of ARBOCK_held_out and ARBOCK_shuffled pairs in the benchmark

	random100_FREX – negative pairs		**top100PSAP_FREX – negative pairs**		**OLIDA_total – positive pairs**	
	True negative	False positive	True negative	False positive	True positive	False negative
DiGePred	32,937	16	18,341	7	44	25
DIEP	32,125	828	18,026	322	59	10
VarCoPP2.0	32,943	10	17,908	440	34	35
ARBOCK.excl.pheno	22,897	10,056	12,776	5,572	60	9
ARBOCK.incl.pheno	30,428	2,525	17,248	1,100	64	5

Supplementary Table S4: Categorization of random100_FREX and top100PSAP_FREX and pairs OLIDA_total in the benchmark

	random100_FREX_clinvar100_AD – negative pairs		random100_FREX_clinvar100_AR – negative pairs	
	True negative	False positive	True negative	False positive
DiGePred	70	0	74	0
DIEP	66	4	70	4
VarCoPP2.0	68	2	73	1
ARBOCK.excl.pheno	38	32	52	22
ARBOCK.incl.pheno	56	14	65	9

Supplementary Table S5: Categorization of random100_FREX_clinvar100_AD and random100_FREX_clinvar100_AR in the benchmark

	top100PSAP_FREX_clinvar100_AD – negative pairs		top100PSAP_FREX_clinvar100_AR – negative pairs	
	True negative	False positive	True negative	False positive
DiGePred	69	0	71	0
DIEP	63	6	70	1
VarCoPP2.0	63	6	64	7
ARBOCK.excl.pheno	33	36	51	20
ARBOCK.incl.pheno	50	19	61	10

Supplementary Table S6: Categorization of top100PSAP _FREX_clinvar100_AD and top100PSAP _FREX_clinvar100_AR in the benchmark

COLLEGE SCIENCES
DOCTORAL DE LA VIE
BRETAGNE ET DE LA SANTE

Université de Bretagne Occidentale

Titre : Stratégies statistiques exploitant les données de la population générale pour aider au diagnostic des maladies rares

Mots clés : Maladies rares, hétérogénéité génétique, génome non-codant, digénisme

Résumé : La forte hétérogénéité génétique et les modes de transmission complexes des maladies rares posent le défi d'identifier le variant causal si un seul patient le porte, en utilisant des données de séquençage et des méthodes d'analyse standard. Pour aborder ce problème, la méthode PSAP utilise des distributions nulles par gène de scores de pathogénicité CADD pour évaluer la probabilité d'observer un génotype donné dans la population générale. L'objectif de ce travail était de répondre au manque de diagnostic des maladies rares grâce à des méthodes statistiques. Nous proposons PSAP-genomic-regions, une extension de la méthode PSAP au génome non codant, en utilisant comme unités de test des régions prédéfinies reflétant la contrainte fonctionnelle à l'échelle du génome entier. Nous avons implémenté PSAP-genomic-regions et sa version initiale PSAP-genes dans Easy-PSAP, un workflow Snakemake intuitif et adaptable, accessible aussi bien aux chercheurs qu'aux cliniciens. Appliqué à des familles touchées par de l'infertilité masculine, Easy-PSAP a permis la priorisation de variants candidats pertinents dans des gènes connus et nouveaux. Nous nous sommes ensuite concentrés sur le digénisme, le mode le plus simple de transmission complexe, qui implique l'altération simultanée de deux gènes pour développer une maladie. Nous avons décrit et évalué les méthodes actuelles publiées dans la littérature pour détecter le digénisme et proposé de nouvelles stratégies pour améliorer le diagnostic de ce mode de transmission complexe.

Title : Statistical strategies leveraging population data to help with the diagnosis of rare diseases

Keywords : Rare diseases, genetic heterogeneity, non-coding genome, digenism

Abstract : High genetic heterogeneity and complex modes of inheritance in rare diseases pose the challenge of identifying an n-of-one patient's causal variant using sequencing data and standard analysis methods. To tackle this issue, the PSAP method uses gene-specific null distributions of CADD pathogenicity scores to assess the probability of observing a given genotype in a healthy population. The goal of this work was to address rare disease lack of diagnosis through statistical strategies. We propose PSAP-genomic-regions an extension of the PSAP method to the non-coding genome, using as testing units predefined regions reflecting functional constraint at the scale of the whole genome. We implemented PSAP-genomic-regions and the initial PSAP-genes in Easy-PSAP a user-friendly and versatile Snakemake workflow, accessible to both researchers and clinicians. When applied to families affected by male infertility, Easy-PSAP allowed the prioritization of relevant candidate variants in known and novel genes. We then focused on digenism, the most simple mode of complex inheritance, which implicates the simultaneous alteration of two genes to develop a disease. We reviewed and benchmarked current methods in the literature to detect digenism and put forward new strategies to improve the diagnostic of this complex mode of inheritance.