# Development of gene-prioritising methods using statistical genetics and clinical annotation for rare genetic disorders

Antoine Favier



Université Paris Cité

*Learning Planet Institute - École doctorale Frontières de l'Innovation en Recherche et Éducation (474)*

IHU Imagine – Institut des maladies génétiques – Clinical Bioinformatics Lab

# Development of gene-prioritising methods using statistical genetics and clinical annotation for rare genetic disorders

Par Antoine FAVIER

Thèse de doctorat de Génétique, omiques, bioinformatique et biologie des systèmes

Thèse menée sous la direction du Dr Antonio RAUSELL

Thèse présentée et soutenue publiquement le 15/12/2022 devant un jury composé de :

| | |
|---|---|
| Pr Juan Antonio GARCIA-RANEA | Rapporteur – Université de Málaga |
| Dr Emmanuelle GÉNIN | Rapportrice – Université de Bretagne Occidentale |
| Dr Aurélie COBAT | Examinatrice – Université Paris Cité |
| Dr David-Alexandre TRÉGOUËT | Examinateur – Université de Bordeaux |
| Dr Antonio RAUSELL | Directeur de thèse – Université Paris Cité |
| Dr Loredana MARTIGNETTI | Membre invité – Sorbonne Université |

# Titre : Développement de méthodes de priorisation de gènes pour les maladies génétiques rares grâce à un test paramétrique de statistique génétique et à l'annotation clinique

## Résumé :

À ce jour, près de 70% des patients atteints de maladies mendéliennes demeurent sans diagnostic après séquençage de leur ADN. Il est nécessaire d'étudier les causes génétiques de ces maladies à l'aide des nouveaux outils génomiques et de bio-informatique pour mettre en place de potentielles stratégies thérapeutiques. Les nouvelles méthodes de séquençage d'exome et de génome ont grandement amélioré la précision des études cliniques sur les maladies rares. La nouvelle ère de la génomique a permis une meilleure compréhension du génome humain et en particulier des variants génétiques associés pour un grand nombre de maladies rares et communes, ouvrant la voie à la médecine de précision. Cependant, le diagnostic et l'élaboration de stratégies thérapeutiques demeurent extrêmement compliqués du fait de l'hétérogénéité clinique et génétique, des défis statistiques associés et de la complexité de l'architecture génétique des maladies. L'ensemble des mécanismes génétiques et des artéfacts techniques peuvent brouiller les signaux statistiques, rendant le diagnostic et la recherche de thérapies très compliqués. Les méthodes de priorisation de gènes sont une solution pour simplifier ce problème.

Au cours de cette thèse, j'ai développé une première stratégie consistant à agréger plusieurs variants d'intérêt au sein d'une région spécifique et à évaluer l'importance de leur accumulation dans une cohorte de patients par rapport à une cohorte contrôle, plutôt que de tester chaque variant individuellement. Néanmoins, les individus contrôle sont rarement séquencés conjointement aux patients et cela peut conduire à des biais d'analyse. Pour contrer cet effet, j'ai développé une stratégie de test statistique de type "burden" sans contrôle, en utilisant les données publiques de Genome Aggregation Database (gnomAD) comme paramètre. L'hypothèse de ma stratégie a été testée sur les données du projet 1000 Génomes et appliquées dans le cadre clinique d'une cohorte de patients souffrant de ciliopathies.

La seconde stratégie développée a été d'utiliser les données cliniques renseignées par les médecins dans les dossiers médicaux pour prioriser les gènes et gagner en puissance statistique lors de l'évaluation de leur association au génotype. Des analyses guidées par le phénotype grâce aux ontologies phénotypiques pour définir de nouveaux diagnostics dans les maladies du développement ont déjà été menées et ont montré leur efficacité. J'ai travaillé sur la fiabilité des méthodes de

similarité sémantique utilisant les termes de l'ontologie dite « HPO » (de l'anglais, Human Phenotype Ontology) pour construire des groupes de patients cliniquement similaires grâce à ses similarités sémantiques afin de prioriser les variants génétiques. Ceci a été étudié sur des exomes du projet Deciphering Developmental Disorders (DDD).

Les résultats de cette thèse montrent d'une part qu'une stratégie de type "burden" peut fonctionner dans un cadre clinique et identifier des variants causaux sans *a priori* dans une cohorte hétérogène. Le test implémenté, de type "burden" sans contrôle appariés a été déployé comme logiciel open-source appelé COBT (Case-Only Burden Test) et il est désormais mis à disposition de la communauté scientifique. D'autre part, mes résultats montrent que les termes HPO utilisés pour grouper des patients similaires souffrant de maladies hétérogènes telles que les maladies du développement sont trop peu fiables pour systématiquement prioriser les variants exoniques de toute une cohorte dans le contexte évalué. En revanche, pour les patients les plus proches, la similarité sémantique calculée avec les termes HPO apparait comme un bon indicateur de la proximité génétique. Mon travail sur la priorisation de variants guidée par la similarité clinique pourra servir à la communauté scientifique pour améliorer les méthodes existantes et la précision de l'ontologie.

**Mots clés** : bioinformatique, génomique, test paramétrique, biostatistiques

# Title: Development of gene-prioritising methods using statistical genetics and clinical annotation for rare genetic disorders

## Summary:

To this day, near 70% of patients suffering from Mendelian diseases remain without any diagnosis after DNA sequencing. There is a need to study those disorders regarding their potential genetic causes with the newest genomic and bioinformatics tools to find potential treatments. Whole exome sequencing (WES) and whole genome sequencing (WGS) have improved the precision of rare disease diagnosis in clinical research. The new era of genomics has opened the door for a better understanding of the human genome and disease-associated variants in order to identify new diagnostic and therapeutic strategies for a growing number of rare and common diseases, thus giving birth to the concept of precision medicine. However, diagnosis and disease-associated variant discovery remain an overly difficult task because of clinical, genetic and disease heterogeneity as well as the analytical challenges they bring. Collectively, the various genetic mechanisms and technical artefacts can blur the statistical association signals, making the diagnosis and drug discovery very complex. Gene prioritisation methods can alleviate this task.

In this thesis, I first implemented a strategy consisting in aggregating variants in a specific genomic region to test for the accumulation of multiple rare variants across patients versus controls, rather than testing each variant individually. However, controls are rarely sequenced at the same time as cases, which can lead to batch effects. To overcome this limitation, I developed a case-only burden strategy that relies on a parametric test using publicly available sequence data from the Genome Aggregation Database (gnomAD). I have tested the hypothesis of the framework on the well-studied 1000 Genomes Project and applied it on a heterogeneous cohort of patients suffering from multiple cilia-related disorders.

Second, I used clinical data as a strategy to gain statistical power in the study of genotype-phenotype associations. Such phenotype-driven approach has already been successfully performed, leading to new genetic diagnoses in patients with developmental disorders. In this thesis, I worked on the evaluation of the potential of phenotypic terms from the human phenotype ontology “HPO” in order to prioritise variants shared by clinically similar patients from the Deciphering Development Disorder cohort (DDD) profiled with exome sequencing data.

The results presented in this thesis showed on the one hand that a case-only burden statistical test can be used in a clinical setting to identify causal variants in a heterogeneous cohort without *a priori*

knowledge. The implemented case-only burden test (COBT) has been released as an open-source software. On the other hand, I showed that the HPO used to generate groups of similar patients suffering from heterogeneous diseases such as developmental disorders were not reliable to systematically prioritise exome variants in the evaluated setting. However, the closest patients assessed with HPO semantic similarity do share rare genetic events and thus, phenotypically-guided variant prioritisation could be used in such case. Notwithstanding, my work on variant prioritisation assessed with clinical proximity will help the scientific community to improve existing methods for semantic similarity and accuracy of HPO terms.

**Keywords**: bioinformatics, genomics, parametric test, biostatistics

# Résumé substantiel :

À ce jour, près de 70% des patients atteints de maladies mendéliennes demeurent sans diagnostic après séquençage de leur ADN. Il est nécessaire d'étudier les causes génétiques de ces maladies à l'aide des nouveaux outils génomiques et de bio-informatique pour mettre en place de potentielles stratégies thérapeutiques. Les nouvelles méthodes de séquençage d'exome et de génome ont grandement amélioré la précision des études cliniques sur les maladies rares. La nouvelle ère de la génomique a permis une meilleure compréhension du génome humain et en particulier des variants génétiques associés pour un grand nombre de maladies rares et communes, ouvrant la voie à la médecine de précision. Cependant, le diagnostic et l'élaboration de stratégies thérapeutiques demeurent extrêmement compliqués du fait de l'hétérogénéité clinique et génétique, des défis statistiques associés et de la complexité de l'architecture génétique des maladies. L'ensemble des mécanismes génétiques et des artéfacts techniques peuvent brouiller les signaux statistiques, rendant le diagnostic et la recherche de thérapies très compliqués. Du fait de ces difficultés techniques, il n'est pas possible d'appliquer les mêmes méthodes d'identification de gènes candidats que pour les maladies communes. Des méthodes de priorisation de gènes ont donc été proposées pour simplifier ce problème et proposer des diagnostics.

Parmi ces méthodes, l'une des plus utilisées est une stratégie de mesure de l'accumulation de variants délétères au sein d'une cohorte de patients versus une cohorte contrôle. On parle de méthodes de type « burden » en anglais. En effet, il s'agit d'agréger plusieurs variants d'intérêt au sein d'une région spécifique et d'évaluer l'importance de leur accumulation dans une cohorte de patients par rapport à une cohorte contrôle plutôt que de tester chaque variant individuellement. Cela permet de pallier le déficit de puissance statistique dû au faible nombre de patients dans la cohorte. De nombreuses méthodes statistiques reposant le plus souvent sur une régression logistique ont été développées à ces fins. Néanmoins, les individus contrôle sont rarement séquencés conjointement aux patients et cela peut conduire à des biais d'analyse. Pour se prémunir contre cet effet, j'ai développé une stratégie de test statistique de type "burden" sans contrôle COBT (pour Case-Only Burden Test), en utilisant les données publiques de Genome Aggregation Database (gnomAD) comme paramètre. En effet, gnomAD est à ce jour la base de données publique contenant le plus grand nombre de génomes et d'exomes complets. Les individus de cette base de données peuvent collectivement simuler « l'attendu neutre » de la variation attendue dans la population générale et donc servir de paramètre au test statistique. L'intuition derrière cette méthode est qu'un gène accumulant plus de variants supposément pathogéniques au sein d'une cohorte de patients que l'attendu neutre est probablement associé à la maladie. Les avantages de cette méthode sont qu'elle ne nécessite pas de contrôles (et donc élimine

les biais dus à des contrôles non appariés), mais aussi qu'elle peut être utilisée pour prioriser des gènes candidats ou des groupes de gènes pertinents (tels que des voies métaboliques). Avant tout, un processus d'analyse bio-informatique comprenant différents filtres de référence a été mis en place pour éliminer tout potentiel biais d'analyse. L'hypothèse de cette stratégie « burden » a été testée sur les données du projet 1000 Génomes phase 3 en sélectionnant uniquement les individus européens non Finlandais pour éviter tout biais populationnel. Grâce à plusieurs tests de qualité d'ajustement, j'ai montré qu'une loi de Poisson était adaptée pour modéliser l'accumulation de variants rares dans un gène au sein d'une cohorte. En sélectionnant les variants synonymes, qui sont généralement neutres, j'ai montré que très peu de gènes étaient enrichis en variants rares dans la cohorte du projet 1000 Génomes. Néanmoins, certains gènes ont été significativement enrichis en variants synonymes, témoignant d'une trop grande sensibilité du test. D'autres méthodes de type « burden » sans contrôle existent déjà : « Test Rare vAriants with Public Data » (TRAPD) et plus récemment, « consistent summary counts based rare variant burden test » (CoCoRV) ont été développés avec la même ambition que COBT. Ces tests reposent sur la même hypothèse mais effectuent un test statistique comparant le nombre de patients mutés et non le nombre de variants accumulés dans les gènes, ce qui ne correspond pas *stricto sensu* à la définition d'un test de type « burden ». J'ai appliqué ces méthodes au jeu de données du projet 1000 Génomes. TRAPD et CoCoRV se sont révélés beaucoup plus sensibles que COBT. Différentes hypothèses ont été testées pour expliquer la détection de gènes enrichis en variants neutres. Une simulation de Monte-Carlo dans laquelle le jeu de données a été séparé en deux a été répétée mille fois pour relancer l'analyse et a permis de mettre en avant la sensibilité du test en détectant les mêmes gènes mutés plus souvent qu'attendu par chance dans une fraction des simulations. Ceci indique que ces gènes sont détectés par le test lorsque les individus mutés sont tirés au sort dans le jeu de données et montre donc la sensibilité du test. La structure de sous-populations a été testée à l'aide d'Analyse en Composante Principale (ACP ou PCA en anglais) et de régression mais elle ne justifie pas la détection de ces gènes qui résulte d'un artefact dû à la trop grande sensibilité du test. Le test statistique COBT a également été appliqué sur une cohorte hétérogène de patients atteints de ciliopathies séquencés sur différentes puces de ciliome. Une fraction seulement de ces patients avait un gène causal identifié. COBT a permis de réidentifier le gène causal de 26% des patients pour lequel il était connu en plus de proposer de potentiels variants modificateurs pour d'autres patients. COBT possède certaines limites en dehors de son applicabilité restreinte à la précédente version du génome (GrCH37). La puissance du test est très dépendante du nombre de patients dans la cohorte puisque l'analyse par sous-groupe de maladies n'a pas permis d'identifier d'autres gènes causaux que ceux réidentifiés avec la cohorte complète. Bien que l'analyse par voie métabolique n'ait pas révélé d'enrichissement en variant délétère biologiquement pertinent, la possibilité d'application pour COBT demeure très prometteuse pour les réseaux biologiques. Une autre piste d'amélioration de la méthode

serait d'adapter COBT aux variants non codants grâce aux domaines d'association topologiques puisque des méthodes de type « burden » existent d'ores et déjà à ces fins.

La seconde stratégie développée pour la priorisation de gène a été d'utiliser les données cliniques renseignées par les médecins dans les dossiers médicaux. Cette stratégie permet de guider la priorisation de gènes grâce à la proximité clinique et donc de gagner en puissance statistique lors de l'évaluation de l'association des gènes au génotype. Des analyses guidées par le phénotype grâce aux ontologies phénotypiques pour définir de nouveaux diagnostics dans les maladies du développement ont déjà été menées et ont montré leur efficacité. J'ai étudié la fiabilité des méthodes de similarité sémantique utilisant les termes de l'ontologie dite « HPO » (de l'anglais, Human Phenotype Ontology) pour construire des groupes de patients cliniquement similaires grâce à ses similarités sémantiques afin de prioriser les variants génétiques. Pour cela j'ai travaillé sur les données d'exomes complets de 4 286 trios du projet Deciphering Developmental Disorders (DDD). J'ai étudié l'influence de la méthode de calcul des différentes composantes permettant d'évaluer la proximité phénotypique : tout d'abord, la méthode de calcul de la quantité d'information contenue dans un terme ontologique, la méthode de calcul de la similarité sémantique et enfin la méthode d'agrégation des données de chaque terme lors du calcul de la similarité. Parmi les nombreuses méthodes de calcul de similarité sémantique, les cinq plus utilisées en recherche clinique ont été évaluées. Pour cela, grâce à l'outil Cohort Analyzer, j'ai d'abord montré que la qualité des données HPO de la cohorte DDD était suffisamment bonne : c'est-à-dire que les patients étaient phénotypés avec suffisamment de termes HPO et que ces derniers étaient assez précis. Ensuite, en considérant les gènes contenant au moins un variant *de novo* pour chaque patient, la similarité clinique a pu être comparée à la similarité phénotypique. J'ai montré que l'influence de la méthode de calcul de la quantité d'information contenue dans un terme HPO et la méthode de calcul de la similarité sémantique étaient très proches et qu'aucune ne reflétait la proximité génétique plus précisement que les autres. En revanche, la méthode d'agrégation des scores de similarité clinique dite « Best Match Average » est significativement plus précise que la méthode « Best Match Sum » ou le maximum des scores pour refléter la similarité génétique. J'ai montré que la sélection des couples de patients les plus proches phénotypiquement effectuée grâce aux termes HPO permettait de retrouver une plus grande proximité génétique que celle des couples choisis aléatoirement dans le jeu de données. En effet, les couples les plus proches selon le calcul de similarité utilisant les termes HPO partagent plus souvent des gènes contenant des variants *de novo* que les autres couples et cette proportion est grandement supérieure lorsque la méthode est étendue aux couples partageant des variants dans les mêmes voies métaboliques. Néanmoins, lors de la généralisation de la méthode à l'ensemble du jeu de données en définissant des couples significativement proches, ce résultat n'est pas reproduit : la proximité génétique des couples de

patients significativement proches et celle des autres patients est équivalente. La méthode est donc uniquement valable pour les couples de patients présentant le plus de traits phénotypiques semblables et ne permet donc pas en l'état la priorisation de gènes pour l'ensemble d'une cohorte. L'utilisation des termes HPO pour la similarité clinique entre tous les individus d'une cohorte pourrait ne pas fonctionner à cause des méthodes de calculs de similarité sémantique trop peu précises. Il est également possible que la structure même de l'ontologie soit en cause et que l'inexistence de termes précis pour les maladies développementales empêche l'utilisation des termes HPO pour rapprocher des individus cliniquement et refléter une réalité génétique.

Les résultats de cette thèse montrent d'une part qu'une stratégie de type "burden" sans cohorte contrôle peut fonctionner dans un cadre clinique et identifier des variants causaux sans *a priori* dans une cohorte hétérogène. La stratégie a été éprouvée formellement et son utilité clinique a été prouvée sur une cohorte hétérogène de patients atteints de ciliopathies. Une telle stratégie peut donc être utile dans le cadre d'une étude de patients hétérogènes souffrant de maladie génétique rare pour laquelle les outils traditionnels n'ont permis d'identifier aucun gène candidat. Le test implémenté, de type "burden" sans contrôle appariés a été déployé comme logiciel open-source appelé COBT (Case-Only Burden Test) et il est désormais mis à disposition de la communauté scientifique. D'autre part, mes résultats montrent que les termes de l'ontologie HPO utilisés pour grouper des patients similaires souffrant de maladies hétérogènes telles que les maladies du développement sont trop peu fiables pour systématiquement prioriser les variants exoniques de toute une cohorte dans le contexte évalué. En revanche, pour les patients les plus proches cliniquement, la similarité sémantique calculée avec les termes HPO apparait comme un bon indicateur de la proximité génétique. Mon travail sur la priorisation de variants guidée par la similarité clinique pourra servir à la communauté scientifique pour améliorer les méthodes existantes et la précision de l'ontologie.

# TABLE OF CONTENT

# TABLE OF FIGURES

# Table of tables



# Table of supplementary figures

# ACKNOWLEDGEMENTS

Doing a thesis was not the path I initially wanted to follow during my studies or even when I joined the clinical bioinformatics laboratory. But eventually, I decided to become a PhD student to work on an interesting project I was proposed, with the hope of learning more technical bioinformatics skills. Retrospectively, this has been a wise choice but for different reasons: not only have I mastered new bioinformatics skills and improved my scientific knowledge, but I have also grown up as a scientist and as a person. Following a doctoral path has not been an easy task, especially towards the end of the road. I have faced several moments of discouragement, particularly during lockdown, which have been mentally challenging times. These hardships have taught me patience when confronted to failure and to sometimes take a step back to get a glimpse of the bigger picture. I had the chance to benefit from the help of talented scientists and be supported throughout this journey by amazing people. Therefore, I would like to take here the opportunity to thank them.

First, I would like to thank Antonio Rausell who trusted me with this project and gave me the opportunity to work on a very interesting topic under great conditions. I am very grateful for the mentoring and most importantly for all the support provided during these years. I would like to thank Loredana Martignetti for following my work and advising me as well during this thesis.

I will probably not have enough kind words to thank my PhD student colleagues and friends who shared with me all the moments of joy, doubt and laugh every day: Barthélémy who helped me countless times with my code, suffered many middays at the gym with me and helped me move twice, Akira who introduced me to so many new musical horizons, Francisco with whom we had so many fascinating debates, Nigreisy and Romain with whom I had many passionate discussions at midday and that I am sad to have known so late in my thesis. I would like to thank also all the present and past members of the lab that I had the chance to work with and learn from: Yufei, Stefani, Stefania, Alejandro, Thibaut, Alexia, Marceau, Sara, Mathis. I am very thankful to each one of them for creating so many good memories in the lab with me. A special thanks goes to the members of the bioinformatics platform who helped me and gave precious pieces of advice, especially Fabienne Jabot-Hanin, Frédéric Tores and Patrick Nitschké.

I would also like to thank the people who collaborated with me during this work, including all scientists involved with the Cil'lico project, Pr. Higueras, and the members of my thesis committee Hugues Aschard and Chloé-Agathe Azencott who kindly followed my work every year and offered precious feedback.

Last but not least, I would like to thank the persons who know me the most and brought their emotional support during all these years: my dear friend Thibault who also followed the PhD path at the same time and helped me countless times in so many ways, my parents and my brother who have always trusted in me and supported me during all my studies and education, and of course my fiancée Laure, who offered an unconditional support every single day throughout this long journey.

# LIST OF ABBREVIATIONS

- **APPRIS**: Annotating PRincipal splice ISoforms
- **ASD**: Autism Spectrum Disorder
- **BMA**: Best-Match Average
- **BMS**: Best-Match Sum
- **BWA**: Burrows–Wheeler Alignment (BWA)
- **BWT**: Burrows–Wheeler Transform
- **CADD**: Combined Annotation Dependent Depletion
- **CMH**: Cochran–Mantel–Haenszel
- **COBT**: Case-Only Burden Test
- **CoCoRV**: Consistent summary Counts based Rare Variant burden test
- **DAG**: Directed Acyclic Graph
- **DD**: Developmental Disorder
- **DDD**: Deciphering Developmental Disorder
- **DNM**: *De Novo* Mutation
- **EGA**: European Genome-phenome Archive
- **HER**: Electronic Health Record
- **ERIC**: Emission-Reception Information Content
- **ESP**: Exome Sequencing Project
- **ExAC**: Exome Aggregation Consortium
- **FDR**: False Discovery Rate
- **gnomAD**: Genome Aggregation Database
- **GO**: Gene Ontology
- **GRC**: Genome Reference Consortium
- **GSEA**: Gene set enrichment analysis
- **GWAS**: Genome Wide Association Study
- **GWAVA**: Genome-Wide Annotation of VAriants
- **HWE**: Hardy-Weinberg Equilibrium
- **IC**: Information Content
- **IFT**: IntraFlagellar Transport
- **IGSR**: International Genome Sample Resource
- **Indel:** insertion/deletion
- **LD:** Linkage Disequilibrium
- **LOD**: Logarithm of the Odds
- **LoF**: Loss-Of-Function
- **MAF**: Minor Allele Frequency
- **MICA**: Most Informative Common Ancestor
- **MPO**: Mouse Phenotype Ontology
- **MRI**: Magnetic Resonance Imaging
- **MSigDB**: Molecular Signatures Database

- **NGS**: Next Generation Sequencing
- **NLP**: Natural Language Processing
- **OMIM**: Online Mendelian Inheritance in Man
- **PCA**: Principal Component Analysis
- **PCR**: Polymerase Chain Reaction
- **PHIVE**: PHenotypic Interpretation of Variants in Exomes
- **PID**: Primary immunodeficiencies
- **PRS**: Polygenic Risk Score
- **PTV**: Protein Truncating Variant
- **QC**: Quality Control
- **REVEL**: Rare Exome Variant Ensemble Learner
- **RFLP**: Restriction Fragment Length Polymorphism
- **SNP**: Single Nucleotide Polymorphism
- **SNV**: Single Nucleotide Variant
- **TAD**: Topologically Associated Domain
- **TPR**: True Positive Rate
- **TRAPD**: Test Rare vAriants with Public Data
- **VCF**: Variant Call Format
- **VEP**: Variant Effect Predictor
- **WES**: Whole Exome Sequencing
- **WGS**: Whole Genome Sequencing

# CHAPTER 1. INTRODUCTION

## 1.1 HISTORICAL CONTEXT OF HUMAN RARE DISEASES AND CLINICAL GENETICS

Clinical genetics is a medical field which aims at finding diagnosis and therapeutic strategies for hereditary diseases. This section of the introduction provides a brief historical overview of the research in human and clinical genetics fields starting from the first genetic linkage analyses to the changes brought by Next Generation Sequencing.

### 1.1.1 Before Next Generation Sequencing

#### *1.1.1.1 A brief history of genetic research and key discoveries*

In the nineteenth century, Gregor Mendel the father of modern genetics, brought light on the laws of heredity by crossbreeding peas. He introduced the model of parental transmission and the concepts of dominant and recessive alleles. This first milestone introduced the novel and strong idea that the existence of an invisible factor was passed down by both parents to the offspring and explained the phenotypic similarity. To this day a trait whose mode of inheritance can simply be described by a dominant, codominant, or recessive model at a single locus is usually referred to as "Mendelian" as a reference to Mendel's pioneering discoveries. Arguably, Charles Darwin could also be considered as the father of genetics with the evolution theory that gave birth to a whole domain called evolutionary biology. His theory allowed for biostatisticians to establish models and laws to further explain the transmission of traits. At the second half of last century, renown biologists and chemists Wilkins, Franklin, Watson and Crick (building on the work of Miescher, Levene, Chargaff that are often forgotten[1]) discovered the molecular structure of this invisible transmission factor, now known as DNA. As we will see in the next section, the discovery of the transmission and molecular structure of the genetic material allowed clinicians to identify the first genetic elements responsible for genetic diseases.

#### *1.1.1.2 Genetic mapping & linkage analysis*

Mendelian diseases have been studied quite early, even before the discovery of DNA structure, thanks to a simple method: genetic mapping, which consists in comparing the inheritance pattern of chromosomal regions to the one of a disease trait. Genetic linkage methods are based on the principle that physically close loci segregate together more often than loci far away from each other or located on different chromosomes. Researchers applied these methods to define a genetic distance, measured in centimorgan (*i.e.*, distance between two chromosomal positions resulting in one recombination out of 100). When the structure of DNA became available, researchers got access to natural DNA variation and thus to genetic markers of better quality. Genetic markers, such as RFLP (Restriction Fragment Length Polymorphism), microsatellites or later Single Nucleotide Polymorphisms (SNP), allowed

researchers to use polymorphisms to track down the inheritance of many Mendelian diseases[2]. Genetic markers have thus been studied to calculate their recombination rate and used to track down the segregation of genetically (and often physically) close traits. As the number of markers grew all along the genome researchers were able to study their segregation through pedigrees of patients with more and more precision. To apply such an approach successfully, studied traits need to accumulate enough variation in the affected families and be carried by enough individuals (*i.e.*, with a sufficiently large number of meiosis) to reach statistically significant results[3]. While the first parameter is barely an issue, the second parameter is more problematic. Even though the method is not biased by allelic heterogeneity, its power is affected by the potential complexity of the disease (locus heterogeneity) and incomplete penetrance. For instance, to identify the locus responsible for Ehlers-Danlos disease, 72 individuals needed to be genotyped across five generations[4]. Researchers then reported linkage as Logarithm Of the Odds (LOD) scores to show the co-segregation of genetic markers and disease loci. While the application of linkage analyses to specific mendelian diseases allowed the identification of regions harbouring several putative causal genes, large portions of family with a precise phenotype would have to be genotyped in order to reach statistical significance and to identify the causal region.

### 1.1.2 NGS revolution

In 1977, the first DNA sequencing method was developed by Sanger[5]. As we will see in this section, sequencing technologies have been improving at great speed until today, with spectacular methodological evolutions. Even after the huge improvement brought to Sanger's method in 1986 and 1990 using base-specific fluorescent dyes for the four DNA bases[6] with capillary electrophoresis[7], DNA sequencing was still expensive, long, and tedious. One operation of sequencing could only sequence up to 1 kilobase that represent all copies and thus all allelic variations of the target from the sample. Therefore, in the 90s, it was almost impossible to apply genomic sequencing in a clinical routine for more than a few genes and a tedious task to use it to identify causal genes with the financial and time limitations of the method.

Next-Generation Sequencing (NGS) or massively parallel DNA sequencing refer to the same procedure: the sequencing in parallel of small chunks of DNA to exponentially accelerate the speed of the sequencing process. NGS refers to the follow-up of the first generation of DNA sequencing technology after Sanger sequencing. In a single run of NGS, millions of targets of around 250 base pair are sequenced in parallel on an array and are cloned through Polymerase Chain Reaction (PCR, *i.e.*, fast technique aiming at replicating a specific DNA sequence up to millions of copies). Once the sequencing process is done, multiple chunks of overlapping DNA need to be reassembled. If no genome reference exists (*e.g.*, when sequencing a new emerging virus), then the process is called a *de novo* assembly. Raw DNA fragments, called reads, can overlap by a number $k$ of base pairs or k-mers. Therefore, they

need to be assembled in oriented reads comprising two or several overlapping reads. These so-called contigs are then assembled end-to-end to form scaffolds. The big scaffolds can finally be joined into the chromosomes of the organism. Several *de novo* assembly software applications have been created using different algorithms[8]. For Humans (and other species as well later) the whole genome has been deciphered with the Human Genome Project launched in 1990 and completed in 2003[9–12]. Thus, a first draft[9] followed by several corrections of the build of the human genome have been released by the Genome Reference Consortium (GRC)[11]. The most recent build is the 38th called GRCh38 (for Genome Research Consortium human build 38)[13]. Once raw data is provided by the sequencer, it must be aligned on this reference genome with the help of alignment programs. The most famous that has become a gold standard is Burrows-Wheeler Alignment (BWA) tool that uses Burrows–Wheeler Transform (BWT) algorithm[14].

Several companies developed their own NGS machine each with its specificities. Illumina proposed a plate-based technology (called reversible termination) whereas Life Technologies/Applied Biosystems and Roche used a bead-based solution to bind DNA primers (with methods respectively called sequencing by ligation and pyrosequencing)[15]. Due to the competition and the demand, prices collapsed very rapidly and made NGS affordable to a lot of laboratories, allowing for thousands of research teams to study genomic sequences in many new ways. More recently, a third generation of NGS, allowing to sequence longer reads have been developed. The most famous is Oxford's Nanopore MinION, approximatively the size of a USB stick, which can sequence a whole DNA molecule in one read. Such a revolution now allows geneticists to perform sequencing on-site and at a greater speed, for example to react to a pandemic.

The main issue with NGS is that as compared to Sanger sequencing, they are prone to error. Sanger sequencing has an error rate between $10^{-4}$ to $10^{-5}$ per call whereas the error rate for NGS is between $10^{-2}$ to $10^{-3}$ [15].

### 1.1.3 Rare Mendelian diseases

#### *1.1.3.1 Definitions*

Rare diseases are disorders affecting a small proportion of the population. This definition can differ according to the country or organization: according to the United States Rare Disease Act of 2002, a rare disease is a "disease or condition affecting fewer than 200,000 persons in the United States" whereas the European Commission defines them as "life-threatening or chronically debilitating diseases which are of such low prevalence that special combined efforts are needed to address them"[16]. Low prevalence is vague, but European Union seems to agree on a prevalence lower than 5 per 10 000 inhabitants[17]. There are dozens of other definitions in the different national official and

scientific literature. Most of them are linked to the prevalence and revisit the number depending on the data from the country. Sometimes rare diseases can be referred to as orphan diseases, but according to the European Union and the United States, this is neither strictly nor legally the same definition.

A key point in genetics that became clear with NGS is that rare variants (usually considered below 1%) are rare at the individual-level but are very common at the population scale. A rare variant in a disease-associated gene has good chances of segregating in the family. With NGS it has become easier to sequence several individuals such as full trios for family linkage analysis. However, the true potential of NGS for rare disease lies in the reproducibility of the analysis: it becomes feasible to apply WES or WGS and analyse hundreds of patients suffering from the same disease.

#### *1.1.3.2 Online resources*

Rare variants have a strong role in the mechanisms of Mendelian diseases. Hence, to help physicians and researchers, multiple references databases have been created for rare diseases and genetic variants. The Online Catalog of Human Genes and Genetic Disorders (OMIM) is the first catalogue of known human mendelian disorders (https://www.omim.org). It was created in 1966 and was made publicly available in 1987 before being easily accessible on the Internet in 1995. OMIM gathered over 24,600 entries in 2018 with 6,259 identified phenotypes and 3,961 associated genes[18]. In total, OMIM now gathers information on more than 16,000 genes. Every entry in OMIM has a unique identifier. Specific variants representing disease-causing mutations (or at least having a positive correlation with the disease) can have entries in OMIM if they respect the inclusion criteria, namely being the first mutation to be discovered, having a high population frequency, a distinctive phenotype, a historic significance, an unusual mechanism of mutation and pathogenetic mechanism, and a distinctive inheritance[19]. Researchers can thus look for specific variants of interest when studying a gene or a disorder as well as adding their own findings.

Orphanet is another resource that was created just before NGS, at the advent of the Internet in 1997, in order to gather and share information on rare diseases in France (https://www.orpha.net). In the early 2000s, this became a European effort and a network involving 41 countries, funded by the European commission. One of the missions of Orphanet is to simplify the work of researchers and clinicians by providing them with a rare disease nomenclature: the ORPHA code which gives to each rare disease a unique identifier aligned with most other clinical databases such as OMIM and Human Phenotype Ontology (HPO) which will be detailed later in this introduction. Orphanet serves as a reference and encyclopaedia for rare diseases description and updated information on them. Rare

disorders are classified in a nomenclature providing useful information for researchers about the clinical symptoms as well as available genetic information[20].

Many other resources have been developed for rare disorders studies, opening the door for a new era of data sharing. NGS and data sharing have given access to information about very low frequency variants for rare diseases.

#### *1.1.3.3 Primary ciliopathies: an example of family of genetic rare diseases*

Ciliopathies are a category of rare diseases that affect cells' cilia, which are complex organelles with a hair form present in almost every cell type. Cilia can be divided in two categories: non-motile (or primary) cilia and motile cilia. Few cell types differentiate with specialized motile cilia. They rely on a machinery attached to a central pair of microtubules (Figure 1) such as in sperm cells (for cell locomotion) or cells of the respiratory tract (with a role in respiratory airway clearance). Most cells carry primary cilia which have a role in signal transduction or chemosensation. Most disorders that will be discussed in this manuscript are primary ciliopathies (*i.e.*, ciliopathies affecting primary cilia).

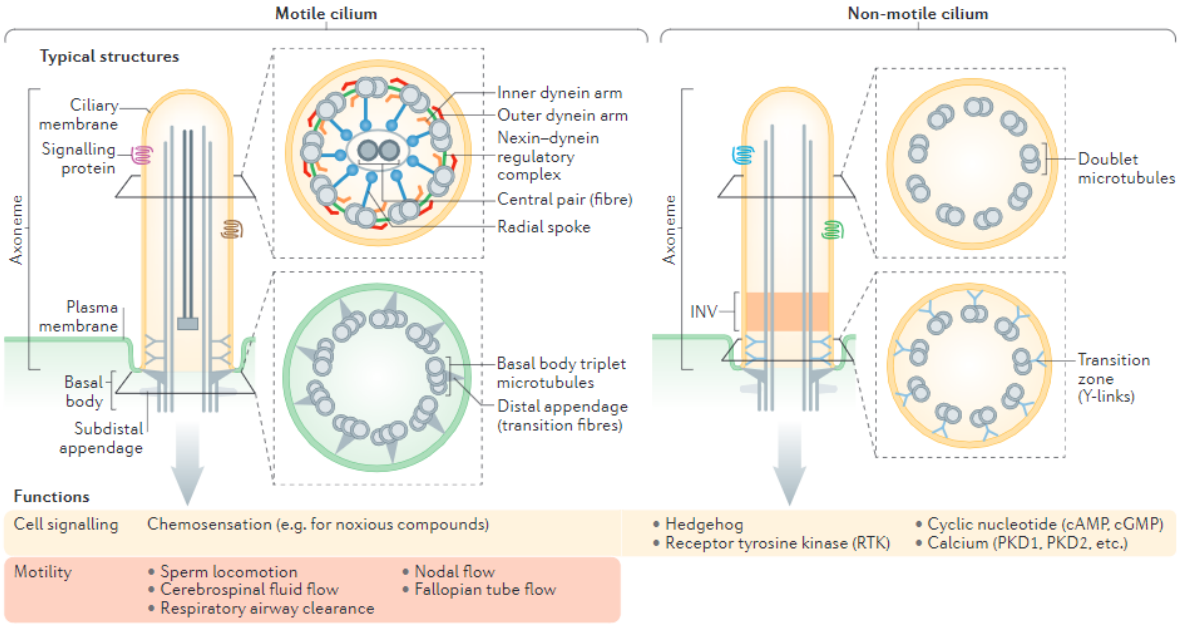



*Figure 1: Schematic representation of structures and functions of primary and motile cilia*
*Source: Reiter J, Leroux M. (2017)*[21]

Primary ciliopathies can affect various organs with different symptoms such as ataxia, epilepsy and mental disability for the brain.

Ciliopathies as a class of diseases is a rather new concept: the term was first used in 1984[22]. It has then been widely used to describe ciliary dysfunction which can be split in several classes depending on:

- The protein affected

    - First-order ciliopathies are caused by a dysfunction of a ciliary protein such as intraflagellar transport (IFT) components disruption that leads to Jeune syndrome
    - Second-order ciliopathies are caused by a dysfunction of a non-ciliary protein
- The type of cilia affected
    - Motile ciliopathies result from a dysfunction of ciliary motility such as Primary Ciliary Dyskinesia (PCD)
    - Sensory ciliopathies result from a dysfunction of cilia signaling such as Joubert syndrome

Ciliopathies discussed in this manuscript are first-order sensory ciliopathies, *i.e.*, primary ciliopathies affecting ciliary proteins. Jeune syndrome for example is characterized by several symptoms which include a narrow thorax with short ribs, trident acetabulum (*i.e.*, small bony projections in the socket of the hip bone), cone-shaped epiphyses (*i.e.*, the ends of the long bones) and polydactyly. It can result from a mutation in DYNC2H1 gene which codes for chain 1 of cytoplasmic dynein 2 that is part of the IFT complex[23]. Different mutations can affect the IFT complex and lead to multiple clinical manifestations such as chronic infections of the airways or chondrodysplasias with obligate renal involvement and thus different syndromes such as Jeune syndrome, Mainzer–Saldino syndrome or Sensenbrenner syndrome[24]. However, even if it not usual, patients suffering from Jeune syndrome might have renal or retinal symptoms but rarely when DYNC2H1 is mutated[23]. Likewise, Joubert syndrome is characterized by muscular hypotonia, cerebellar ataxia, abnormal eye movements and breathing pattern in infancy and cognitive impairment, yet patients with dysmorphic features such as hypertelorism (*i.e.*, increased distance between the eyes) or a broad forehead have been described in the literature[24]. In this particular case, there is a pathognomonic symptom (*i.e.*, symptom that allows by itself to diagnose a disorder) called molar tooth sign "defined by an abnormally deep interpeduncular fossa; elongated, thick, and mal-oriented superior cerebellar peduncles"[25]. It looks like a molar tooth on axial brain Magnetic Resonance Imaging (MRI) at the junction between the midbrain and the hindbrain. However, in most ciliopathies, there is a high phenotypic variability. This heterogeneity is also genetic: INVS and CEP290 genes can both carry recessive mutations that will result in nephronophthisis (*i.e.*, "renal ciliopathy characterized by reduced ability of the kidneys to concentrate solutes, chronic tubulointerstitial nephritis, occasional presence of cysts, and progression to end stage renal disease"[26]) and can be associated with other ciliopathies such as Joubert syndrome or Senior–Løken syndrome with added symptoms on other organs (*e.g.*, retinal degeneration for Senior–Løken syndrome).

The majority of ciliopathies are homozygous recessive diseases that are clinically and genetically heterogeneous. Hence, they are complicated to classify and consequently, few epidemiological data

are available. The prevalence of most ciliopathies is unknown[24]. Nevertheless, two lists of ciliopathies are available: the list from J. Reiter and M. Leroux[21] and the CiliaCarta gene list[27]. Reiter and Leroux's list includes 187 genes implicated in 35 ciliopathies and 241 genes associated with ciliary structures and functions that lack evidence of ciliopathy implication but would probably lead to ciliopathies if disrupted. The 187 established ciliopathy genes come from the manually curated database of the SysCilia consortium (http://syscilia.org/index.shtml). Candidate genes were identified from multiple studies aiming at uncovering the ciliome, most of which are gathered in the CilDB database[28]. CiliaCarta comprises 956 cilia-related genes including 209 putative new ones. They were assessed through literature, annotation, and genome-wide Bayesian integration of experimental data. By integrating data from proteomics, genomics, expression, evolutionary data plus two public datasets, CiliaCarta provides a Bayesian score: CiliaCarta score, which is the likelihood of a gene to be ciliary. Scores vary from -8.78 to 11.70: a positive score of 1 means the gene is twice as much likely to be cilia-related, and this to be involved in a ciliopathy.

## 1.2 Bioinformatics methods for gene-disease association identification

Clinical genetics gather two main distinctive fields that require specific methodologies, *i.e.*: complex diseases such as cancers or diabetes on the one hand and rare diseases on the other hand. Furthermore, disorders can have a multitude of specific mechanisms. Hence, clinicians and researchers have worked on tools and methods to decipher and identify causality of those different diseases. In this section, we will briefly describe the basis of the main diseases' differences and associated methods.

### 1.2.1 Variant frequency and association methods

A popular hypothesis among geneticists is that common variants are responsible for common and complex diseases whereas rare variants cause rare mendelian diseases[29]. This hypothesis is mainly based on Kimura's neutral theory of evolution[30]: highly deleterious variants will be removed by purifying selection, therefore, variants responsible for rare life-threatening diseases should subsist at very low frequency in the population. Even though this theory has been proven to be only partially true and is still prone to debate in the community[31], it has led researchers to apply different methods to detect specific classes of variants along the variant frequency spectrum (Figure 2). For complex disorders and more generally complex traits, researchers have investigated the role of common variants through Genome-Wide Association Studies (GWAS) and successfully identified significantly associated loci and complex disease phenotypes. Those methods are applied on variants with a high frequency in the population. The assumption is that common variants cause complex common diseases while rare variants are responsible for rare mendelian diseases. Common complex diseases can be

highly polygenic, and variants implicated usually have a low effect size. Since the mid-2000s lots of SNPs and loci have been associated with numerous complex traits including type 1 and type 2 diabetes, Crohn's disease, and other auto-immune diseases[32] as well as Parkinson's disease and various cancers. Two key factors of the popularity of the method were its simplicity and its cost. Nevertheless, variants identified by GWAS explain at most up to 5-10% of heritability of the diseases. This is often referred to as the so-called missing heritability problem[33].

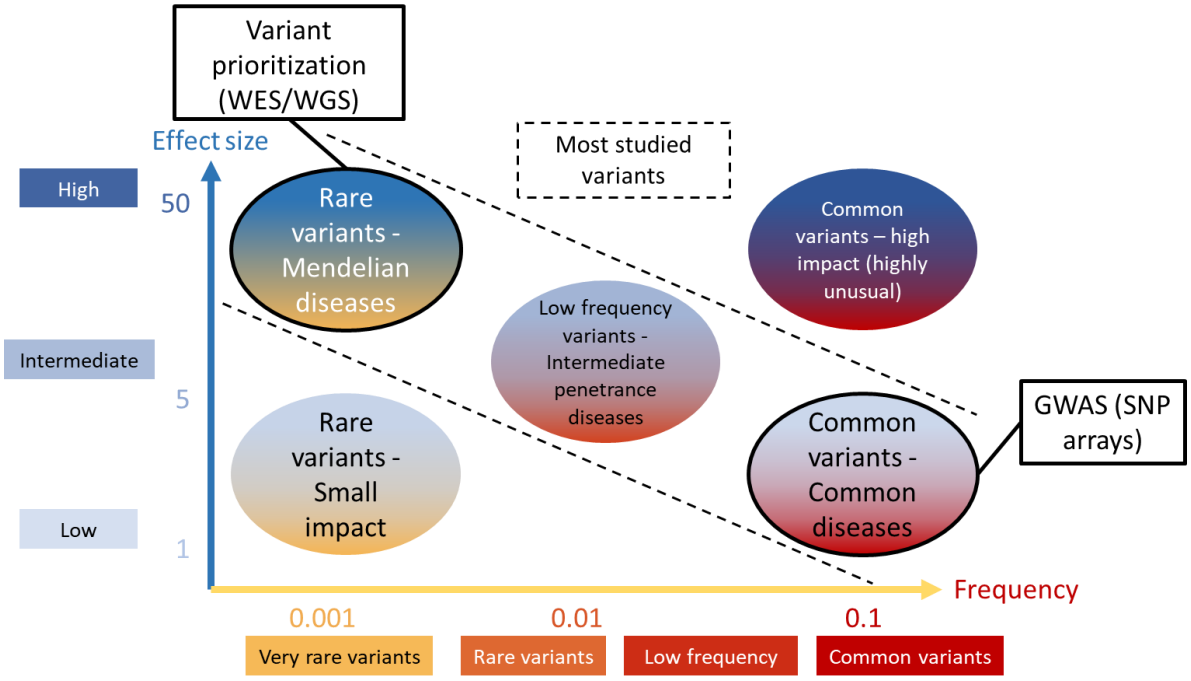



*Figure 2: Effect size according to the allele frequency of variants*

GWAS allowed to identify many SNVs implicated in complex diseases, however it is now admitted that other factors including epigenetics, CNVs or other structural variants as well as rare variants play a significant role in such disorders [34].

### 1.2.2 Population genomics and clinical data sharing

Next Generation Sequencing techniques have opened the door for a new era in genetics: the era of big data. As researchers identified new disease-causing rare variants, data sharing became more systematic. Many accessible valuable resources became available, giving more precision in allele frequencies for researchers, and providing access to low allele frequency variants. With this new opportunity, new challenges as well as new solutions emerged. The huge quantity of data that quickly became available have been aggregated in comprehensive databases such as ClinGen[35] and ClinVar[36] respectively referencing clinical relevance of genes and variants for precision medicine. Additionally, large-scale sequencing projects gave to the community access to their data. In 2008, the 1000 Genomes Project was initiated with the aim of creating the first and largest catalogue of genetic

variants of the human genome. The project had several milestones and releases, with 2,504 individuals from 26 populations at the end of phase 3[37] but most importantly gave an easy and free access to all researchers to their whole genome data with the creation of the International Genome Sample Resource (IGSR)[38]. One of the key points of the 1000 Genomes Project is that individuals are supposed to be healthy and can thus be used as controls. Shortly after, in 2009, the National Heart, Lung, and Blood Institute (NHLBI) released the Exome Sequencing Project (ESP) to identify rare variants associated with heart, lung and blood diseases and traits[39]. This study gathered more than 6,500 exome sequences that has been used in multiple studies[40–42]. In most studies though, variant and gene-level association tests were underpowered and with state-of-the-art methods, researchers estimated that more than 100,000 samples would be necessary to detect true rare variant-disease associations across the exome[39]. Hopefully, in 2017 the Broad Institute provided the community with Exome Aggregation Consortium (ExAC), a browser and database gathering 60,706 exomes[43]. A couple of years later, it evolved into the Genome Aggregation Database (gnomAD) gathering 125,748 whole exomes and 76,156 whole genomes in the version 3.1[44]. All those databases have allowed the identification of hundreds of thousands of new variants in different human sub-populations.

### 1.2.3 Genetic variation and mode of inheritance

Whole Genome and whole exome sequencing produce a colossal amount of genomic: between 3.5 and 4.4 million SNVs for WGS[45,46] and up to 22,000 for WES[45,47]. Therefore, variant filtering and prioritisation have become mandatory. Different strategies can be proposed for this purpose.

#### *1.2.3.1 De Novo variants*

A common scenario in rare diseases such as developmental disorders[48], is the occurrence of *de novo* variants in patients (*i.e.*, variants found in the proband but absent in both unaffected parents). The estimated number of *de* novo variations is estimated at 74 SNVs per generation[49] as the human germline mutation rate is estimated at $1.18 \times 10^{-8}$ per position along the genome[50] (even though mutagenesis is not random along the genome and so-called hot spots of mutations have been identified[51]). Considering only exomes, this brings to an approximation of 1 to 2 *de novo* variants per individual. However, the number of *de novo* mutations (DNM) per family is not stable: there is a paternal-origin bias among germline DNMs arising in families as compared to those from maternal-origin[49,50]. Paternal age is also correlated with the number of DNMs[48]. Sequencing whole exomes of family trios and analysing the impact of DNMs in proband patients has proven to be a very effective strategy to decipher the mechanisms of rare mendelian diseases such as Kabuki syndrome[52], Bohring-Opitz syndrome[53] or Schinzel-Giedion syndrome[54] and complex diseases from a whole range of developmental disorders including autism, epilepsy, or intellectual disabilities[48]. *De novo* variation analysis in WES can help providing a molecular diagnosis for 25 up to 61% of unresolved cases of

patients[55,56]. This method is thus very effective, especially in paediatric diseases with a dominant mechanism. For late-onset diseases and other disease mechanisms including compound heterozygous variants for example, even if DNMs cases have already been reported, *de novo* variants are unlikely[45].

#### *1.2.3.2 Molecular impact of genomic variation*

The molecular consequence of the variants on the viability of the protein and its efficacy in playing its role is obviously impactful on the phenotype. Studying these variant consequences is a way to discard harmless variants and select the most damaging variants that will most probably be linked to the studied disorder. Because of the redundancy of the genetic code, some single nucleotide mutations can lead to no amino acid change. Those are the so-called synonymous variations. They often would not be under selective pressure. While this assumption is correct most of the time, there is evidence that it is not an absolute truth as synonymous variants can be under purifying selection because of their impact on exon inclusion[57]. In clinical genetics, synonymous variants are usually discarded from the analysis. Variants that eventually change the protein amino acid are called missense variants and have a more challenging interpretation as they can be more or less harmful: some might reduce the efficacy of the protein (by changing the protein conformation, leading to a diminution in the specificity to the connection to a ligand for example) and some might not affect it at all. Loss-of-function variants (LoF) can result from an early stop codon, a shift in the reading frame or the removal of a splice site. They have a direct impact on the viability of the protein and are often considered the most damaging ones. Missense and LoF variants are thus susceptible of undergoing selective pressure. Tools that will be further detailed later in this introduction like REVEL[58] or CADD[59] can help discriminate between damaging and benign missense and LoF variants. Non-coding variants are usually overlooked as well as they are hard to interpret but they are getting more and more interest since they have been proven to be linked with genetic disorders[60]. The use of *in silico* annotation tools such as the Variant Effect Predictor (VEP)[61], ANNOVAR[62] and SnpEff[63] has helped researchers evaluating the molecular impact of genomic variants on the proteins, hence the phenotype. As mentioned, knowing the molecular consequences of variants is not enough: even if among the 22,000 exome variants, only a fraction will be loss-of-function or missense, one still needs to discriminate between benign and damaging variants to lower the false positive rate. Moreover, a gene may lead to different protein isoforms and consequently different phenotypes, making the interpretation of a variation even more complicated. Nonetheless, not all isoforms are biologically relevant. This is why the GENCODE consortium[64] developed the Annotating principal splice isoforms (APPRIS) database[65] to help researchers selecting the most relevant isoform of the protein and thus select a transcript from the annotation process. The rationale behind this idea is that the main isoform of a protein shows the biological reality of a protein-

coding gene. To annotate every protein-coding gene, APPRIS uses protein structural and functional features as well as cross-species conservation in order to define the most relevant “principal” isoform.

#### *1.2.3.3 Diseases modes of inheritance*

Humans are diploid organisms that carry two copies of each gene on their autosomes. The combination of the two alleles at a locus defines the phenotype. The main models explaining this interaction for each allele are:

- Dominant allele fully responsible for the phenotype (if an individual has allele A and allele O for the blood type gene locus on homologous chromosomes, type A is expressed: A is dominant over O)
- Recessive allele that is present but whose effect is masked by the dominant allele (in the above example O is recessive over A; to have an O blood type, one needs to have both O homologous alleles)
- Co-dominant alleles that both are expressed in the phenotype (if an individual has A and B alleles in the blood type example, blood type will be AB: they are codominant)

Thus, in Mendelian disorders especially, the interaction between the two alleles is highly impactful. Mendelian disorders can be autosomal or X-linked and dominant or recessive. A disease with a dominant mechanism needs only one pathogenic heterozygous variant whereas a recessive disease requires two homozygous variants, one on each copy. However, a gene with two different variants that are not in phase on different haplotypes affecting both copies can also trigger a disease mechanism. This means that two heterozygous variants can also lead to specific disorders. Such variants are called compound heterozygotes.

Other Mendelian diseases can have incomplete penetrance, meaning that some individuals might bear the risk genotype without suffering from the disease while others, with the same genetic configuration, do[66]. Famous examples of dominant autosomal incomplete penetrance are the mutations in BRCA1 and BRCA2 genes responsible for 3% of breast cancers breast cancer[67]: 57% of females carrying a mutation in BRCA1 and 49% in BRCA2 have a risk of contracting a breast cancer; in other words, the penetrance is 57% for BRCA1 and 49% for BRCA2[68].

## 1.3 BURDEN TEST STRATEGY

To associate a phenotype with a genotype in rare diseases, GWAS method is not appropriate as it relies on common variants and is testing the effect of multiple single variants all along the genome. Other single-variant-based statistical tests have very low statistical power due to the low number of cases

and low effect sizes. Hence, a strategy relying on an aggregation of relevant variants has been proposed: burden tests.

### 1.3.1 Burden test conceptual idea

Initially, to test for association between genotypes and phenotypes in rare diseases, biostatisticians used single-marker and multiple marker tests (*i.e.*, tests respectively evaluating single SNP-to-phenotype association or multiple SNPs-to-phenotype association) such as $\chi^2$-test, Fisher's exact test or regression models[69]. However, statistical power when testing low-frequency variants for a trait was usually low[70]. In GWAS, the contribution of each variant is assessed by the test. Nevertheless, for rare variants, even with the same effect size as with common variants, association tests will be less powerful[71,72]. Furthermore, there are more rare than common variants, which means that multiple testing correction is mandatory to compensate the increase of likely-erroneous inferences that will be performed. Therefore, a solution is to aggregate variants to test for the accumulation of multiple rare variants in a specific region across patients versus controls rather than testing each variant individually in a single burden variable[73,74]. Aggregation tests and burden tests are not strictly identical methods. An aggregation test evaluates the cumulative effect of a combination of variants within a gene or a region rather than each variant individually (as in GWAS), but for all samples collectively, whereas a burden test takes into account the accumulation of variants in a gene for each sample. A burden test is thus an aggregation test in which the information brought by every variant and each sample is summarized in a single variable.

One biological assumption for applying a rare-variant aggregation method is that allelic heterogeneity exists in the genes tested for disease-association. This suggests that each patient involved with a rare variant in the tested gene would explain only a fraction of the variability and that the method evaluates the cumulative contribution of all patients and alleles to associate the phenotype to the causal gene. Assessing the presence of missense or loss-of-function variants in patients versus healthy individuals can lead to gene-disease association. Hence, the use of a control cohort is mandatory to assess the background variation of the genes and detect unusual genetic load[75]. Different statistical frameworks that will be thoroughly discussed in the next paragraph can be used to compare cases and controls. Before going through association testing, the first step consists in ensuring that cases and controls are indeed comparable through sample selection, pruning and dataset harmonization of the qualifying variants, *i.e.*, collection of variants that pass all established filters (*e.g.*, related to sequencing quality control, damaging predicted consequence, allele frequency filter, etc) among both cases and controls. Ideally, cases and controls should be sequenced with the same technology (same sequencing technique, machine, chip version) by the same technicians and be processed jointly. This is often not the case for budget and practical reasons.

### 1.3.2 Pipeline for data pre-processing

Before applying any sort of statistical test, the genomic data needs to be processed both to ensure minimal batch effect and biological relevance of the aggregated variants usually referred to as "qualifying variants". It is critical to apply a common bioinformatics pipeline for cases and controls that will prevent any bias and make sure that any difference between them is biologically relevant. The strategy is summarized in Figure 3.

#### *1.3.2.1 Patients' and controls' selection and filtering*

The very first step is to gather cases for the studied disorder and controls. The latter have to be healthy individuals if possible or at least without comorbidity for the traits that could be involved in the disease. As previously stated, when publicly available cohorts are used for an association test, specific confounders can bias the analysis: namely ethnicity of patients and controls as well as factors directly correlated to the technical and analytical artifacts such as sequencing techniques in different platforms, various base calling and alignment procedures, different read depth etc[76]. To reduce the impact of subsequent batch effect, quality control (QC) steps need to be applied: samples could be removed from the analysis because of low capture region specificity or low coverage. The two main reasons for sample pruning are genetic relatedness and population stratification. Kinship between cases can bias the analysis by giving more weight to variants non-associated with the studied trait. Another risk in rare disease studies is that implicated variants might be discarded due to a too high allele frequency in the family. To avoid this, it is recommended to only keep unrelated samples in the case cohort. Finally, population structure is probably the most impactful bias to care for: if all cases share a common ancestry while controls are heterogeneous (or worse, from another close subpopulation), any signal measured would most certainly be related to it rather than to the disease. Variants present at high frequency in population from European ancestry might be at very low frequency in samples from East Asian ancestry. Even with recent progress in genomic data sharing and publicly available databases, non-European ancestry groups are under-represented[77,78]. Even if it does not preclude the necessity to control for population stratification, the number of variants of interest will be higher in under-represented populations simply because less sequenced individuals are available: allele frequencies of rare variants will tend to be lower and provide false positives for qualifying variants[78]. Synonymous variants are supposedly neutral and can be used to verify that no batch effects are observed between cases and controls. A sample carrying significantly more synonymous variants than the others in cases or controls should be excluded from further analyses. Restricting to synonymous variants is also a mean to evaluate any deviation between the number of variants between cases and controls to ensure that no other populational bias remains. If no population information is available for cases or control samples, one method to get rid of populational

bias is to use Principal Component Analysis (PCA). A first possibility is to perform the PCA on common variants of samples from a publicly available database with known ancestry, such as the 1000 Genomes Project and to project cases or controls samples on it (assuming all necessary quality checks have been applied to the data). Then, outliers coming from a different ancestry than the rest of the cohort can be discarded from the analysis. The other possibility is to apply the PCA solely on the cohort (with cases, controls or jointly) to directly identify outliers without further information about the ancestry. Analogous methods have also been successfully used to correct for population stratification in GWAS studies [79].

#### *1.3.2.2 Qualifying variants quality control*

Despite controlling for populational biases, technical artifacts can still affect association tests, particularly coverage depth and genotype calling. Quality control filters need to be applied at the variant level. First, to avoid huge discrepancies in the number of qualifying variants in one cohort or the other. For example, if the sequencing coverage depth is dramatically lower in controls than cases in a particular gene, the test might result in an artefactual excess of qualifying variants. Several strategies can be used to avoid such bias. The more robust, but also computationally intensive way of controlling this bias is to test for rare-variant association on modelled sequencing reads without calling for genotypes[80]. This method uses a likelihood-based approach modelling sequencing reads for burden testing with a bootstrap strategy to assess significance. This is a way to compensate potential imbalance of SNVs due to the different read depths between cases and controls. This method is precise and can handle complex scenarios but is computationally demanding and not very flexible in terms of the set of burden tests that are applicable. The other method consists in retaining the regions where the depth of coverage is high enough both in cases and controls, independently of whether variants have been called or not. This strategy aims at minimizing the artificial impaired balance of qualifying variants. Different filters or thresholds can be applied to that aim. However, what seems to be accepted as a golden standard is to keep genomic regions where 90% of samples are sequenced with at least 10X of coverage depth[81–83]. Another possibility is to statistically test for independence between case/control status and coverage depth at 10X[84]. Additionally, classical variant quality control filters can be applied to remove potential sources of bias, such as the Phred quality score (*i.e.*, Phred-scaled posterior probability that samples are all homozygous), the genotype Phred quality score (*i.e.*, Phred-scaled posterior probability that a base is incorrectly called)[85], the Phred quality score normalized by depth or a more refined score, the Variant quality score log-odds (*i.e.*, logarithm of the odds of the variant being true as compared to a trained Gaussian mixture model)[86].

#### *1.3.2.3 Qualifying variants filtering and prioritisation*

After quality control of the data, qualifying variants can be selected based on their molecular consequence. Usually non-coding variants (UTRs and intronic variants) as well as synonymous variants are filtered out when the goal of the analysis is to measure the accumulation of variants susceptible to modify the protein, mainly driven by LoF, missense and indels (insertion/deletion) variants. A common filter is to use an allele frequency threshold as mentioned previously, following the rare variant-rare disease hypothesis. With the availability of hundreds of thousands of WES and thousands of WGS of supposedly healthy individuals, it has become feasible to utilise databases such as gnomAD for allele frequency filtering; even though some ethnicities are under-represented as compared to Caucasian populations. GnomAD 2.1 release enables the identification of variants as low as $8 \times 10^{-6}$ of allele frequency. This filtering strategy is often referred to as Minor Allele Frequency (MAF) filtering method. This MAF is commonly referred to as the external MAF. The internal MAF would be the MAF of variants within the cohort. It is also important to examine the allele frequency of variants considering cases and controls collectively. Usually, variants are considered rare below 1% of MAF in the population, thus a MAF filter of 0.01 is often applied for rare diseases to remove common variants.

Additional filtering of variants can be based on their predicted pathogenicity. A way to identify potentially harmful variants is by aligning human sequences with other species and derive conservation scores. Highly conserved regions across species are less likely to accumulate new mutations. Therefore, several conservation scores such as GERP[87], phastCons[88] and phyloP[89] allowing to identify such evolutionary conserved regions which have been used widely to predict the potential harmfulness of variants. Polyphen-2[90] and SIFT[91] (Sorting Intolerant From Tolerant) are probably the most famous tools and pioneers in the coding variant scoring field. They both predict the effect of non-synonymous variants. SIFT computes the likelihood that the amino acid change will affect the protein function by aligning homologous sequences: the assumption behind the method is that highly conserved regions tend to be less tolerant to mutations. SIFT provides a normalized score with values ranging between 0 and 0.05 being predicted to alter the function of the protein. Polyphen-2 works in a similar fashion: it calculates a naïve Bayes posterior probability that the mutation will be damaging thanks to a classifier trained on human variation data and provides a category for the variants (benign, possibly damaging, or probably damaging). The Combined Annotation Dependent Depletion score (CADD)[59] has quickly become a gold standard. The latter is using multi-species alignment to derive proxy-neutral variants with changes between humans and ancestors versus simulated deleterious variants (*i.e.*, proxy-deleterious *de novo* matching set of variants drawn from the proxy-neutral variants) and train a logistic regression model on those annotated neutral and putative deleterious variants. The variants are annotated with functional prediction, genetic context (such as GC or CpG

content) and conservation scores including GERP++[92], phastCons and phyloP. One of the most recent and broadly used scoring tools is the Rare Exome Variant Ensemble Learner (REVEL)[58] that aims at providing a pathogenicity score for missense variants. The latter is based on a machine learning model (random forest) that uses numerous different features including already described tools (CADD, SIFT, PolyPhen-2, phastCons, GERP...). The use of prioritising tools and filtering strategy dramatically increases the specificity of the tests[93].

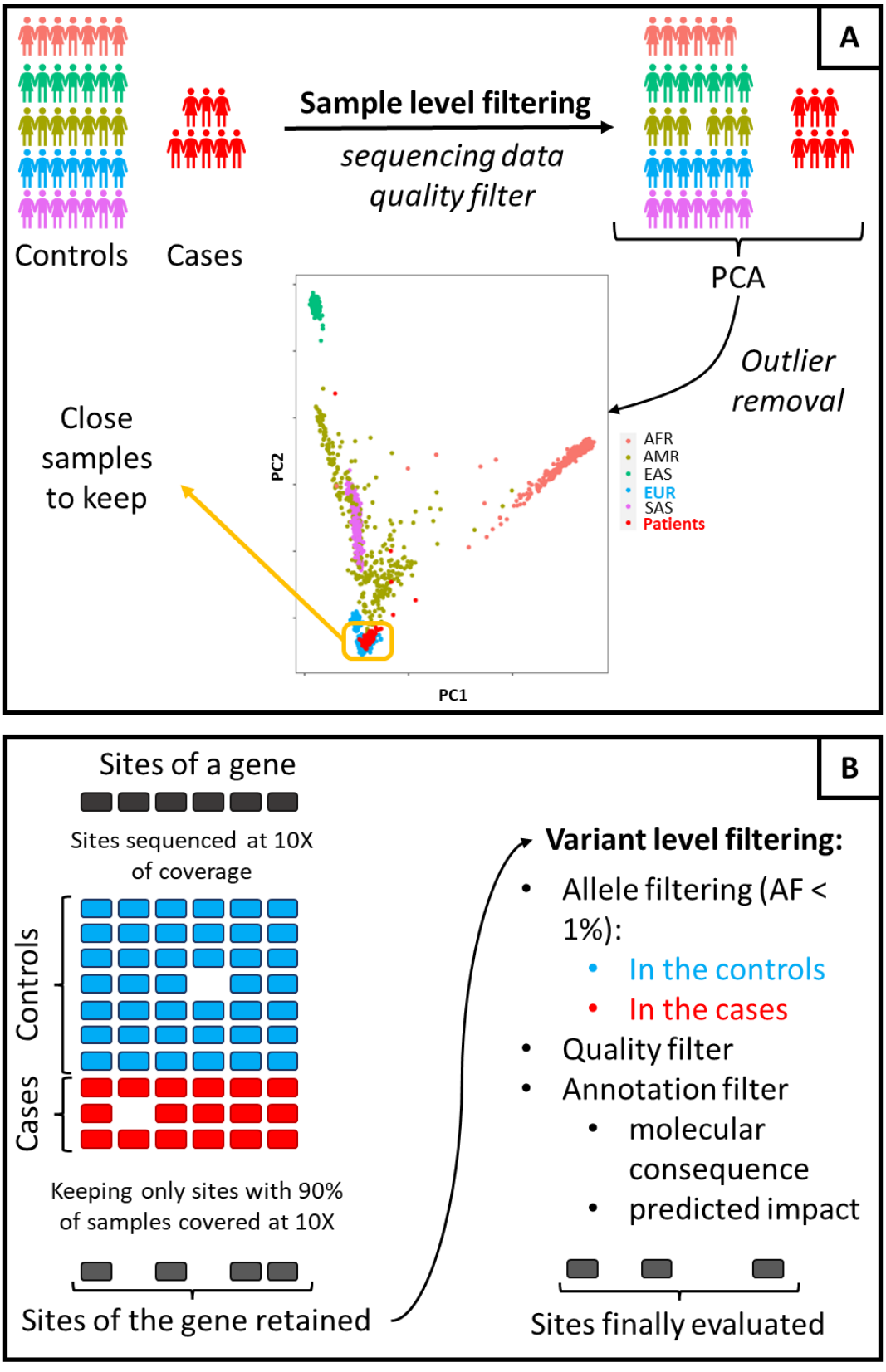



*Figure 3: The two main steps of data pre-processing for burden tests*
*Description of the main steps of burden test data pre-processing including sample-level filtering based on sample data quality and ethnicity (A) and variant-level filtering based on genotyping data quality variant molecular and pathogenicity predictions (B)*

### 1.3.3 Burden testing: different association tests

Once genomic data is filtered to remove potential biases, the burden tests *per se* can be applied. Several methods and statistical frameworks have been proposed, each with pros and cons. This section will be dedicated to a short summary of the most popular available methods.

#### *1.3.3.1 Classical burden test framework*

As in GWAS, the goal of the burden test is to associate the phenotype (case or control) with the genotype. The burden test is based on a regression framework. The basic equation of a linear regression for quantitative phenotypes is:

$$Phenotype = \alpha_0 + \beta \times Genotype \quad [1]$$

Where:

- $\beta$ is the coefficient of the regression and serves as a measure of the association by providing its effect size (absolute value); a p-value can be derived to assess significance with Wald test or likelihood-ratio test
- $\alpha_0$ is the intercept of the regression
- Genotype is the array of allele count value for each individual and can be assessed according to different models:
    - Additive: all variants are summed whether they are homozygous for the alternative allele (2 variants) or heterozygous for the alternative allele (1 variant) or homozygous for the reference allele (0 variant); under this model, the homozygote mutant is twice as likely to suffer from the disease as the heterozygote mutant
    - Dominant: a single allele is assumed to be enough to justify the phenotype, so both homozygous and heterozygous alternative alleles count as 1 (0 for homozygous for reference allele)
    - Recessive: 2 alternative alleles are needed to explain the phenotype, so homozygous for the alternative allele count as 1, 0 otherwise
- Phenotype is the array of quantitative values per sample

This simple model can be adapted to deal with binary outcomes, such being a “case” or a “control”. Here, the logistic regression is often used, in which the probability of a binary variable is assessed as follows:

$$logit\big(\mathbb{P}(Phenotype = Case)\big) = \alpha_0 + \beta \times Genotype \quad [2]$$

Where $logit(p) = \log\left(\frac{p}{1-p}\right)$ ; $0 \leq p \leq 1$. In the previous equation, $p$ is the probability of the outcome (*i.e.*, probability of being case) and the logit function is the logarithm of the odds (*i.e.*, the odds being the ratio of suffering from the disease or not). As can be seen is Figure 4A, the probability of suffering from a disease or not (probability of being a case or a control) can be modelled as a continuous function (inverse of the logarithm of the odds). This function is bounded between zero (*i.e.*, predicted to be a control) and one (*i.e.*, predicted to be a case). Then, the logistic regression model allows to represent the logarithm of the odds as a function of the genotype (Figure 4B). The advantage of a regression model is that it allows to accommodate co-variables, interaction terms, and complex designs (*e.g.*, mixed models with fixed effects and random effects). Moreover, it is easy to convert the logit into a probability. Here, in this hypothetical exaggerated example, a patient with a variant of homozygous alternative genotype (2 on the x axis of Figure 4B), the logit would be near 5.2 (following the dotted blue lines), which would translate in a high p-value of suffering from the disease (following the blue dotted line on Figure 4A).

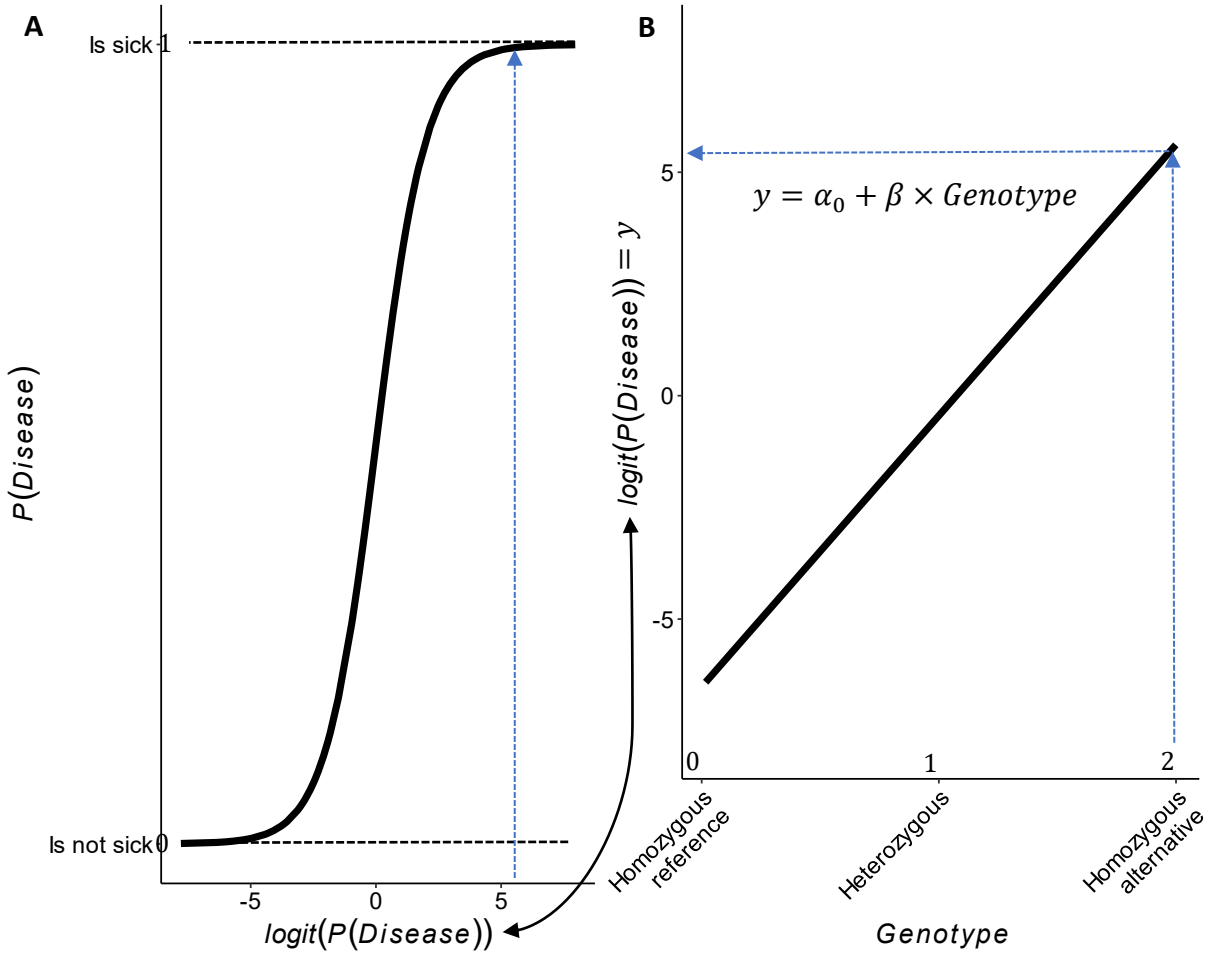



*Figure 4: Logistic regression example plots*
*Plots displaying an example of logistic regression for a case-control study with the probability of suffering from the disease (being case) to the logarithm of the odds (A) and logarithm of the odds to the genotype (B)*

Even if, supposedly, all possible confounders have been taken care of in the pre-processing step, some factors can still influence the test such as the sex of the individuals, their age which is often hard to control for or the ancestry. To accommodate to those confounders, covariates can be added to the regression:

$$logit\big(\mathbb{P}(Phenotype = Case)\big) = \\ \alpha_0 + \alpha_1 \times Sex + \alpha_2 \times Age + \alpha_3 \times PC_1 + \alpha_4 \times PC_2 + \cdots + \beta_1 \times G_1 + \cdots + \beta_m \times G_m \quad [3]$$

In the above equation, sex and age are considered in the model as well as ancestry with the first principal components of the PCA (often 10, but it can be less). Therefore, assuming $n$ samples, $m$ variants and $c$ covariates for an individual $i$, let $y_i$ represent their phenotype (where $y_i = 1$ means the sample is case and 0 otherwise), $G_i = (G_{i1}, \dots, G_{im})$ their genotypes and $X_i = (X_{i1}, \dots, X_{ic})$, the vectorized notation of the logistic regression model will be written as follows:

$$logit\big(\mathbb{P}(y_i = 1)\big) = \alpha_0 + \alpha^c \times X_i + \beta^m \times G_i \quad [4]$$

Where $\alpha^c = (\alpha_1, \dots, \alpha_c)$ is the vector of covariates' coefficients and $\beta^m = (\beta_1, \dots, \beta_m)$ the vector of regression coefficients. The covariate coefficients can be regarded as random variables. The null hypothesis of the test is that all regression coefficients are null: $H_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0$. As mentioned previously, a *sensu stricto* burden test is a test in which all the information is collapsed into a single variable. To do that, the effect of each variant needs to be computed with $S_j$, the score of the marginal model for variant $j$ for all samples $i$:

$$S_j = \sum_{i=1}^{n} G_{ij}(y_i - \widehat{\mu_0}) \quad [5]$$

Where $\widehat{\mu_0}$ is the predicted mean of the phenotype $y_i$ under the null hypothesis, *i.e.*, $\widehat{\mu_0} = \widehat{\alpha_0} + X_i \times \hat{\alpha}$ where $\widehat{\alpha_0}$ and $\hat{\alpha}$ are estimated by regressing $y_i$ on the covariates $X_i$. Thus, $S_j$ is positive when variant $j$ associated with the disease risk and negative when it is protective. The summary score statistic of each gene is then calculated as follows:

$$Q_{burden} = \left(\sum_{j=1}^{m} S_j\right)^2 \quad [6]$$

The summary score is positive or null. To compute a p-value, a $\chi^2$-test is applied with one degree of freedom. In Figure 5, 8 variants have an associated S score. Three of them are dramatically higher than the others, meaning they are truly associated with the evaluated trait and probably increase disease

risk. Overall, the $Q_{burden}$ statistic score will be high; therefore, the gene will be associated with the disease. The burden test can be used under the three assumptions for the disease mechanism: additive, dominant or recessive depending on how the number of rare variants is counted as explained previously.

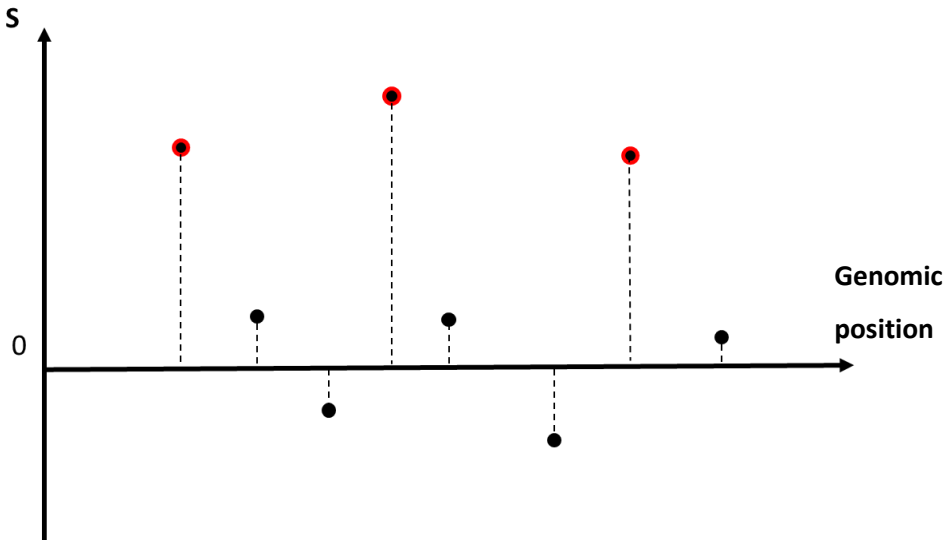



*Figure 5: Distribution of the score statistic of a gene in a weighted burden test*
*Example of a distribution of a burden test score statistic with circled red dots showing a strong association to the disease*

#### *1.3.3.2 Weighted burden test*

A simple and intuitive way to improve the classical burden framework is to add weights to the coefficients. In a linear regression framework, weights can be added to each coefficient. In a similar manner it can be added to a logistic regression model:

$$logit\big(\mathbb{P}(y_i = 1)\big) = \alpha_0 + \alpha^c \times X_i + \beta^m \times w^m \times G_i \quad [7]$$

Where $w^m = (w_1, \dots, w_m)$ are the weights of each coefficient and can be assigned in different ways:

- according to the variant MAF (*i.e.*, weight of 1 for variant below a certain threshold and 0 otherwise[94], or a continuous weight function[95]
- according to the variant pathogenicity (*i.e.*, pathogenicity score like SIFT, PolyPhen-2 or CADD or polygenic risk scores[96])

#### *1.3.3.3 Adaptive burden tests*

Several refinements to the logistic regression model have been proposed to make the method more robust to null variants or consider specific scenarios[73]. Han and Pen for example proposed a data-adaptive burden test which uses permutation to reassign the weights of the model in order to consider extreme genetic heterogeneity[97]. Other methods have been developed to refine the computation of

the weights to simply removing negatively associated variants[98], estimating the regression coefficients of the variants to use them as weights[99] or computing the weights according to the mutation patterns of each variant with a hyper-geometric kernel[100]. Although less prone to directional effects, those burden tests are most of the time computationally more intensive than classical methods, mainly due to the permutation part of the algorithm that can take time and resources.

#### *1.3.3.4 Variance-component tests*

One of the problems of the classical linear regression burden test framework is that it allows for enrichment of variants both for cases and controls. Thus, a burden test can theoretically be significant for an enrichment of both disease-risk variants and protective variants (Figure 6). To overcome this limitation, variance-components tests have been proposed.

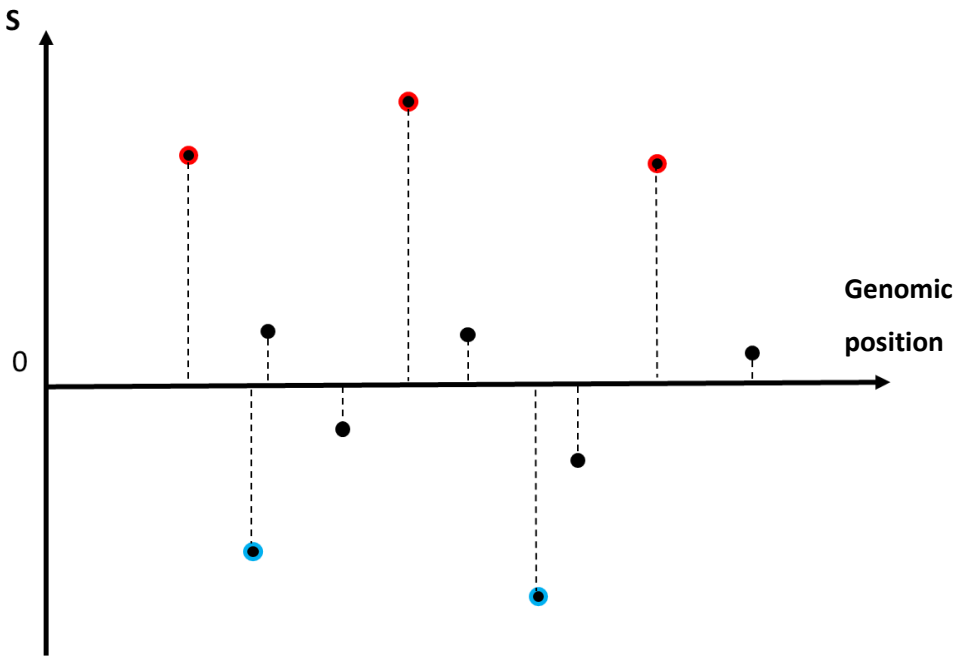



*Figure 6: Distribution of the score statistic of a gene in an adaptive burden test*
*Example of a distribution of a burden test score statistic with circled red dots showing variants with a strong association to the disease and circled blue dots showing variants with a strong protective effect*

As parameters in the burden test result from random variables (random sampling), a random effect model, or variance-component model is appropriate. This will allow to estimate the contribution of the different variables to the global variability of the data through variance. This category of test also accounts for antagonistic effects (protective and harmful variants). Variance-component tests evaluate the distribution of aggregated scores instead of testing the aggregated variants. Several tests have been developed, such as C-alpha[101] and the sum of squared score[102], but the most popular is the sequence kernel association test (SKAT)[103]. SKAT is a method based on a multiple regression model and tests the regression coefficients of the variants through a kernel association test using a variance-

component score test within a mixed-model. The null hypothesis is not the same as in a regular logistic regression: it assumes each $\beta_j \in (\beta_1, \dots, \beta_m)$ follows a distribution of mean 0 and variance $w_j \times \tau$ where $\tau$ is a variance component. SKAT is also robust to directional effects as the test statistic collapses ${S_j}^2$ instead of $S_j$:

$$Q_{SKAT} = \sum_{j=1}^{m} {w_j}^2 \times {S_j}^2 \quad [8]$$

Another advantage of this method is that it is computationally fast as compared to most adaptive burden tests.

#### *1.3.3.5 Combined tests: omnibus*

Variance-component and adaptive burden tests manage to take into account the directional effects of variants and are thus more powerful than classical logistic regression-based burden tests in such scenario. However, if no protective variants are present, they will be less powerful than classical burden tests. Such information is not available before the application of the test. Therefore, a combination of the tests would be an ideal solution. Several methods have been proposed to achieve that, the most famous being SKAT-O[104]. The latter is a linear combination of SKAT and the logistic regression burden $Q$ statistic:

$$Q_\rho = (1 - \rho) \times Q_{SKAT} + \rho \times Q_{burden} \quad [9]$$

Where $0 \leq \rho \leq 1$ and is a correlation factor between all $\beta_j \in (\beta_1, \dots, \beta_m)$. The optimal value of the $\rho$ factor is estimated by computing the minimum p-value of $\rho \times S$. Such combined tests are called omnibus. In extreme cases they can be less powerful than classical frameworks or variance-component tests. But overall, without any prior knowledge about the disease mechanism, they achieve very good results.

### 1.3.4 Limits of the case-control burden test strategy and case-only burden tests

In case-control association and especially in burden test frameworks, researchers have tried to remove confounding factors through filtering and prioritization as well as addressing power issues and directional effects. However, it may remain difficult to associate a genotype with the phenotype for rare diseases cohorts. One probable reason is that despite the dramatic decrease in exome sequencing cost (100 to 1,000-fold in ten years)[105], the number of patients suffering from a rare disease going through WES or WGS is relatively small. In addition, in most cases, no controls will be sequenced at the same time as patients. Therefore, many studies are involving controls extracted from public databases such as the 1000 Genomes Project. This may result in batch effects (*i.e.*, results explained by technical differences rather than biological factors). Hence, case-only study designs characterized by the

absence of sequencing data from a matched control cohort or from relative individuals represent a common scenario in the study of rare diseases. To overcome this limitation, case-only burden test designs have been developed over the past decade. This next section is dedicated to present existing case-only burden tests.

#### *1.3.4.1 Background mutation rate method*

To study patients with autism spectrum disorders (ASD), Samocha *et al.* focused on *de novo* mutations[42]. ASD are very heterogeneous disorders, making the detection of damaging variants particularly complicated, even within *de novo* variants. Thus, to assess the causality of genes, they proposed a method based on the modelling of the background distribution of variants in each gene. As stated in their work, multiple studies had already identified a significant excess of *de novo* LoF variants in ASD patients in a restricted number of genes. These studies lacked the evidence to implicate genes in ASD. As the background mutation can greatly differ between genes, it is hard to know if the number of rare variants observed in the cohort is meaningful to incriminate the gene. Nevertheless, a gene linked to a disease should carry more damaging variants than expected by chance. Therefore, Samocha *et al.* first developed a model that predicts the background distribution of rare variants at each gene (or genomic region) in the general population, which can then be used in a second stage to evaluate the eventual excess on mutation *de novo* rates in a case cohort. This model allows to determine if a gene carries more *de novo* variants in a cohort than expected by the neutral mutational model.

The model is using the sequence nucleotidic context as it has been determined that estimating the mutability of a base at the centre of a trinucleotide is more precise than a single base[106]. Each trinucleotide's probability to mutate into one of the three other possibilities (the three other centre nucleotides) were computed by tallying the number of occurrences of the mutation in orthologous Human/Chimp intergenic regions of the 1000 Genomes Project, considering the chimp allele as the ancestral. Using intergenic regions allows to minimize the impact of selection. The probabilities of mutation are then summed per gene for four mutation categories: synonymous, missense, nonsense, and splice site. Then the model has been adjusted for several confounders:

- Coverage, as the probability to observe a *de novo* variant is highly dependent on the capacity to call them
- Divergence between humans and primates, as the whole model is built on orthologous genomic alignments
- Replication timing, which has been associated with mutation rate[107]

In order to apply this framework to a real setting, the authors also adjusted the probability with the number of patients in the study. Therefore, for a given type of mutation, significance of the observation is assessed with Poisson distribution probabilities parametrized by the sum of the probabilities, adjusted by the number of samples. Using ESP data, Samocha and colleagues showed that specific genes had a deficit in missense variants as compared to the expectation from the model which is consistent with purifying selection's evolutionary constraint. The model allowed to detect genes enriched in *de novo* LoF variants in small genes for which a case-control test would have been underpowered.

Intuitively, the longer a gene, the higher its probability of carrying a *de novo* variation. This has been verified in Samocha and colleagues' work. The correlation between the number of rare synonymous variants in the National Heart, Lung, and Blood Institute Exome Sequencing Project (ESP) patients and the gene length was almost as high as Samocha *et al.*'s mutational model: $r = 0.880$ for gene length and $r = 0.940$ for the mutational model[42]. Thus, the gene length by itself appears as a good proxy to estimate a gene's susceptibility to mutate.

#### *1.3.4.2 Filtus*

In 2016 FILTUS, a software aiming at searching Mendelian disease-causing variants was released[108]. It allows users to filter annotated genomic variant files, detect *de novo* variants but also provides a statistical model to evaluate the gene enrichment in rare variants. This test relies on a method developed by Zhi and Chen and consists in a binomial test parametrized with the total number of mutations and the number of candidate genes[109]. Let $n$ be the number of unrelated patients, $M$ the number of candidate genes and $C$ the count matrix. $C_{ij}$ is the count of rare variants at gene $j$ for individual $i$. The null hypothesis is that gene $j$ is not associated with the studied disease, thus all variants $m$ are random variants unrelated to the disorder. By assuming each gene is of average length, the probability that a variant fall on that gene would be $p = \frac{1}{M}$. The test statistic is thus computed as follows:

$$T = \sum_{i=1}^{n} X_i \sim Binomial(n \times m, p) \quad [10]$$

This is obviously a simplistic model. It can be adapted for gene length: by assuming that a gene is $w$ times the average length, the probability becomes $p = \frac{w}{M}$. Therefore, in an additive model, the test statistic $T_a$ is computed as previously described by multiplying the probability of mutation with the gene length factor $w$. To compute the test statistic $T_d$ for a dominant model, as $m \ll M; 1 - \frac{m}{M} \approx 1$ for a Bernoulli variable, the probability can empirically be approximated by $p \approx \frac{m}{M}$. For a recessive

model, the test statistic $T_r$ can be computed with a Bernoulli variable that can be approximated by $p \approx \frac{m \times (m-1)}{M^2}$. The test is applied on a gene level $M$ times, thus a multiple testing correction is necessary. The test is corrected by Bonferroni correction.

The test makes very strong assumptions: that mutations occur randomly in a gene (thus depending only on the gene length), and that no Linkage Disequilibrium (LD) takes place between rare variations, and it does not consider populational biases nor evolutionary constraints. Highly conserved regions would be less tolerant to variations and thus small number of variants should be sufficient to detect a gene significantly enriched in rare events whereas a gene in a less conserved region would require a higher number of variants to be successfully identified. Moreover, even if FILTUS allows users to filter the data and prioritise variants of interest, the burden testing is not considering technical confounding factors such as coverage depth.

#### *1.3.4.3 TRAPD*

Large-scale publicly available sequencing databases have grown extremely big in recent years as mentioned in previous sections. Guo *et al.* have worked on a burden test using aggregated data from such databases as controls[81]. The method uses a big WES database, ExAC or gnomAD, to extract samples allele counts and perform a Fisher's exact test per gene with a contingency table comparing the number of mutated individuals between cases and controls. TRAPD includes a full pipeline to perform quality filters, pathogenicity filters/prioritising of variants, individual pruning (by studying ancestry) and finally gene-based burden testing. Most of the steps of the pipeline are based upon gold standard filters for rare variants case-control burden testing that have been described in previous sections. Interestingly, the pipeline includes an adjustment for read depth to harmonize regions that are well-covered in cases and controls (10X of coverage in 90% of the samples). The pathogenicity filters include a MAF filter, and a molecular consequence filter based on the protein prediction. In gnomAD, the user can choose to use the global MAF or a subpopulation MAF whereas in cases, Guo and colleagues propose to perform a PCA on common variants (MAF > 5%) to evaluate the ancestry and remove outlier samples.

The major issue that this method is facing is that no individual data is provided in gnomAD v2.0.2 (which was the last version of gnomAD in 2018 when the study was published). Since then, gnomAD v2.1.1 was released, including more patients as well as gnomAD v3.1 where individual data is available. However, individual data is only available for GRCh38. To overcome the unavailability of individual data, TRAPD defines qualifying variants that are variants meeting all quality and pathogenicity criteria. The test is then performed gene-wise and compares different parameters for cases and controls summarized in Table 1.

*Table 1: Summary table of TRAPD parameters for dominant and recessive tests*
*The parameters displayed are tallied for each gene*

| | Dominant test | Recessive test |
|---|---|---|
| **Cases** | Sum of the number of cases carrying at least one qualifying variant | Sum of the number of cases carrying at least two qualifying variants |
| **Controls (ExAC or gnomAD)** | Sum of the allele counts of each qualifying variant | Sum of the number of individuals carrying homozygous variants and the estimated number of samples carrying homozygous variants at each qualifying variant |

For cases in the dominant test, the number of samples carrying at least one qualifying variant is assessed whereas for the recessive test, the number of cases carrying at least two qualifying variants is computed. For controls, the number of mutated individuals needs to be approximated. Therefore, for the dominant test, the number of mutated individuals is approximated as the sum of allele counts of every qualifying variant in the gene. For the recessive test, the number of individuals carrying homozygous variants are summed at each qualifying variant and, to estimate the number of compound heterozygous samples, the cumulative frequency of individuals carrying heterozygous variants are squared and multiplied by the total number of controls. The authors underline the fact that this last approximation can be an overestimate that would make the test more conservative.

Then for each gene, a contingency table between the number of cases and controls carrying qualifying variants is built and a Fisher's exact test is applied (with Bonferroni correction). The problem with this method is that it is not *per se* a burden test in the sense that it does not measure an excess of variants for a given gene within a cohort but rather measures the number of mutated patients in the cohort compared to an approximated number of mutated patients in the global population. In the case where most cases would accumulate several damaging variants, they would still count only for one. Therefore, a gene under balancing selection gathering several variants in gnomAD but multiple damaging variants in the cases would not be detected by the test.

#### *1.3.4.4 CoCoRV*

The most recent framework to prioritising genes through a burden test from aggregated control data was released in 2022. Chen and colleagues proposed COnsistent summary COunts based Rare Variant burden test (CoCoRV): a method using genotype summary counts from gnomAD as a proxy for controls[82]. CoCoRV is very similar to TRAPD, following the same pipeline including filtering and an analogous statistical test with additional improvements in false positive control: the test takes LD into account and reassesses the expected counts with a better estimation of inflation factors. The framework follows an analogous pipeline to TRAPD but adds the possibility of removing from the

analysis a blacklist of specific problematic genomic regions. This blacklist was generated with gnomAD v2.1 filtering data by applying 3 variant filters:

- Variants that failed both WES and WGS data quality controls
- Variants failing QC in either WES or WGS platform and were absent in the other
- Variants with a substantial AF difference between WES and WGS platforms

Additionally, CoCoRV proposes two different statistical tests: a Fisher's exact test (similarly to TRAPD), and a Cochran–Mantel–Haenszel (CMH) exact test. CMH is analogous to Fischer's exact test on stratified data. Here, it is used to integrate systematic inflations due to the ethnicity-stratified data. The method can be used under a dominant and a recessive model. In a dominant model, a sample is counted when the allele count of a variant in the gene is at least 1 and at least 2 in a recessive model. The same problem as in TRAPD applies for the lack of individual data in gnomAD, hence, the same solution is used to estimate the counts by assuming Hardy-Weinberg Equilibrium (HWE) and computing the probability of the dominant or recessive model. Let's assume that a gene has $m$ variants and the genotype is coded $0, 1, 2$ for homozygous reference, heterozygous and homozygous alternative, let $p_{iG}$ be the frequency of the genotype $G$ for variant $i$. The probability of the dominant model is modelled as:

$$p_{Dom} = 1 - \prod_{i=1}^{m} p_{i0} \quad [11]$$

The probability of the recessive model is then:

$$p_{Rec} = \sum_{i=1}^{m} p_{i2} \times \sum_{j \neq i} p_{j0} + p_{2Het} \quad [12]$$

Where $p_{2Het}$ is the probability of being double heterozygous (as two variants could be on the same haplotype, this is not exactly the same as compound heterozygous) and computed as follows:

$$p_{2Het} = \sum_{i=1}^{m-1} \times \sum_{j=i+1}^{m} \times \sum_{k \neq i, k \neq j} (1 - p_{i0}) \times (1 - p_{j0}) \times p_{k0} \quad [13]$$

In these calculations all variants are assumed to be independent. This changes when LD information is incorporated. Thus, another novelty of CoCoRV over TRAPD is the identification and exclusion of high-LD variants to reject false positives. These high-LD variants were detected using gnomAD and are available for users. As the method highly depends on the count of variants from gnomAD, LD is a very important factor that needs to be controlled. When qualifying variants in control samples are in LD, only the variant with the highest AF is considered for the test. CoCoRV also provides an adjustment to

the usual inflation factor: the simulated null p-values are usually derived from a uniform distribution, but this is no longer accurate for discrete count data. This would lead to an underestimate of the inflation of the data. The inflation factor proposed by Chen *et al.* is based on an empiric resampling method of the data to construct a distribution of p-values for each gene based on the resampling. For each gene, a random number of cases with rare alleles is sampled through a hypergeometric distribution and the p-value of the test is computed. The process is repeated $N$ times (1,000 in the example), p-values are sorted according to their ranks and the average of each rank is selected to form a simulated distribution of expected p-values. The empiric factor can then be estimated the same way as in TRAPD by regressing the sorted lower 95% quantile $-log10(p-values)$. The slope of this regression becomes the empiric inflation factor (called $\lambda_{95}$ in Guo *et al.*'s study and $\lambda_{emp}$ in Chen *et al.*'s). The final improvement of CoCoRV over TRAPD is their False Discovery Rate (FDR) computation that is based over the resampling method, making it more accurate with regards to their null estimation. Two resampling methods are proposed: point estimate and upper limit estimate. The rationale is to adjust the p-values based on the simulated p-values under the null hypothesis. Let $p$ be the threshold value, $R^*(p)$ be the number of genes defined by $p \leq p-value_{simulated}$, with mean $M^*(p)$, $Q^*_\beta(p)$ its $1-\beta$ quantile and $R(p)$ defined by $p \leq p-value_{observed}$. The difference between point estimate and upper limit estimate is the use of the threshold value $s(p)$: the condition for the point estimate is $s(p) \geq Q^*_\beta(p)$ and $s(p) > 0$ for upper limit.

$$p-value_{adjusted} = \begin{cases} E_{R^*} \times \left(\frac{R^*(p)}{R^*(p)+s(p)}\right); if\ condition\ fulfilled \\ E_{R^*} \times (R^*(p) > 0)\ otherwise \end{cases} \quad [14]$$

Where $E_{R^*}$ is the expectation over all null p-values replicates.

## 1.4 Phenotype-driven methods of prioritisation

The idea to use clinical and consequently phenotypic information to identify disease-related genes is not new among bioinformatics strategies (GWAS for example). However, in the past, computers were not as omnipresent in hospitals as they are nowadays. Physicians are filling medical reports on computers on a daily basis. This provides a huge opportunity for data scientists: with the increasing performance of machine learning methods for text mining, over 24,600 diseases were described in OMIM in 2018[110]. The symptoms of multiple disorders are obviously alike but were not always described the same way. To avoid this and simplify the disease identification for both researchers and physicians, a standardized vocabulary was released in 2008: the Human Phenotype Ontology (HPO)[111].

### 1.4.1 Human Phenotype Ontology

#### *1.4.1.1 An ontology for phenotypic abnormalities*

The HPO is an ontology data model (*i.e.*, a data model which gathers concepts and links between them to represent the knowledge a given field) that takes the form of a Directed Acyclic Graph (DAG): a directed graph (*i.e.*, a graph made up of vertices linked by directed edges) without directed cycle (*i.e.*, vertices forming a closed loop), meaning it can and has to be topologically ordered. The Gene Ontology (GO) is perhaps the most famous ontology which extensively describes molecular functions, biological processes, and cellular locations. HPO is similar to GO but gathers concepts showing phenotypic abnormalities that can help describe symptoms of various disorders. Parent terms of a child term are less specific, links between terms represent "is a" relationships: a patient annotated with a given term is implicitly also annotated with all its ancestor terms. In other words, the deeper a term is in the ontology, the more precise it is.

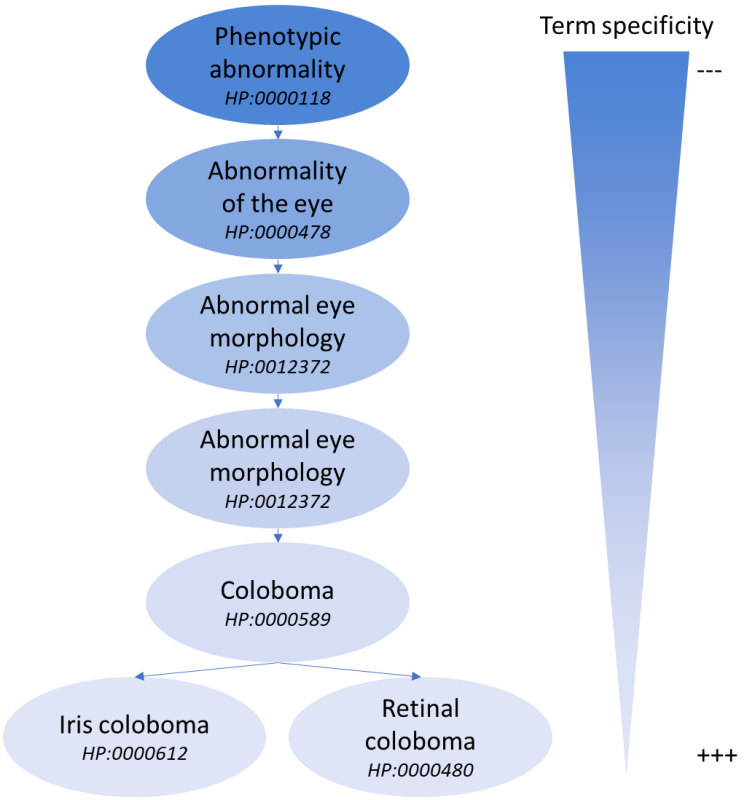



*Figure 7: Example of HPO term specificity*

A child term can have multiple parents (direct ancestors). In Figure 7 for example, a patient annotated with "retinal coloboma" would implicitly be annotated with all its ancestor terms, namely "coloboma", "abnormal eye morphology" etc which are less specific terms. This is "true-path rule" applied to all HPO terms and is based upon GO[112]. This is one key advantage of the HPO framework: clinicians do not need to exhaustively write all symptoms of a patient; a few symptoms should be able to summarize

the whole phenotype. HPO is updated and corrected on a regular basis and covers all phenotypic abnormalities recorded in OMIM and occurring at least once. HPO gathered more than 13,000 terms covering over 156,000 annotations to described hereditary disorders[113]. As HPO terms were created based on OMIM disorder descriptions, all of them are linked to at least one OMIM disease. Orphanet identifiers have been added as well (*cf.* Figure 8). Researchers identify gene-disease relationship which are also incorporated into the HPO web interface, allowing for phenotype-genotype relationship with HPO as will be discussed in the next sections.

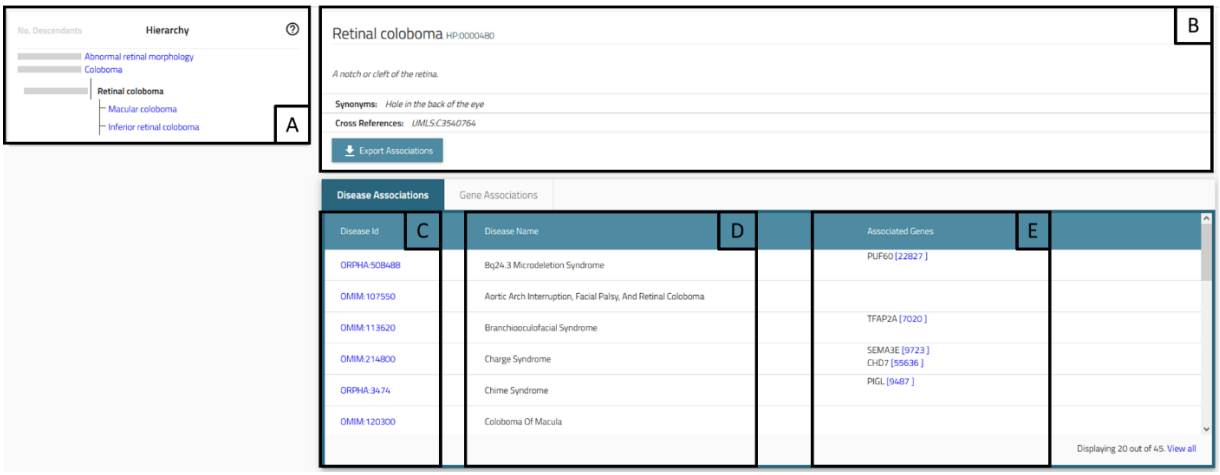



*Figure 8: Example of the HPO web interface*
*Screenshot of a search on the HPO website (https://hpo.jax.org) for "retinal coloboma" providing the close hierarchy of the term (A); the phenotypic description of the term (B); the OMIM and Orphanet known associated disorders codes (C) and names (D); the identified associated genes (E)*

#### 1.4.1.2 Information content and phenotypic similarity

The biggest interest of HPO is that every term has an identifier like in GO. This allows users to process the data computationally. In a DAG, to evaluate how informative a concept is, the notion of Information Content (IC) can be used. For a term $t$, it is computed as the negative logarithm of its frequency in the corpus:

$$IC(HPO_t) = -\log(p_t) \; where \; p_t = frequency \; of \; term \; t \begin{cases} in \; the \; ontology \\ in \; the \; dataset \; or \; a \; database \end{cases}$$

Where $f_t$ is the frequency of term $t$, which can differ depending on the used corpus. The most common way to compute the IC is to consider all HPO terms and count the number of OMIM diseases that are annotated with a term. The rationale is that terms that are used to annotate few diseases are more informative than terms annotated in many diseases. And the closer to the root of the HPO, the more general and thus more used terms should be. Using the same intuition, another way to compute the term frequency for the IC is to count the number of descendants a term has. This removes the bias of already known diseases: suppose you study a cohort of patients suffering from an unknown disease, you may still be able to have known symptoms, but they might not be linked to a close OMIM disease.

The disadvantage of this method is that because of the way a DAG is built, there are far more HPO terms at the end of the branches than at the root. Another solution is also to take solely terms that are used in the study cohort and compute the frequency based on the subset of HPO terms. Whatever the method, the rarer the term is within the corpus, the lower its frequency, the higher the IC and thus the more informative the HPO term.

### 1.4.2 Semantic similarity computation

The motivation for calculating the information content is to compute semantic similarity: the similarity between two terms. Computing several semantic similarities between couples of terms in two lists and then summarizing the information is the way to assess the similarity between two sets of HPO terms such as the one of a patient and the list of HPO terms of a disease. The semantic similarity between two terms is the IC of the Most Informative Common Ancestor (MICA, *i.e.*, the IC of the term ancestor that is the closest to the two terms in the HPO). For example, in Figure 7 the MICA of retinal (*HP:0000480*) and iris coloboma (*HP:0000612*) is coloboma (*HP:0000589*). Then, to compute the semantic similarity between two terms in a DAG, several methods have been developed. The simplest and most widely used was developed in 1995 by Resnik[114]:

$$sim_{Resnik}(t_1, t_2) = IC\left(MICA_{t_1,t_2}\right) ; sim_{Resnik} \in [0, +\infty[ \quad [15]$$

The higher the IC of the MICA, the highest the Resnik score and thus the more similar HPO terms are. Lin brought an improvement to this semantic similarity computation was to normalize the score so that the upper bound is zero[115]:

$$sim_{Lin}(t_1, t_2) = \frac{2 \times IC(t_{MICA})}{IC(t_1) + IC(t_2)} ; sim_{Lin} \in [0,1] \quad [16]$$

On top of this last idea, Jiang and Conrath proposed a weighted path approach for ontologies that takes into account the individual IC of the evaluated nodes and that performs better than Lin's computation[116]:

$$sim_{JC}(t_1, t_2) = \frac{1}{1 + IC(t_1) + IC(t_2) - 2 \times IC(t_{MICA})} ; sim_{Lin} \in [0,1] \quad [17]$$

Li and colleagues proposed another semantic similarity called Information Coefficient score for GO[117]. This score is based upon Lin's score and takes into account the structure of the DAG in addition to the IC. The number of child nodes to each node has an impact on the score: the more children a node has, the lower the score:

$$sim_{IC}(t_1, t_2) = sim_{Lin}(t_1, t_2) \times \left(1 - \frac{1}{1 + IC(t_{MICA})}\right); sim_{IC} \in [0,1] \quad [18]$$

To the best of my knowledge, the most recent semantic similarity computation score was proposed in 2019. Emission-Reception Information Content (ERIC) is an improvement to Resnik's score to be more resilient to noise and imprecision data (*i.e.*, noisy and imprecise HPO terms in a clinical record)[118]:

$$sim_{ERIC}(t_1, t_2) = \max[0; 2 \times IC(t_{MICA} - \min(IC(t_1), IC(t_2))] \in [0, +\infty[ \quad [19]$$

As for Lin, Jiang-Conrath and Information Coefficient scores have a penalty for imprecise terms (*i.e.*, terms used to describe a phenotype and for which a more precise child term exist), thus, Resnik and ERIC scores are supposed to be more robust to imprecise annotation. The authors showed in a simulated dataset that ERIC score is also more robust than all other scores in presence of noise (*i.e.*, random HPO terms that are not linked to the studied disease). There exist several other scores, but the five aforementioned scores seem to be the most common in HPO-related studies.

As previously stated, patients' health records can be summarized as lists of HPO terms as well as disorders and genes for which established relationships exist. To compute the similarity between HPO lists, a similar approach as the one developed for protein semantic similarity within GO can be applied[119]. Three protein semantic similarity approaches are proposed:

- The maximum

$$sim_{max}(A, B) = \max_{t_1 \in HPO(A), t_2 \in HPO(B)} (\mathrm{sim}(t_1, t_2)) \quad [20]$$

- The average

$$sim_{average}(A, B) = \mathrm{mean}_{t_1 \in HPO(A), t_2 \in HPO(B)} (\mathrm{sim}(t_1, t_2)) \quad [21]$$

- The Best-Match Average (BMA)

$$sim_{BMA}(A, B) = \frac{mean_{t_1}(\max_{t_2}(sim(t_1, t_2))) + mean_{t_2}(\max_{t_1}(sim(t_1, t_2)))}{2}; \quad t_1 \in HPO(A), t_2 \in HPO(B) \quad [22]$$

The maximum method has the advantage of supposedly getting rid of the noise as only the HPO terms with the highest IC of each list between A and B will be used. However, if this term is by chance part of the noise, the final result is completely biased. Using the average method allows to bypass this limitation at the cost of highly blurring the signal with all the potential noise contained in the list. The BMA takes advantage of both previous methods: it computes all the possible maxima of all pairs of

HPO terms and then averages it in both directions (*i.e.*, if the best match of $HPO_a$ within list A towards list B is $HPO_b$ not necessarily the same the other way around: the best match of $HPO_b$ can be $HPO_a$'). The BMA is averaging two asymmetric semantic similarity computations defined as follows for list A against list B:

$$sim_{asymmetric}(A \rightarrow B) = \frac{1}{|A|} \times \sum_{\mathrm{a \in A}} \max_{\mathrm{b \in B}} [sim(t_a, t_b)] \quad [23]$$

Then the opposite is computed, and the BMA can be assessed as follows:

$$sim_{BMA}(A, B) = \frac{sim_{asymmetric}(A \rightarrow B) + sim_{asymmetric}(B \rightarrow A)}{2} \quad [24]$$

In Li's article for ERIC score, they propose to use the Best-Match Sum (BMS) rather than the Bast-Match Average:

$$sim_{BMS}(A, B) = \sum_{\mathrm{a \in A}} \max_{\mathrm{b \in B}} [sim(t_a, t_b)] \quad [25]$$

No precise justification is provided to justify this choice. The rationale can be that as ERIC score is not prone to noise and imprecision, only the best scores should be considered despite the potential presence of imprecise and noisy terms. In either case of semantic similarity computation, the higher the score, the closer the two HPO lists.

### 1.4.3 Semantic similarity evaluation

Several studies have presented a comparison of the semantic similarity scores, but they are complicated to interpret and compare as the framework is usually very different as well as the results. For example, Li and colleagues studied the gene-prioritisation through phenotype-disease semantic similarity using a leave-one-out approach on all OMIM gene-disease known relationships[120]. They chose to compare Resnik and Lin score (with BMA) and several other scores, including one called Tanimoto that are not described above and that are not seen as often in the literature. They evaluated the results by comparing the methods' mean-rank ratio (*i.e.*, mean rank of the causal gene among all candidates) as well as the top-ranking genes (top1, top5…) with a True positive Rate (TPR). According to this study, the Tanimoto score performs better than Resnik and Lin score. However, in Gong *et al.*'s study, starting from 44 complex dysmorphology syndromes with detailed phenotypes, the authors created a dataset of 1,100 simulated patients and compared Resnik, Lin, Jiang-Conrath and Information Content scores to their own new score Relative Best Pair[121]. Just as in Xrare's paper, they introduced noise and imprecision and with this parameter, Relative Best Pair performed way better than Resnik score which itself performed better than Information Coefficient, Lin and Jiang-Conrath being the worst scoring method. Several other studies found that Jiang-Conrath is not performing as good as Lin

and Information Content and that Resnik performs better[119,121,122]. Not all studies measure performance with the same indicator. Furthermore, Li's Xrare study and Gong's are the only one to test the consider imprecision and noise and Xrare is the only one testing its framework on real clinical exome dataset. In the real case scenario, the performance is not nearly as good as in the simulation. Plus, Xrare is hardly comparable as it uses BMS aggregation computation whereas all other studies are using BMA. It seems hard to strictly define a better semantic similarity computation based on those studies.

### 1.4.4 Gene prioritising and semantic similarity

Most bioinformatics methods using HPO semantic similarity compute the phenotypic proximity of patients-diseases or patients-genes couples to find the closest ones. They either aim at facilitating the diagnosis by finding a suitable disease for the phenotype of the patient or prioritising genes to find a potential causal relationship with the clinical description. The next sections will be dedicated to a brief description of some of the most famous software programs utilising HPO and semantic similarity computation.

#### *1.4.4.1 Phenomizer*

Köhler and Robinson who worked on the HPO framework developed Phenomizer to simplify the diagnosis of patients. The aim of Phenomizer is to help physicians by computing p-values of patients-to-disease semantic similarity. Users enter a list of phenotypic description in the form of HPO terms and the software returns a list of diseases from OMIM and Orphanet ordered by p-values[123]. The semantic similarity computation uses Resnik score and is aggregating the score by BMA. The query (*i.e.*, the patient's HPO list) is compared to all known HPO-described diseases and the similarity scores are computed. As the number of HPO terms of the query varies, it is impossible to decide whether a score is high or not on its own, because the BMA computation highly depends on this. To overcome this limitation, authors proposed to rely on Monte-Carlo simulations: a number of HPO terms of the same size as the query are randomly drawn $n$ times and the semantic similarity scores with all diseases are computed. In theory, the majority of randomly selected lists of HPO terms should not be closer to the diagnosis than the phenotype of the patient. P-values can be computed following the Monte-Carlo random sampling by computing the number of scores that are lower than the query and are then adjusted for multiple testing by Benjamini-Hochberg correction[124]. As there are numerous different diseases and infinite number of possible queries, the computation can be extremely long. Therefore, as the software is available online through a user-friendly interface to be used by any physician or researcher (https://compbio.charite.de/phenomizer/), random computations have been pre-computed. As the number of combinations grows exponentially with the amount of query HPO terms, all possible scores have been computed and stored for all queries of one to ten HPO terms. The p-

values for more than ten terms are computed using the values for ten terms. As diseases are also linked with genes in HPO, the web interface also provides a list of associated genes. Using gene panels for Primary immunodeficiencies (PID) and knowing the causal genes of their patients, a team was able to recover 33% of them using Phenomizer[125]. One of the disadvantages of the method is that it is very sensitive, and an excessive number of candidate disorders and genes are provided by the method.

#### *1.4.4.2 eXtasy, Exomiser (PHIVE) and PhenIX*

With a random forest classifier trained on HGMD using haploinsufficiency prediction scores and gene-to-disease-gene similarity assessed with HPO, Sifrim *et al.* proposed a new way to use phenotypic information to prioritise variants with WES data[126]. Robinson and colleagues developed a collection of tools for the same aim also using HPO semantic similarity, namely, Exomiser or Phenotypic Interpretation of Variants in Exomes (PHIVE)[127] and Phenotypic Interpretation of eXomes (PhenIX)[128]. The PHIVE method which is available through Exomiser's Server (https://www.sanger.ac.uk/tool/exomiser/) filters variants based on MAF (keeping rare variants), location (needs to be in an exon), inheritance mode; then it ranks them according their pathogenicity in the gene in which they are located and the phenotypic relevance score of the latter (Figure 9). Exomiser computes a variant score using the MAF, SIFT, PolyPhen-2, and MutationTaster. A phenotypic score is then computed using semantic similarity score between OMIM-annotated diseases and the mouse equivalent to HPO, MPO (Mouse Phenotype Ontology), from the 28,176 annotated mice from the Sanger Mouse Genetics Project[129]. The rationale is to provide a lot more data than with human. Even though the phenotypic relevance score is computed for mouse genes with an ortholog human correspondence, genes are similar but not with an exact match in term of phenotype. The semantic similarity score between clinical features and mouse model is averaged across all pairwise computations between human and mouse phenotypes. The model finally provides a set of high scoring gene-disease couple. Exomiser has been very popular among researchers and physicians because of its simplicity and its completeness as it gathers both genomic filtering of variants and variant-gene prioritisation. PhenIX which was developed by the same team as PHIVE, relies on the same principle: ranking candidate genes through rarity, predicted pathogenicity of variants and clinical relevance of the genes assessed with HPO[128]. The method also takes into account the mode of inheritance of the disease (to remove incompatible diseases). It requires users to provide a Variant Call Format file (VCF) and a list of HPO terms. The difference with Exomiser is that PhenIX is not designed to identify new diseases genes for a patient but intends to help clinicians in the diagnosis process. It is even more user-friendly than Exomiser.

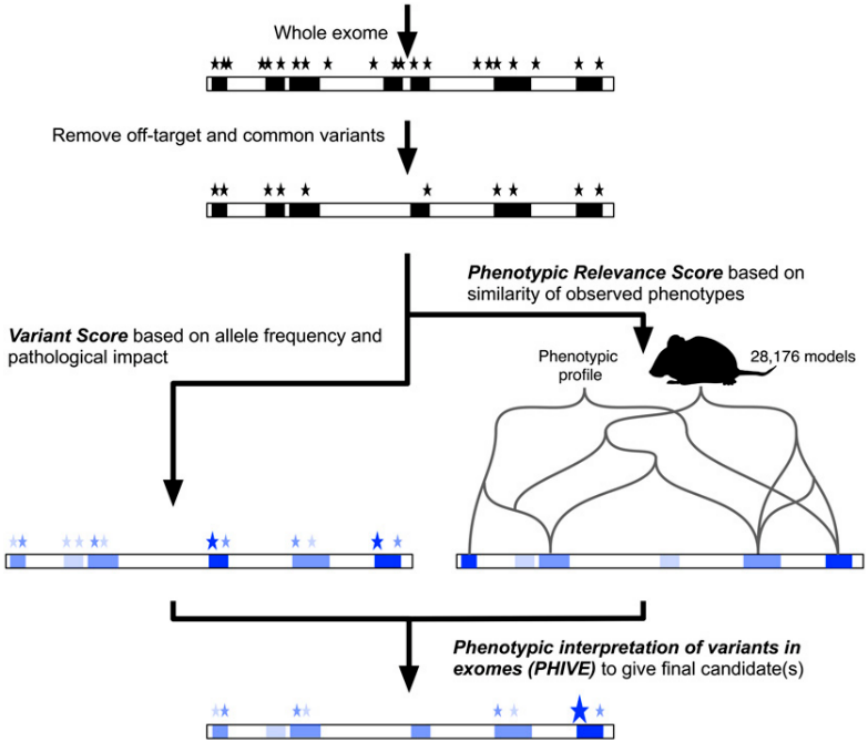



*Figure 9: Schematic representation of Exomiser's method*
*source: Robinson* et al. *(2014)*[127]

#### *1.4.4.3 Phevor*

To improve previous approaches, Singleton and colleagues proposed to combine the information brought by multiple biomedical ontologies, namely HPO, MPO, GO and the Disease Ontology (DO)[130] to calculate a disease-gene association score[131]. Researchers and physicians can thus use terms coming from either ontology database. Phevor will make a list of associated genes to the term queried by running up the ontology if the latter are not linked to any gene. Then, it iterates with the genes found in the other ontologies that were not in the query and the final gene list is saved. After that, the genes are used as starting nodes in each ontology to backpropagate and find new potentially involved genes through a so-called ontological propagation. This strategy allows to identify new candidate genes that might be involved in a particular disease. Phevor finally combines gene-scoring using variant-prioritisation tools such as SIFT and PhastCons to the previously described methodology through a final score. Just as PHIVE, Phevor is a complete method for gene and variant prioritisation. Even though it is built upon already existing knowledge of gene-disease associations, it can unveil new associations. Therefore, Phevor is using the ontologies' structure to improve the performance of gene-disease associations however it does not require information content.

#### *1.4.4.4 Xrare*

One of the most recent attempts to use HPO semantic similarity for rare-disease variant prioritisation is a software program called Xrare[118]. The latter is a machine learning approach using 51 features gathered in different categories such as population allele-frequency features (including allele frequencies from 1000 genomes, ESP, and ExAC), *in silico* prediction scores of pathogenicity (splicing prediction from dbscSNV, CADD, PhyloP…), gene-level constraints (pLI, probability of dominant gene from ExAC…) and gene-phenotype similarity scores computed with the ERIC scores developed by the same team. Xrare model uses a gradiant boosting decision tree XGBoost[132] to learn from the 49,021 pathogenic variants selected from ClinVar: 41,590 variants were used to train the model for knowledgebase, 6,576 served as positives for model training and 855 positives for model evaluation. The main feature with the highest importance in the model was one of the ERIC computation score between patients' lists of HPO terms and OMIM diseases' lists of HPO terms. In other words, the phenotypic similarity is a huge decision in the variant prioritisation. Their model outperformed Exomiser and PhenIX both in simulations and real clinical setting to classify the causal gene as top 1.

### 1.4.5 Deciphering Developmental Disorders

The UK National Health Service and the Republic of Ireland recruited 4,293 patients in the so-called Deciphering Developmental Disorders study (DDD). Those patients were suffering from severe developmental disorders (DD) but remained undiagnosed, being the only affected family member in most cases. Patients as well as both parents when available were whole-exome sequenced and deeply phenotyped with HPO terms. This allowed the consortium to detect and analyse a high-sensitivity set of 8,361 candidate *de novo* variants. Using the framework developed by Samocha and colleagues, researchers from the DDD consortium were able to identify probably pathogenic PTVs and missense variants in a set of known DD-associated genes for 33% of the patients. 93 genes significantly enriched in DNMs as compared to the expected null-mutation model developed by Samocha *et al*[42] have been identified. Among them 80 had already been demonstrated to be associated with DD. Researchers also found that females had a higher chance than males to carry a probably pathogenic DNM in DD (as had already been observed in autism). The authors explored the integration of phenotypic data to their methodology by using the semantic similarity computation between the patients' clinical records HPO terms and genes sharing DNMs. This allowed to increase the significance for some DD-associated genes (at the cost of decrease the significance of others). Their study showed that half of severe DDs could be associated with LoF and missense DNMs. They proved that using a burden-like metric parametrized by external data was a useful and successful strategy on exome data. They also provided the community with a precious database of deeply phenotyped exome-sequenced patients.

## 1.5 CONTEXT AND MOTIVATION OF THE THESIS

To this day, near 70% of patients suffering from Mendelian diseases remain without any diagnosis after whole genome sequencing[133]. There is a need to study those disorders regarding their potential genetic causes with the newest genomic and bioinformatics tools to uncover potential therapeutic strategies. However, it is an overly complex task because of both clinical and genetic heterogeneity, the complexity of the genetic architecture of such diseases and the statistical challenges they bring. Whole exome sequencing (WES) and whole genome sequencing (WGS) are much more powerful in terms of coverage than the previous genotyping methods and have improved the precision of rare disease analyses for clinical research. Precision medicine and genomics opened the door for a better understanding of the human genome and specifically disease-associated variants to identify new diagnostic, and therapeutic strategies for a growing number of rare and common diseases[134]. Around 1% of the human genome is coding proteins, collectively involving more than 19,000 protein-coding genes[135,136]. Many diseases are linked to mutations in those genes, in opposition to environmentally related diseases. Mendelian disorders might be due to *de novo* mutations, but many of them are transmitted by the parents. Collectively, all the genetic effects that can blur the signal make the diagnosis and thus drug discovery extraordinarily complex.

A frequent problem in statistical genetics is the lack of statistical power to identify relevant variants and to associate them to a correct diagnosis. Thus, to associate specific diagnoses with potential causal variants while reaching statistical significance, methods for integrative data analysis have been developed. Different solutions have been proposed as has been described in the previous sections with their pros and cons. Building on this knowledge, new strategies can be implemented with the aim of providing the community with perhaps solutions that would perform better.

In rare diseases clinical assessment of causality, many methods rely on variant and gene prioritising. Among these methods, collapsing strategies, so-called “burden tests”, have been proven successful in identifying disease-related genes with well-defined case-control cohorts. However, in rare diseases, so-called burden tests require cases and controls preferably sequenced jointly. But for practical and financial reasons, this scenario is uncommon: most of the time controls are sequenced in a different setting, introducing batch effects. To counter this, case-only frameworks have been proposed. The first method was introduced by Samocha *et al.* in 2014 and proven efficient for *de novo* variants. Then, TRAPD and earlier this year CoCoRV proposed new methods taking advantage of the ever-growing genomic public databases. Although their studies proved the usefulness of these methods to identify probable causal genes, they are not burden tests *per se*: they allow the user to count the number of patients that were mutated at a particular locus; the test does not take into account the actual

accumulation of variants as a burden test would. This can restrict the application of the method for particular cases of rare diseases with heterogeneous cohorts.

The use of HPO has been popularized among researchers and clinicians to annotate genes and to standardize clinical description of patients in electronic health records (EHR). It is now used in many clinical facilities to prioritise genes for diagnosis with the help of the Exomiser tool, Phevor, PhenIX or Xrare. Phenotype-driven analyses to assess new diagnoses in developmental disorders have already been led and have been proven successful despite all previously described difficulties. However, most methods are taking advantage of already existing phenotype-gene and gene-disease associations. For previously unknown causal genes and cohorts of heterogeneous exome-negative cases, such methods would not provide satisfactory results.

# CHAPTER 2. HYPOTHESIS AND OBJECTIVES OF THE WORK PRESENTED

## 2.1 HYPOTHESIS

Based on the state of the art and the motivation described above, the working hypotheses of this thesis are:

I. Rare genetic diseases are fundamentally monogenic or oligogenic.
II. Rare genetic diseases are characterized by varying degrees of allelic and locus heterogeneity.
III. Rare genetic diseases are driven by rare genetic variants with diverse inheritance modes including additive models, where several variants affecting diverse genomic sites on the same or different genes of an individual may accumulate their phenotypic consequences effects.
IV. The rare frequency of causal variants together with the often-small size of rare disease cohorts challenge the statistical power of statistical genetics approaches for the identification of genotype-phenotype associations.
V. Burden genetic analyses based on aggregation strategies at the gene or gene set levels may cope with the allelic or locus heterogeneity of rare diseases, accounting for additive effects and increasing statistical power rates.
VI. Allelic frequencies observed on the general population may be used to infer neutral mutation rates and to provide probabilistic estimates of the number of variants that are expected to find in a random sample of putatively healthy individuals.
VII. Clinically similar patients are likely to share common genetic factors and their associated molecular mechanisms. Thus, the identification of clinically similar patients may help prioritising genetic variants in unsolved cases.
VIII. Ciliopathies are a group of rare diseases presenting both phenotypic and genetic heterogeneity where the application of burden variant tests could help identifying novel gene-phenotype associations.

## 2.2 OBJECTIVES

Considering the previous working hypotheses, the main objective of this thesis is the development of gene-prioritising methods on rare disease patients using statistical genetics and clinical similarity approaches. Specific aims are:

I. Design a statistical test and pre-processing pipeline for rare variant burden analysis on case-only experimental designs using aggregated counts of genetic variants observed on the general population. The null hypothesis of the test being that the total amount of variants observed in

a given gene across the individuals in a cohort is not different from what is expected from a random sampling of the general population.

II. Evaluate the statistical assumptions of such approach and its behaviour on real genome sequencing data from an actual random sample of individuals from the general population, where no rejections of the null hypothesis are expected.

III. Apply the approach on a clinically heterogeneous cohort of ciliopathy patients profiled through ciliary exome-targeted sequencing in order to evaluate the capacity of the approach to recapitulate already known gene-phenotype associations in a statistically significant manner as well as to unveil potential previously unknown associations.

IV. Implement a computational pipeline to evaluate clinical similarity among rare disease patients based on a controlled phenotypic ontology.

V. Evaluate the ability of clinical similarity metrics to identify pairs of individuals sharing pathogenic genetic factors and their associated molecular mechanisms.

# CHAPTER 3.A GENE-BASED RARE VARIANTS ASSOCIATION TEST FOR CASE-ONLY RARE-DISEASE STUDY DESIGNS USING AGGREGATED GENOTYPES FROM PUBLIC REFERENCE COHORTS

## 3.1 METHODS

### 3.1.1 Sequencing data

Two *target* cohorts were considered. First, whole-genome sequencing data from non-Finnish European individuals from the 1000 Genomes project[46] Phase 3, based on genome reference GRCh37/hg19 version, were downloaded in Variant Calling Format from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. Second, ciliary exome-targeted sequencing, referred in the text as *ciliome sequencing*, was conducted in an in-house cohort of 496 patients suffering from various ciliopathies, under Ethical approval committee *Comité Éthique et Scientifique pour les Recherches, les Études et les Évaluations dans le domaine de la Santé (CESREES)*, approved on September 3rd, 2020, number #2201437. Patients included in the final study after QC represented 478 individuals with 224 females and 243 males (11 unknown), from diverse ethnicity origins. Genomic DNA was isolated from blood lymphocytes and subjected to exome capture using a custom SureSelect capture kit (Agilent Technologies) targeting 4.5 Mb of 20,168 exons (1 221 ciliary candidate genes)[137,138]. Sequencing performed on SOLiD5500XL (Life Technologies) and HiSeq2500 (Illumina) was done on pools of barcoded ciliome libraries. Paired-end reads were generated (75+35 for SOLiD, 100+100 for HiSeq) and sequences were aligned to the reference human genome hg19 with Illumina's processing software ELAND (CASAVA 1.8.2), the Burrows–Wheeler Aligner (Illumina) or mapread (SoliD). Downstream processing was carried out with the Genome Analysis Toolkit (GATK), SAMtools, and Picard Tools, following documented best practices (http://www.broadinstitute.org/gatk/guide/topic?name=best-practices). Variants were jointly called with GATK4 Haplotypecaller. As the *reference* cohort representing the general population, exome sequencing data from the Genome Aggregation Database (gnomAD) was used throughout the study (Version 2.1, hail tables available from https://gnomad.broadinstitute.org/downloads/). Only non-Finnish europeans samples from gnomAD (n= 56,885) were considered for the evaluation of the 1000 Genomes data, whereas all samples (n=125,748) were used for the ciliome analysis.

### 3.1.2 Genomic regions and sample filtering

Analyses were restricted throughout the study to human protein-coding genes mapping autosomal chromosomes as obtained from BioMart Ensembl[139] release 108 (GRCh37.p13). Protein-coding genes were defined as those containing an open reading frame. In addition, previously-described signal-artifact blacklisted regions of the human genome[140] (as provided at https://github.com/

BoyleLab/Blacklist/raw/master/lists/hg19-blacklist.v2.bed.gz) were filtered out. Blacklisted regions were assessed with bioinformatics tools based on mappability: anomalous, unstructured, high signal/read counts in independent cell-lines from ENCODE and then curated by hand. Furthermore, a gene exclusion list was considered, including 2,157 genes with highly polymorphic regions and characteristics of assembly misalignments identified in exome data from 118 individuals from 29 families[141], obtained from https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fhumu.22033&file=Table_S7_gene_exclusion_list_final.txt. For genomic coordinates considerations, only the principal isoform for each gene were considered, as provided by APPRIS database[65] based on GRCh37/hg19 and Gencode v19 (https://appris.bioinfo.cnio.es/#/downloads/). In addition, as suggested for rare-variant collapsing analyses[74], genomic regions were further restricted to those mapping so-called "trustworthy regions", defined as those covered at 10X or more in at least 90% of the samples in both the reference (*i.e.*, gnomAD) and the target cohort (either the 1000 Genomes or the ciliome cohort described above). In the case of the 1000 Genomes data, analyses were further restricted to regions passing "strict mask" filters, as provided at http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/StrictMask/, and defined as having a total coverage within 50% of the average, having no more than 0.1% of reads with mapping quality of zero, and with an average mapping quality for the position equal or greater than 56. Genes covered in less than 25% of the APPRIS Principal transcript length after intersection among 'trustworthy regions' were filtered out. Samples with genotype quality < 20, depth < 8 or < 95% call-rate (percentage of variants for which no genotype could confidently be called) were discarded. The main filtering steps of the pipeline are summarized in Supplementary figure 1.

### 3.1.3 Genetic variant annotation and filtering

Single nucleotide variants (SNVs) and indels mapping the retained genomic regions described above were considered. Throughout the study, custom variant filtering was performed with hail library[142] 0.2.57. Variant annotation was performed using VEP[61] v100.4. Only variants with Allele Frequency (AF) < 1% in the target cohort and < 0.1% in the reference cohort were retained. Only variants annotated as synonymous, missense or Protein Truncating Variant (PTV) on the APPRIS principal isoform for each gene were considered[65]. PTVs were defined as those annotated as stop-gained, splice acceptor, splice donor and frameshift variants. When a variant mapped 2 overlapping genes, the gene/transcript with the most damaging consequence was retained, with the following preference: PTV > missense variant > synonymous variant. Protein altering variants were defined as those annotated either as missense or PTV.

### 3.1.4 Goodness-of-fit Poisson tests

Let $X = \{x_1, x_2, \dots, x_n\}$ represent a sample of $n$ independent counts, and $s = \sum_{i=1}^{n} x_i$ its sample sum. Several statistical tests allow to determine whether they derive from a Poisson distribution. First, the $u$-test[143] measures the equi-dispersion of the data by determining whether the variance-to-mean ratio is equal to one. The $u$ value associated to the sample $X$ is defined as:

$$u = \left(\frac{V}{m} - 1\right) \times \sqrt{\frac{n-1}{2 \times \left(1 - \frac{1}{s}\right)}}$$

where $V$ is the sample variance and $m$ is the sample mean. P-values for the two-tailed $u$-test are assessed by comparing $u$ values to a normal distribution of mean 0 and standard deviation of 1 with a 5% significance level. P-values are then corrected for multiple testing with Bonferroni correction.

The $D$-test is an exact test evaluating the dispersion of the data[144]. The test is a two-tailed evaluation of the cumulative exact probability of the sum of squares of the supposed Poisson realizations and is derived from the Fischer $\chi^2$-test [145]:

$$D = \frac{\sum_{i=1}^{n} (x_i - m)^2}{m}$$

The $L$-test[143] evaluates the homogeneity of the data, *i.e.*, whether a set of Poisson realizations have the same parameter. The $L$-test is a one-tailed test and is based on the likelihood ratio test. The L-statistic is assessed as follows:

$$L = \sum_{i=1}^{n} x_i \times \log(x_i)$$

P-values for testing dispersion (and homogeneity) are assessed by computing the cumulative probability of the realizations of greater or equal $D$ (or $L$ for the homogeneity)[146].

The CR-test[144] compares the frequency of zeros observed in the sample with the distribution of zeros expected from the Poisson assumptions, $N_0$. Considering $n_0$, the sample number of observed zeros, one can compute the right-tailed CR-test p-value for the null-hypothesis stating that the data is Poisson distributed against the alternative hypothesis that data is zero-inflated:

$$\mathbb{P}(\mathrm{N}_0 \geq \mathrm{n}_0) = \sum_{i=n_0}^{n} \sum_{j=1}^{n} (-1)^{j-i} \times \binom{n}{j} \times \binom{j}{i} \times \left(1 - \frac{i}{n}\right)^S$$

For each test, significancy was assessed after Bonferroni correction. The code to run the previously described tests was adapted from a R shiny app available online (https://manu2h.shinyapps.io/gof_poisson/) and described in Fernández-Fontelo *et al.*[144].

### 3.1.5 Inflation factor estimation

Inflation factor estimation was performed using two alternative expectations about the p-values distribution under the null hypothesis $H_0$. First, assuming a uniform distribution of p-values. And second, empirically sampling the null Poisson distribution under the null hypothesis. Here, following Higueras *et al.*[146], for each evaluated gene, we randomly sampled the number of variants from a Poisson distribution with parameter $\lambda$, and calculated their associated Poisson p-values. We did so N independent times. For each simulation, we sorted the P-values under null across all genes and stored their rank. Finally, the expected sorted P-value at rank $k$ was assessed as the average of the p-values having ranked in position $k$ across each of the $N$ simulations. To estimate the inflation factor, we took the lower 95% quantile of points in the QQ plot and regressed the sorted log10-scaled observed P-values to the log10-scaled expected sorted P values. The slope of the regression against the raw p-values was defined as the inflation factor, denoted by $\lambda_{\Delta 95}^{\text{uniform}}$ or $\lambda_{\Delta 95}^{\text{empirical}}$, depending on whether the uniform distribution or the sampled-null distribution was used, respectively. When inflation factors were assessed upon genomic corrected p-values, the analogous figures where denoted by $\lambda'^{\text{uniform}}_{\Delta 95}$ or $\lambda'^{\text{empirical}}_{\Delta 95}$.

### 3.1.6 Genomic correction of p-values and multiple testing correction

To minimize the effects of confounding factors and especially invisible substructure of the data, genomic control was applied to minimize false positive associations. Following Devlin *et al.*[147], a $\chi^2$-test statistic with 1 degree of freedom is computed at each locus and $\lambda$ is estimated as the median of the $\chi^2$-test divided by 0.456 (*i.e.*, $\chi^2$-test with 1 degree of freedom for a p-value of 0.5). New $\chi^2$-test statistics are divided by $\lambda$ and p-values for a $\chi^2$-test statistic with 1 degree of freedom are computed at each locus. P-values are corrected for multiple testing by Bonferroni correction.

### 3.1.7 Reference databases of pathogenic variants

The relevance of the most damaging variant of patients mutated in the case-only enriched genes was assessed using two variant pathogenicity databases: ClinVar[36] or the Human Gene Mutation Database (HGMD)[148]. CADD[59] was also used to discriminate variants. The most damaging variant of patients were chosen following these criteria:

1. Presence of a pathogenic variant registered in ClinVar or a damaging variant in the Human Gene Mutation Database[149]

2. Presence of a probably pathogenic variant registered in ClinVar (pathogenic/likely pathogenic, likely pathogenic) or a probably damaging variant in HGMD (probably damaging, disease-associated polymorphism)
3. If no record existed in the previous databases, the variant with the highest CADD PHRED score was kept

An arbitrary CADD threshold of 15 was used to define pathogenicity, in accordance with the median value of all possible canonical splice sites and non-synonymous variants of CADD v1.0[150].

## 3.2 RESULTS

### 3.2.1 Formulation of the case-only gene-based rare variants burden test (COBT)

The distribution of the total number $n_j \in \mathbb{N}$ of qualifying rare variants identified in an individual $j$ across $i$ qualifying genes ($i = 1 \dots I$) within a genome may be modelled as a multinomial distribution as follows: let $n_j$ be the number of independent trials (*i.e.*, genetic variants) each leading to a success for exactly one of mutually exclusive $I$ categories (*i.e.*, genes), with each category having a given fixed success normalized probability $p_i$, where $\sum_{i=1}^{I} p_i = 1$. Let $X_1^j \dots X_I^j$ be $I$ random variables representing the non-negative integer number of such successes of each category $i$. Thus, the probability of observing any particular combination of numbers of successes $X_i^j = x_i^j$ at individual $j$ for the various genes $i$ is given by the multinomial distribution described by:

$$\mathbb{P}\left(X_1^j = x_1^j, X_2^j = x_2^j, \dots, X_I^j = x_I^j\right) = \frac{n_j!}{x_1^j! \dots x_I^j!} p_1^{x_1^j} \dots p_I^{x_1^j}$$

where $n_j = \left(\sum_{i=1}^{I} x_i^j\right)!$

Focusing on a specific gene $i$, such distribution may be simply described as a binomial distribution with parameters $n_j$ and $pi$, *i.e.*, $X_i^j \sim B(n_j, p_i)$, where:

$$\mathbb{P}\left(X_i^j = x_i^j\right) = \frac{n_j!}{x_i^j!\left(n_j - x_i^j\right)!} p_i^{x_i^j} (1 - p_i)^{n_j - x_i^j}$$

If multiple individuals are considered, the sum of variants for a specific gene $i$ across $J$ individuals, $X_i$, may be modelled as a random variable resulting from the sum of the realizations of $J$ independent binomial distributions $B(n_j, p_i)$, which behave in turn as a binomial distribution:

$$X_i = \sum_{j=1}^{J} X_i^j \sim B(n \times J, p_i)$$

where $n$ represents the median of $n_j$ across $J$.

When the number $J$ of individuals is high, such binomial distribution may be approximated by a Poisson distribution:

$$X_i \rightarrow Poisson(\lambda);\ where\ \lambda = (n \times J)\ \times p_i$$

The availability of large-scale genome sequencing projects of the general population provides the opportunity to parametrize the fixed success normalized probabilities $p_i$ from the observed distribution of qualifying variants across qualifying genes. The number of rare synonymous and missense variants is highly correlated to gene length[42]. Thus,

$$p_i = \frac{x_i^{ref}}{\sum_{i=1}^{I} x_i^{ref}}$$

where $x_i^{ref}$ represents the number of unique qualifying variants observed in gene $i$ in a reference cohort. In the present study, gnomAD was used as a *reference* cohort throughout the study (*cf.*, 3.1.1).

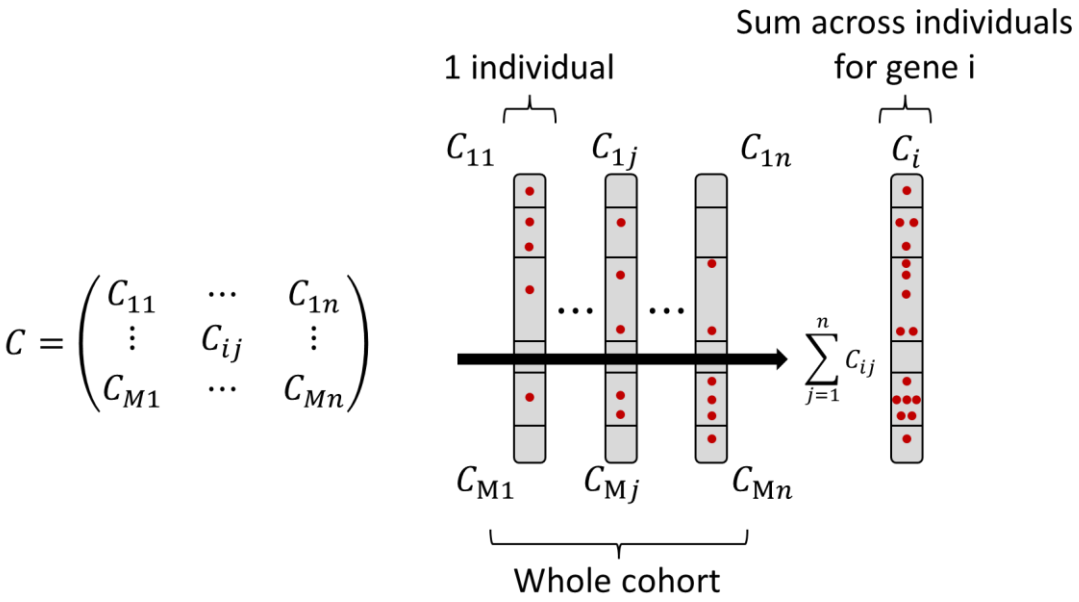



*Figure 10: Conceptual diagram of the intuition behind the case-only burden test*
*Grey rectangles represent evaluated patients as a sum of all their genes and red dots the retained variants within them (all genes collapsed independently from their chromosomal locations) to mimic the $C$ count matrix. The right rectangle is the sum of all variants retained within all patients and all genes to test each gene independently.*

The previous formalisms allow to evaluate the null hypothesis $H_0$ stating that the sum of qualifying variants on a specific gene $i$ observed across the $J$ individuals from a *target* cohort, $x_i$, is not statistically higher from what would be expected if the same number of individuals would be randomly selected from the *reference* cohort. Thus, by establishing a type I error $\alpha$, $H_0$ is rejected when

$$\mathbb{P}(X_i > x_i) = 1 - \mathbb{P}(X_i \leq x_i) = 1 - \sum_{\mu \leq x_i} \frac{\mu^{x_i} e^{-\mu}}{x_i!} < \alpha, reject\ H_0$$

$$where\ \mu = (n \times J) \times p_i\ and\ x_i = \sum_{j=1}^{J} x_i^j \frac{n}{n_j}$$

In the previous formula, qualifying variant counts $x_i^j$ are normalized to account for heterogeneous $n_j$ across individuals, and $x_i$ rounded to the closest integer number.

An additive genetic model was considered throughout this work by which $x_i^j$ represented the total burden of mutated genomic positions in gene $i$ at individual $j$. Thus, heterozygous variants within an individual contributed with one count, while homozygous variants contributed with two counts. However, as done in other rare-variant burden tests in case-control study designs[73], the case-only burden test could be readily adapted to a dominant model without loss of generality, simply by adapting the assessment of $x_i^j$ as the number of qualifying variants for which individual $j$ carries at least one minor allele.

### 3.2.2 Goodness-of-fit of the Poisson distribution for rare variants.

We first characterized the behaviour of COBT on a target cohort constituted of 404 non-Finnish European individuals from the 1000 Genomes Project (*cf.*, 3.1.1). As for the *reference* cohort, which is based on 56,885 non-finish Europeans from gnomAD, this *target* cohort represents a random sampling from the general population. As such, it serves as a negative control for the purpose of this study as, in principle, no rejections of the null hypothesis $H_0$ stated above would be expected. After stringent genomic region filtering, a total of 12,041 qualifying protein-coding genes were retained for downstream analyses (*cf.*, 3.1.2 and 3.1.3). We analysed separately the distribution of rare synonymous and missense variants across the *trustworthy* regions of such genes, with a median number per individual of $n = 80$ and $n = 135$ qualifying variants.

*Table 2: Summary of the Index of dispersion of the number of mutations per gene carrying synonymous and missense mutations in non-Finnish Europeans from the 1000 Genomes Project*

| | Min | 5th per-centile | 10th per-centile | Median | Mean | 90th per-centile | 95th per-centile | Max | Number of evaluated genes | Number of genes rejecting the U-test |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Synonymous*** | 0.940 | 0.998 | 0.988 | 0.998 | 1.021 | 1.000 | 1.158 | 3.469 | 9,266 | 3,108 |
| ***Missense*** | 0.913 | 0.975 | 0.980 | 0.995 | 1.020 | 1.000 | 1.203 | 4.211 | 10,620 | 2,931 |

First, we inspected whether the distribution of qualifying variants across individuals and qualifying genes approximated the assumption of a Poisson distribution, *i.e.*, having a variance to mean ratio close to 1. Here, synonymous and missense variants presented per-gene median values of 0.998 and

0.995, with 10th and 90th percentiles of [0.988-1.000] and [0.980-1.000], respectively (Table 2). The goodness-of-fit of the Poisson-distribution to the observed data was further statistically evaluated in terms of its dispersion, homogeneity and frequency of zeros through the D-test, L-test and CR-test, respectively (*cf.*, 3.1.4). In the case of synonymous variants, a residual 0.4%, 0.5% and 0.2% of the evaluated genes rejected the D, L and CR-test respectively. Similar figures were obtained in the case of missense variants, with 0.1%, 0.2% and 0.1% rejections (Table 3). From the previous results, it can be concluded that the per-gene distribution of rare variants across individuals may be modelled through a Poisson distribution for most of the genes, thus supporting the assumptions of the COBT approach.

*Table 3: Evaluation of the dispersion, homogeneity and zero-inflation of the per-gene distribution of synonymous and missense variants across 404 non-Finnish Europeans from the 1000 Genomes Project*

| | ***Number of evaluated genes (D-test & L-test)*** | ***Number of evaluated genes (CR-test)*** | ***Number of genes rejecting the D-test*** | ***Number of genes rejecting the L-test*** | ***Number of genes rejecting the CR-test*** | ***Number of genes rejecting the 3 tests*** |
|---|---|---|---|---|---|---|
| *Synonymous* | 6,646 | 9,266 | 24 | 30 | 23 | 17 |
| *Missense* | 8,851 | 10,620 | 12 | 18 | 16 | 9 |

### 3.2.3 Diagnosis of p-values distribution and inflation estimation

We then evaluated the distribution of p-values obtained when applying COBT on the target cohort of 404 non-Finnish European individuals from the 1000 Genomes Project. As previously pointed out, the null hypothesis $H_0$ of the COBT test is expected to be true for all genes in this target cohort. Such expectation permits to assess whether the test's p-value distribution is indeed uniform or, alternatively, to diagnose systematic p-value distribution shifts[151]. The inspection of p-value histograms showed a non-uniform distribution of p-values both for synonymous and for missense variants (Figure 11). The associated Quantile-Quantile (QQ)-plots further reflected a systematic inflation of p-values as compared to the expectations from a uniform distribution with $\lambda_{\Delta 95}^{\text{uniform}} = 1.42$ for synonymous variants and $\lambda_{\Delta 95}^{\text{uniform}} = 1.48$ for missense variants (Figure 12 & Figure 13**,** panel A). These results are compatible with recent works pointing out that the assumption of a uniform distribution of p-values is not well suited for statistical tests based on discrete count data[82,152]. Following Zhi & Chen we thus sampled the true null distribution of the discrete test statistics in order to obtain a more realistic estimation of the expected distribution of p-values (*cf.*, 3.1.5). Doing so allowed a more accurate inflation factor estimation method, yet, reflecting a systematic inflation ($\lambda_{\Delta 95}^{\text{empirical}} = 1.20$ and $1.26$ for synonymous and missense variants, respectively) and a significant deviation of p-values from the regression line (Figure 12 & Figure 13, panel B). However, applying a

genomic control of p-values[147,153] (*cf.*, 3.1.6) corrected for global systematic inflation as well as for deviation of observed p-values from the fitted regression lines (Figure 12 & Figure 13, panel C and D), and this assuming a uniform ($\lambda'^{\text{uniform}}_{\Delta 95} = 1.04$ and $1.15$ , for synonymous and missense variants, respectively) as well as an empirical sampled-null distribution ($\lambda'^{\text{empiric}}_{\Delta 95} = 0.36$ and $0.40$, for synonymous and missense variants, respectively).

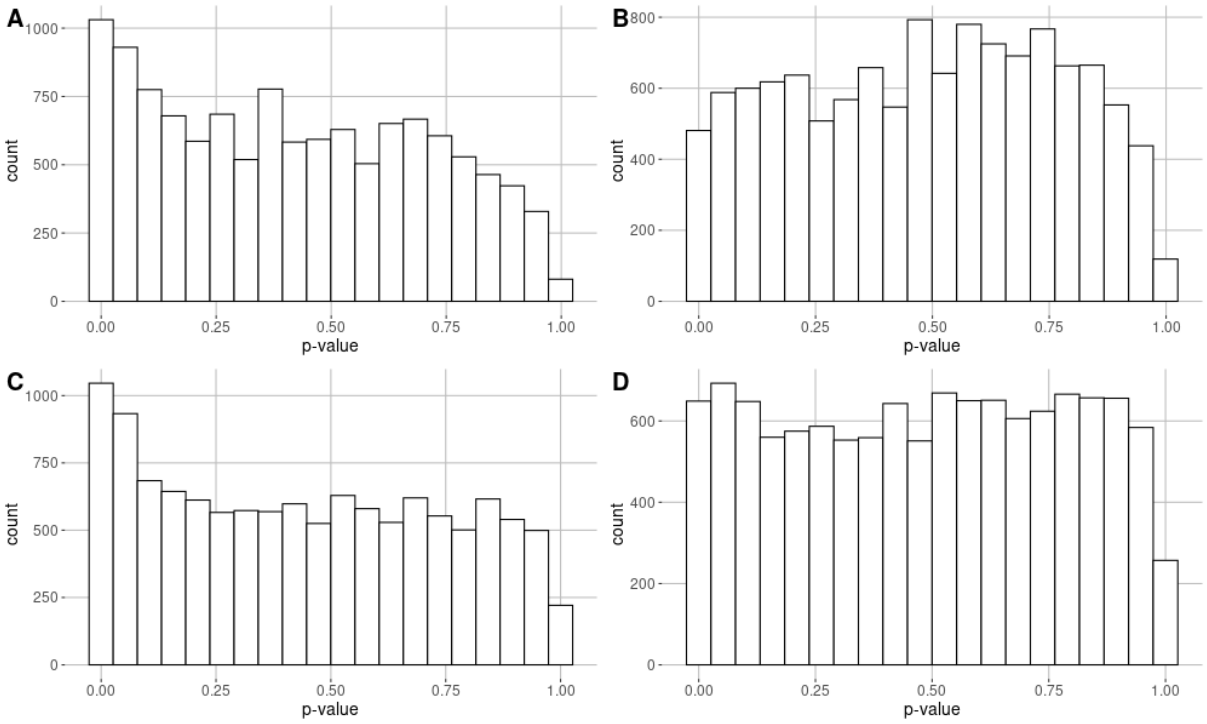



*Figure 11: Distribution of p-values in non-Finnish Europeans' genes from the 1000 Genomes Project for synonymous variants With COBT without genomic control (A) and with genomic control (B); for missense variants with COBT without genomic control (C) and with genomic control (D)*

Along with the previous analyses on the cohort of 404 non-finish Europeans, we further evaluated for comparison the behaviour of two state-of-the-art methods for the identification of gene-disease associations for case-only design studies based on rare variant population frequencies: TRAPD[81] and CoCoRV[82]. Both TRAPD and CoCoRV lead to p-value histograms showing a non-uniform p-value distribution both for synonymous and for missense variants (Figure 14 & Figure 15), although with opposite trends. Thus, TRAPD led to an excess of significant p-values, while CoCoRV led to an excess of p-values close to 1. Furthermore, the corresponding TRAPD QQ-plots reflected a severe inflation ($\lambda^{\text{uniorm}}_{\Delta 95} = 6.13$ and $8.02$ for synonymous and missense variants, respectively). Similarly, CoCoRV QQ-plots reflected as well important inflation levels, while less acute than in the case of TRAPD and in spite of being based on a sampled-null distribution ($\lambda^{\text{empiric}}_{\Delta 95} = 1.27$ and $1.29$ for synonymous and missense variants, respectively). Overall, the previous analyses on the 1000 Genomes data showed that COBT genome corrected p-values were able to control for genomic inflation and kept type I error at a low

rate, as further described in the next section. On the contrary, case-only aggregation methods such TRAPD and CoCoRV, suffered from an inflation of p-values that ultimately translated in an excess of false-positive hits, considering that no enrichments were expected for the large majority of evaluated genes in this analysis.

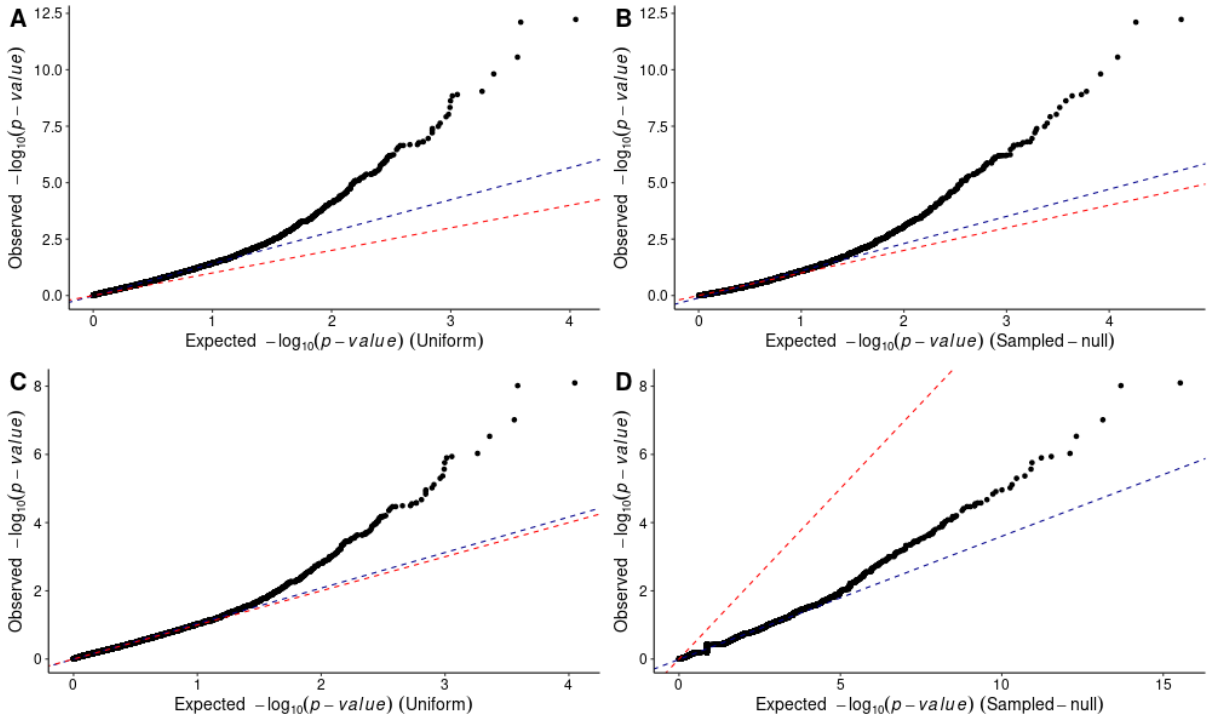



*Figure 12: COBT quantile-quantile plots in non-Finnish Europeans' genes from the 1000 Genomes Project for synonymous variants*
*With expected uniform distribution (A), expected sampled-null distribution (B), with Genomic Control and expected uniform distribution (C) and with Genomic Control and expected sampled-null distribution (D); the red dotted line is the bisector, the blue dotted line is the $\lambda_{\Delta 95}$ (uniform or empiric)*

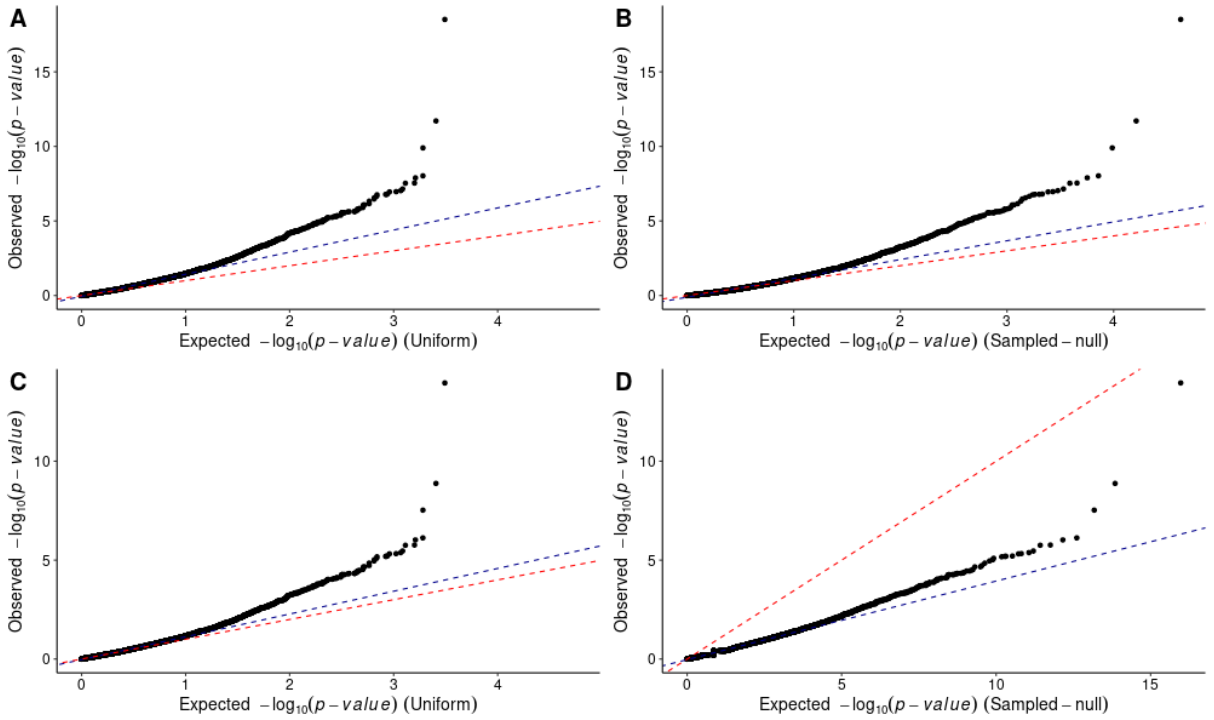



*Figure 13: COBT quantile-quantile plots in non-Finnish Europeans' genes from the 1000 Genomes Project for missense variants With expected uniform distribution (A), expected sampled-null distribution (B), with Genomic Control and expected uniform distribution (C) and with Genomic Control and expected sampled-null distribution (D); the red dotted line is the bisector, the blue dotted line is the $\lambda_{\Delta 95}$ (uniform or empiric)*

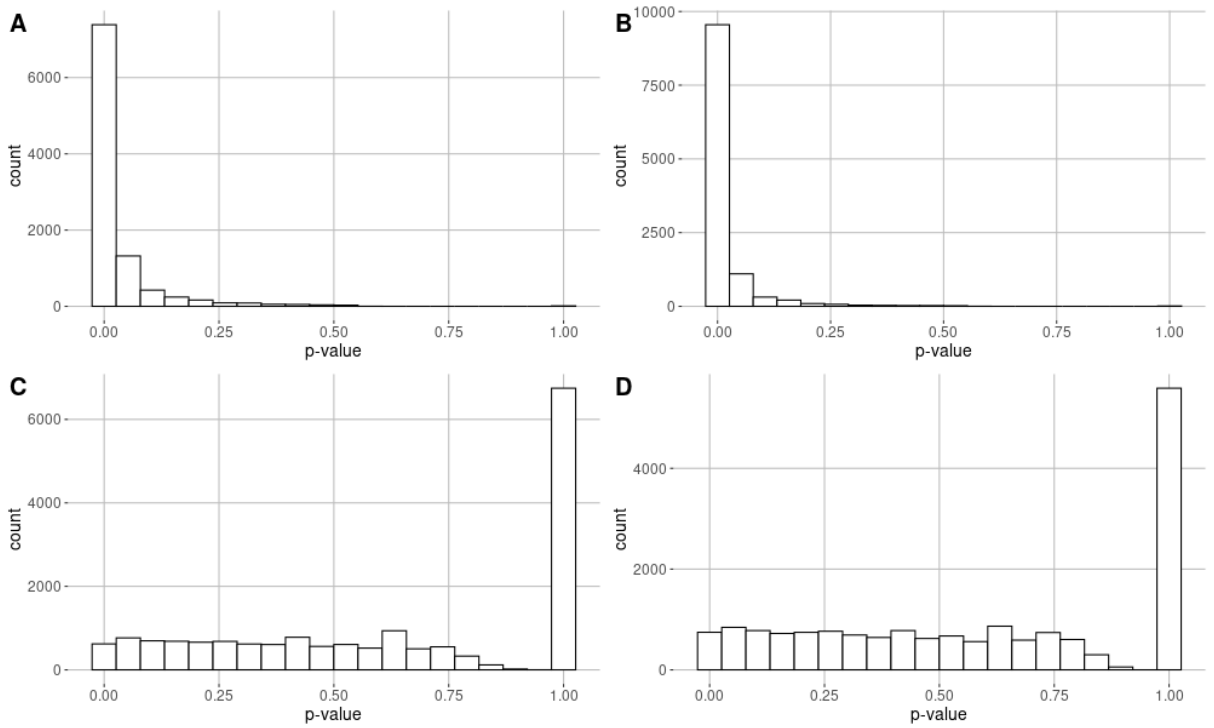



*Figure 14: Distribution of p-values in non-Finnish Europeans' genes from the 1000 Genomes Project for synonymous variants With TRAPD (A) and CoCoRV (B); for missense variants with TRAPD (C) and CoCoRV (D)*

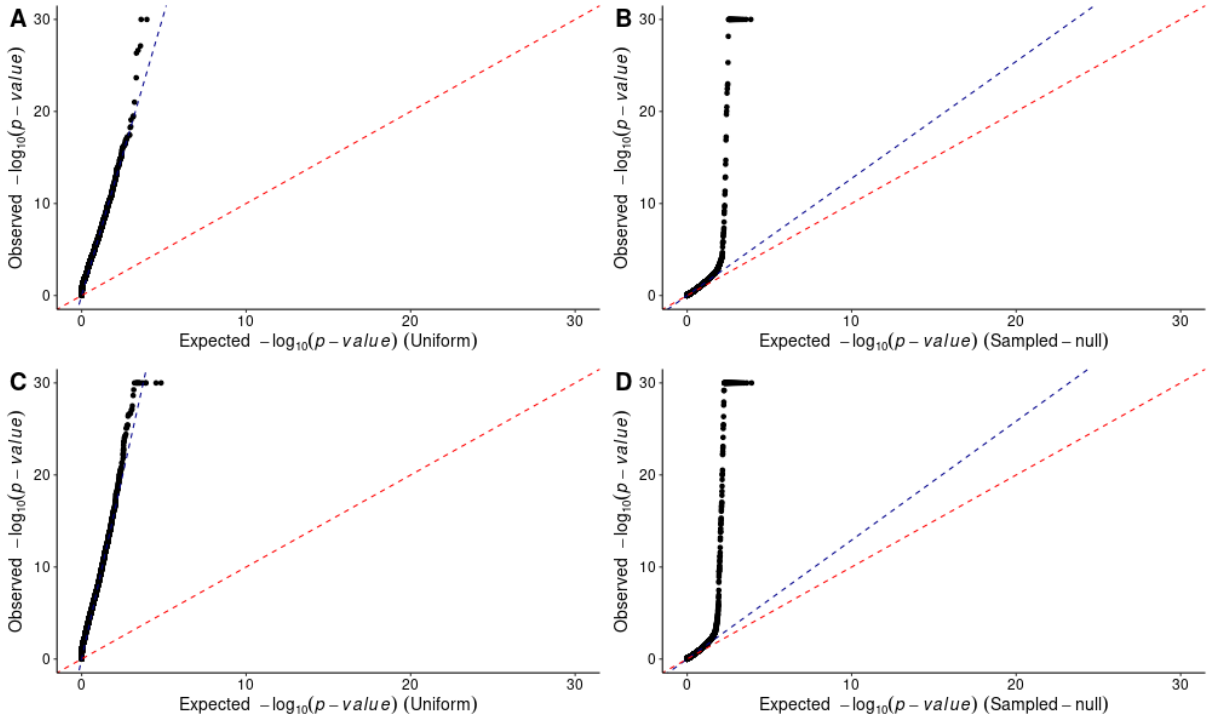



*Figure 15: Quantile-quantile plots in non-Finnish Europeans' genes from the 1000 Genomes Project for synonymous variants With TRAPD (A) and CoCoRV (B); for missense variants with TRAPD (C) and CoCoRV (D); the red dotted line is the bisector, the blue dotted line is the $\lambda_{\Delta 95}$ (uniform for TRAPD and empiric for CoCoRV)*

### 3.2.1 Analysis of gene hits on non-Finnish European individuals from the 1000 Genomes Project

After Bonferroni multiple-testing correction setting a familywise error rate (FWER) of 5%, nine (<0,01%) and eight genes (<0.08%) rejected the null hypothesis of the COBT test for synonymous and missense variants, respectively (Table 4). For those genes, rejecting the null hypothesis means that the sum of variants on a specific gene observed across the individuals from a *target* cohort (*i.e.*, non-Finnish European from the 1000 Genomes), is statistically higher than what would be expected if the same number of individuals would be randomly selected from the *reference* cohort (*i.e.*, non-Finnish European from gnomAD)**.** Further inspection of the actual numbers of qualifying variants and carrier individuals for those genes reflected a clear enrichment in variants among individuals from the 1000 Genomes as compared to the expected counts inferred from gnomAD (Table 4). Such enrichments were not merely explained by the concentration of variants in few carriers, as the ratio of variants per carrier ranged between 1 and 2 for most hits (Table 4). To better characterize the sensitivity of the test to stochastic sampling of the general population, we randomly sampled 202 out of the 404 individuals (*i.e.*, 50%) one thousand times, and applied the COBT approach on each subset. Out of the nine and eight genes initially enriched in synonymous and missense variants, respectively, only HRC appeared as significantly enriched in more than 50% of the re-samplings. However, the rest of such genes only rejected the null hypothesis less than 16% of the times. While a partial replication is expected from

the loss of statistical power associated to a lower sample size, yet the percentage of the N samplings in which hits were replicated positively correlated with the p-value observed on the total set (Wilcoxon rank sum test $p - value < 2.2 \times 10^{-16}$, Figure 16). As a further control, the enrichments found among the 1000 random subsampling were however not explained by a severe unbalance in the retained carrier individuals between the two splits at each sampling. Thus, logistic regression analysis between the 202 individuals considered in COBT and the rest of the 202 individuals left aside at each sampling (mimicking a case-control setting) did not lead to any significant hit after Bonferroni correction at a FWER of 5% in more than 1% of the resampling analysis.

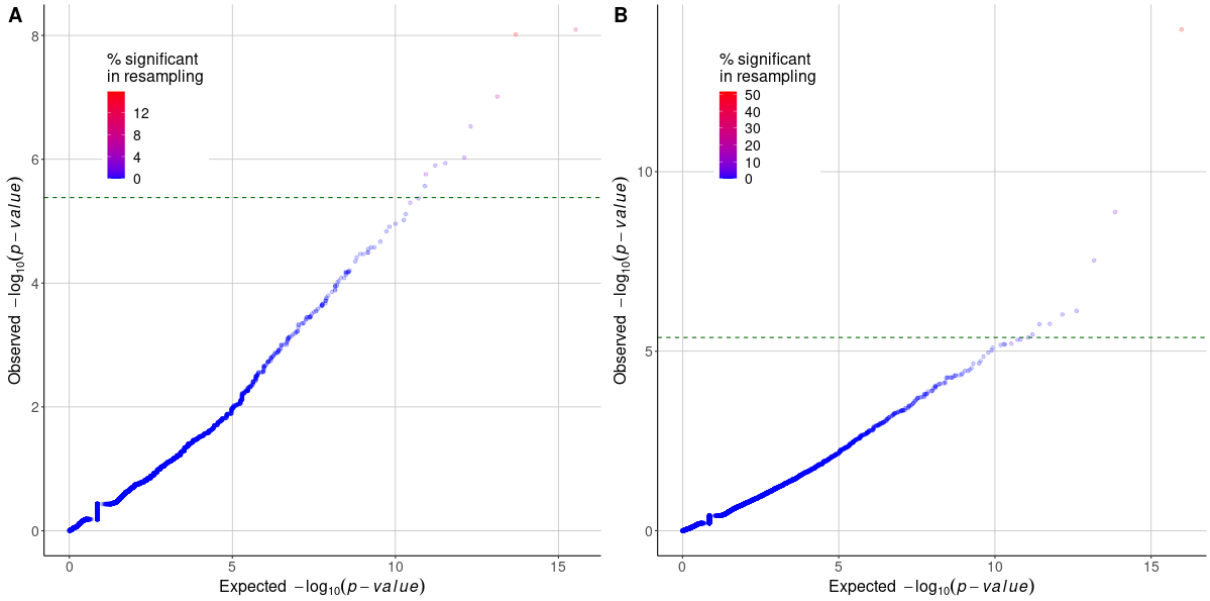



*Figure 16: COBT quantile-quantile plots in non-Finnish Europeans' genes from the 1000 Genomes Project for synonymous variants (A) and missense variants (B)*
*Dots are coloured according to the percentage of resampling in which they were significant; the green dotted line is the Bonferroni threshold*

To further characterize possible hidden factors leading to an enrichment of synonymous and missense variants in specific genes among the 1000 Genomes cohort, we inspected how carrier individuals for such variants distributed in terms of population stratification. Principal Component Analysis (PCA) based on common variants (AF>1%) of the 404 individuals actually reflected the subpopulation origin among non-Finish Europeans, separating the Iberian and the Toscan subpopulations from the Utah residents with Northern and Western European ancestry (CEPH) and the British population (Figure 17). The analysis of the distribution of variants across carriers showed an association with specific European sub-populations for 3 out of nine and 4 out of eight gene hits enriched in synonymous and in missense variants, respectively (Chi-squared test p-value < 0.05; Table 4). The previous results suggest on the one hand, that a fraction of the gene hits, notably those with the highest (yet significant) p-values, may originate from stochastic sampling of the general population; on the other hand, a fraction of gene hits

is associated with specific European sub-populations, suggesting that such subpopulations were unevenly represented in the reference gnomAD Cohort as compared to the 1000 Genomes cohort.

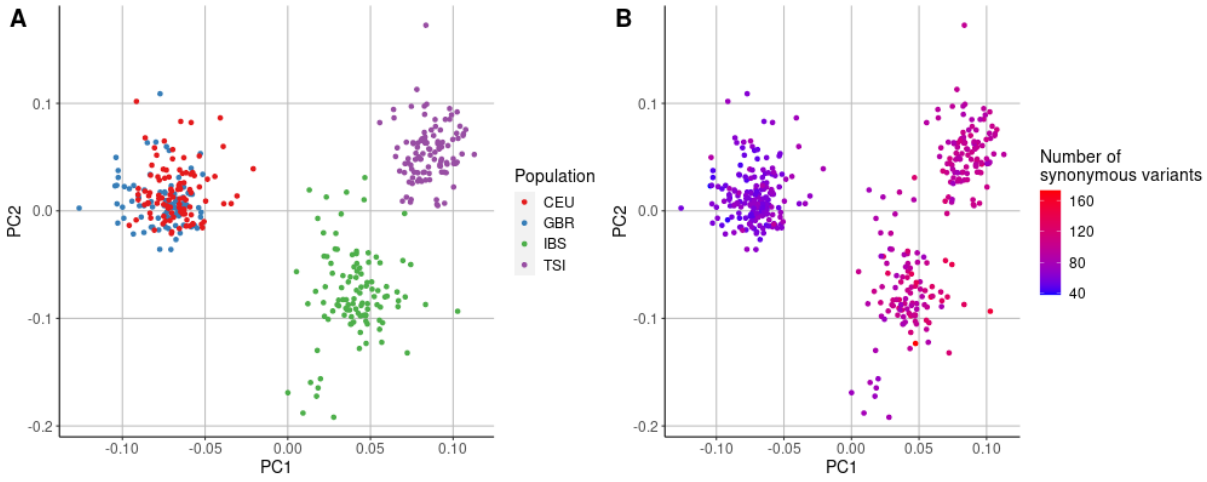



*Figure 17: Principal Component Analysis on common variants from non-Finnish European individuals from the 1000 Genomes Project projected on the first two principal components*
*Coloured by subpopulation (A) and by number of synonymous variants (B); CEU = Utah residents (CEPH) with Northern and Western European ancestry; TSI = Toscani in Italia; GBR = British in England and Scotland; IBS = Iberian populations in Spain*

*Table 4: Summary information of genes significantly enriched in synonymous or missense mutations from the 1000 Genomes Project non-Finnish Europeans individuals*

**Logistic regression testing the association between the mutability of significantly enriched genes and subpopulations: the dash means no subpopulation, when in italic it means significant in the test with CEPH and British grouped, p-value is in parenthesis*

***$\chi^2$-test testing the association between the total number of mutations in significantly enriched genes and subpopulations: the dash means no subpopulation, when in italic it means significant in the test with CEPH and British grouped, p-value is in parenthesis*

*CEU = Utah residents (CEPH) with Northern and Western European ancestry; TSI = Toscani in Italia; GBR = British in England and Scotland; IBS = Iberian populations in Spain*

| Mutation type | Gene symbol | Percentage of time significant in resampling | Number of mutated individuals | Sum of total mutations across individuals | Ratio sum of mutations to number of mutated individuals | P-value in the whole cohort (after Bonferroni correction) | Expected counts (gnomAD number of variants) | Transcript length | Percentage of the transcript covered | Logistic regression on sub-populations* | $\chi^2$-test on sub-populations** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Synonymous** | SPP2 | 8.3 | 12 | 13 | 1.08 | 9.68E-05 | 12 | 860 | 51.40 | - | - |
| | MTUS1 | 0.5 | 13 | 13 | 1 | 1.39E-02 | 30 | 3,731 | 30.69 | - | - |
| | MTRES1 | 1.5 | 6 | 6 | 1 | 1.13E-02 | 6 | 1,434 | 37.10 | - | CEU (p=0.0101) *CEU+GBR (p= 0.0302)* |
| | CHSY1 | 0 | 17 | 17 | 1 | 3.26E-02 | 48 | 4,55 | 44.75 | - | - |
| | USP3 | 2.5 | 9 | 17 | 1.89 | 2.11E-02 | 27 | 2,333 | 52.12 | *TSI (p=0.0164)* | TSI (p= 1.1968E-06) *TSI (9.7536E-05)* |
| **Missense** | PRNP | 15.8 | 8 | 15 | 1.88 | 1.16E-04 | 15 | 2,657 | 25.59 | - | - |
| | TAF15 | 6.8 | 12 | 18 | 1.50 | 1.17E-03 | 35 | 2,191 | 70.33 | - | CEU (p= 0.0278) |
| | GABRG3 | 0.9 | 13 | 14 | 1.08 | 1.52E-02 | 26 | 2,004 | 60.23 | - | - |
| | TNFRSF25 | 1.4 | 13 | 15 | 1.15 | 3.55E-03 | 22 | 1,632 | 55.15 | - | - |
| | TSPAN17 | 1.6 | 9 | 9 | 1 | 2.08E-02 | 17 | 2,361 | 27.83 | - | GBR (p=0.0396) *CEU+GBR (p= 0.0189)* |
| | AAGAB | 0.3 | 13 | 13 | 1 | 4.15E-02 | 33 | 2,841 | 26.65 | - | - |
| | METAP2 | 2.1 | 10 | 12 | 1,2 | 1.15E-02 | 24 | 3,601 | 36.18 | - | *CEU+GBR (p= 0.0087)* |
| | HRC | 51.6 | 15 | 40 | 2,67 | 1.33E-10 | 108 | 2,385 | 78.03 | - | CEU (p= 0.0293) |
| | ADGRF1 | 9.8 | 28 | 29 | 1,04 | 1.60E-05 | 110 | 5,468 | 47.31 | IBS (p= 0.0436) | *CEU+GBR (p=0.0310)* |
| | PACSIN3 | 0.7 | 20 | 20 | 1 | 9.06E-03 | 61 | 1,843 | 69.02 | - | *CEU+GBR (p= 0.0125)* |
| | DNAJB7 | 0.1 | 16 | 16 | 1 | 2.14E-02 | 47 | 2,578 | 35.57 | - | - |
| | LMNTD2 | 4.5 | 20 | 20 | 1 | 3.58E-04 | 58 | 2,084 | 38.96 | - | - |

### 3.2.2 Gene burden analysis of a case-only ciliopathy cohort

We then evaluated the rare variant gene burden in a case-only cohort of 496 patients suffering from diverse ciliopathy types. Clinical evaluation allowed to group patients into 6 categories according to their major characteristic phenotype, *i.e.*: kidney involvement (n=381), neurological defects (n=281), skeletal defects (n=211), eye anomalies (n=182), hepatic defects (n=118) and heart defects (n=61). Yet, a wide phenotypic spectrum could be observed within each group. The sum of patients in each category is higher than the total number of patients because there is some overlap between groups. Ciliary exome-targeted sequencing (hereafter referred as *ciliome sequencing*), was conducted, targeting a total of 1,221 cilia-related genes (*cf.*, 3.1.1). Following ACMG/AMP variant interpretation guidelines [154] clinical geneticist identified a total of 87 different causal genes (*i.e.*, bearing homozygous or compound heterozygous variants classified as pathogenic or likely pathogenic) on $n = 175$ patients (37% of the cohort). The rest of patients $n = 303$ (63% of the cohort) remained genetically unsolved, either because pathogenic or likely pathogenic variants were found in heterozygous state, or because no candidate variants were reported. We thus used COBT on the global cohort of 478 patients to evaluate the per-gene burden of rare variants in order to (i) evaluate its ability to recapitulate know causal genes previously identified in the cohort, (ii) identify novel candidate causal variants that could act as primary or second hits for the genetically unsolved cases.

*Table 5: Summary table for genes significantly enriched in missense variants*
*Patients "involved in the detection" refer to patients mutated in the gene significantly enriched in mutations detected by COBT whereas "total number of patients" refers to any ciliopathy patient in the cohort*

| Mutated gene | Known ciliopathy gene (J.Reiter) | CiliaCarta score | Number of patients mutated | Number of patients with a causal gene identified involved in the detection | Number of patients with the same causal gene as the case-only significant gene involved in the detection | Total number of patients with this identified causal gene in the cohort |
|---|---|---|---|---|---|---|
| *CC2D1A* | No | NA | 14 | 11 | 0 | 0 |
| *DNAJB13* | **established** | 1.26 | 10 | 5 | 0 | 0 |
| *IFT140* | **established** | **4.86** | **43** | **25** | **7** | **11** |
| *TTC30A* | **candidate** | **9.64** | **18** | **8** | **0** | **0** |
| *TUBA1A* | **candidate** | **5.97** | **7** | **5** | **0** | **0** |
| *WDR60* | **established** | **3.61** | **28** | **10** | **3** | **3** |

*Table 6: Summary table for genes significantly enriched in protein altering variants*
*Patients "involved in the detection" refer to patients mutated in the gene significantly enriched in mutations detected by COBT whereas "total number of patients" refers to any ciliopathy patient in the cohort*

| Mutated gene | Known ciliopathy gene (J.Reiter) | CiliaCarta score | Number of patients mutated | Number of patients with a causal gene identified involved in the detection | Number of patients with the same causal gene as the case-only significant gene involved in the detection | Total number of patients with this identified causal gene in the cohort |
|---|---|---|---|---|---|---|
| *CC2D1A* | No | NA | 24 | 11 | 0 | 0 |
| *CCDC40* | **established** | **3.65** | 34 | 18 | 0 | 0 |
| *DNAJB13* | **established** | **1.26** | 10 | 5 | 0 | 0 |
| *EXOC6B* | No | -4.70 | 15 | 6 | 0 | 0 |
| *IFT140* | **established** | **4.86** | **45** | **26** | **8** | **11** |
| *IFT52* | **established** | **4.86** | 14 | 10 | 0 | 0 |
| *KIF3A* | **candidate** | **7.27** | 15 | 8 | 0 | 0 |
| *LTN1* | No | NA | 32 | 20 | 0 | 0 |
| *NADK2* | No | NA | 13 | 10 | 0 | 0 |
| *NPHP4* | **established** | **5.23** | **40** | **24** | **13** | **16** |
| *PDE6H* | No | NA | 8 | 4 | 0 | 0 |
| *RASAL2* | No | NA | 14 | 5 | 0 | 0 |
| *SNX3* | No | NA | 7 | 4 | 0 | 0 |
| *TUBA1A* | **candidate** | **5.97** | 7 | 5 | 0 | 0 |
| *WDR60* | **established** | **3.61** | **28** | **10** | **3** | **3** |

By applying COBT on protein-altering variants (*cf.*, 3.1.3), 15 genes were identified as presenting a higher burden than expected by chance (Bonferroni corrected p-value < 0.05; Table 6). Those genes collectively targeted 194 patients, including 82 genetically resolved cases (Table 9). Among them, 21 individuals (25%) presented a causal gene previously identified through clinical assessment that was actually identical to the one identified by the COBT approach in the affected individuals. Such reidentification ratio is significantly higher that what would be obtained by chance randomly selecting the same number n=15 of cilia-related hits among the ciliome panel (Monte Carlo resampling p-value<0.003). We then evaluated the n=112 unsolved patients presenting variants in the 15 genes identified through COBT, all of which were present in heterozygous state. Of them, 14 presented variants reported as pathogenic or likely pathogenic in ClinVar[36] or HGMD[148], while 49 had variants predicted as potentially pathogenic (CADD score higher than 15; Table 9). Further manual inspection by clinical experts considered that, among such 63 patients, 7 presented phenotypic manifestations compatible with the genes highlighted by COBT, and which could thus constitute candidate primary

hits, yet in heterozygous state. As expected from a decrease in statistical power, the previous figures were partly reproduced when independently considering the 6 major clinical categories described above, as well as when considering only missense variants as the qualifying set (Table 7). On the contrary, no novel hits were identified in such subsets. The previous results showed the capacity of the COBT test to recapitulate previously identified causal genes at statistically significant levels and to identify novel candidate pathogenic variants compatible with the phenotypes under investigation in unsolved cases.

*Table 7: Summary table for genes significantly enriched in missense variants in each disease category*
*Patients "involved in the detection" refer to patients mutated in the gene significantly enriched in mutations detected by COBT whereas "total number of patients" refers to any ciliopathy patient in the cohort*

| Mutated gene | Known ciliopathy gene (J.Reiter) | CiliaCarta score | Number of patients mutated | **Number of patients with a causal gene identified *involved in the detection*** | **Number of patients with the same causal gene as the case-only significant gene *involved in the detection*** | **Total number of patients with this *identified* causal gene *in the cohort*** | Disease category |
|---|---|---|---|---|---|---|---|
| *IFT140* | **established** | **4.86** | **32** | **19** | **6** | **11** | **Kidney involvement** |
| *TUBA1A* | **candidate** | **5.97** | 7 | 5 | 0 | 0 | Kidney involvement |
| *DNAJB13* | established | 1.26 | 6 | 3 | 0 | 0 | Neurological defects |
| *TTC30A* | **candidate** | **9.64** | **14** | **6** | **0** | **0** | **Neurological defects** |
| *TUBA1A* | **candidate** | **5.97** | **5** | **3** | **0** | **0** | **Neurological defects** |
| *DNAJB13* | established | 1.26 | 4 | 4 | 0 | 0 | Skeletal defects |
| *IFT140* | **established** | **4.86** | **24** | **14** | **6** | **11** | **Skeletal defects** |
| *DNAJB13* | established | 1.26 | 4 | 3 | 0 | 0 | Heart defects |

*Table 8: Summary table of variants from patients involved in missense variants enrichment by selecting their most damaging variants (p-values related to the Monte-Carlo simulations)*

| | Number of patients involved in the case-only genes detected | Number of patients with an identified causal gene | Number of patients with a ClinVar/HGMD damaging variant in the causal gene | | | | | Number of patients without ClinVar/HGMD damaging variant but a CADD score in the causal gene | | | | | Number of patients without a causal gene identified | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Number of patients with causal variants and case-only enriched variants in the same gene | | Number of patients with a variant probably more damaging (higher CADD score) in the causal gene than the one in the case-only hit | | Number of patients with a variant probably more damaging (higher CADD score) in the case-only hit than the one in the causal gene | Number of patients with causal variants and case-only enriched variants in the same gene | | Number of patients with a variant probably more damaging (higher CADD score) in the causal gene than the one in the case-only hit | | Number of patients with a variant probably more damaging (higher CADD score) in the case-only hit than the one in the causal gene | Number of patients with a ClinVar / HGMD damaging variant | Number of patients with a CADD score ≥ 15 in case-only variant |
| | | | Total | With the same variant | Total | Number of patients with a CADD score ≥ 15 in case-only variant | Total | Total | With the same variant | Total | Number of patients with a CADD score ≥ 15 in case-only variant | Total | | |
| **Counts** | 120 | 45 | 5 | 5 | 15 | 10 | 4 | 3 | 2 | 17 | 6 | 1 | 3 | 34 |
| | | | | | | | 24 | | | 21 | | | 75 | |
| **P-value** | 0.478 | 0.568 | **0.017** | **0.007** | 0.800 | 0.871 | 0.448 | **0.053** | **0.062** | 0.103 | 0.784 | 0.991 | 0.791 | 0.647 |
| | | | 0.552 | | | | 0.448 | 0.309 | | | | 0.990 | 0.509 | |

*Table 9: Summary table of variants from patients involved in protein altering variants enrichment by selecting their most damaging variants (p-values related to the Monte-Carlo simulations)*

| | Number of patients involved in the case-only genes detected | Number of patients with an identified causal gene | Number of patients with a ClinVar/HGMD damaging variant in the causal gene | | | | | | Number of patients without ClinVar/HGMD damaging variant but a CADD score in the causal gene | | | | | | Number of patients without a causal gene identified | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Number of patients with causal variants and case-only enriched variants in the same gene | | Number of patients with a variant probably more damaging (higher CADD score) in the case-only hit than the one in the causal gene | | Number of patients with a variant probably more damaging (higher CADD score) in the causal gene than the one in the case-only hit | | Number of patients with causal variants and case-only enriched variants in the same gene | | Number of patients with a variant probably more damaging (higher CADD score) in the causal gene than the one in the case-only hit | | Number of patients with a variant probably more damaging (higher CADD score) in the case-only hit than the one in the causal gene | | Number of patients with a ClinVar / HGMD damaging variant | Number of patients with a CADD score ≥ 15 in case-only variant |
| | | | Total | With the same variant | Total | Number of patients with a CADD score ≥ 15 in case-only variant | Total | Number of patients with a CADD score ≥ 15 in case-only variant | Total | With the same variant | Total | Number of patients with a CADD score ≥ 15 in case-only variant | Total | Number of patients with a CADD score ≥ 15 in case-only variant | | |
| **Counts** | 194 | 82 | 11 | 10 | 26 | 24 | 10 | 10 | 10 | 8 | 19 | 10 | 6 | 6 | 14 | 49 |
| | | | | | 47 | | | | 35 | | | | | | 112 | |
| **P-value** | 0.893 | 0.860 | **0.008** | **0.005** | 0.973 | 0.911 | 0.395 | 0.395 | **0.003** | **0.004** | 0.711 | 0.967 | 0.988 | 0.984 | 0.345 | 0.960 |
| | | | 0.669 | | | | | | 0.555 | | | | | | 0.953 | |

*Table 10: Summary table for KEGG pathways significantly enriched in missense variants*

**Number of genes studied in the dataset that are in the pathway*

***Number of genes with a CiliaCarta score ≥ 1 and/or established/candidate gene from J. Reiter's review*

****Number of genes with a CiliaCarta score ≥ 1 and/or established/candidate gene from J. Reiter's review + significant in the case-only burden gene test*

***** Number of genes studied in the dataset that are in the pathway + significant in the case-only burden gene test*

| KEGG Pathway | Number of genes | Number of ciliome genes* | Number of mutated ciliome genes | Number of ciliopathy genes** | Number of ciliopathy genes enriched in the gene test*** | Number of ciliome enriched genes in the gene test**** | Number of different variants | Number of mutated patients | Sum across patients |
|---|---|---|---|---|---|---|---|---|---|
| AMINOACYL TRNA BIOSYNTHESIS | 41 | 1 | 1 | 1 | 0 | 0 | 18 | 24 | 29 |
| AXON GUIDANCE | 129 | 6 | 4 | 3 | 0 | 0 | 20 | 30 | 30 |
| BASE EXCISION REPAIR | 35 | 2 | 2 | 0 | 0 | 0 | 16 | 19 | 20 |
| CELL CYCLE | 125 | 15 | 15 | 5 | 0 | 0 | 112 | 124 | 150 |
| EPITHELIAL CELL SIGNALING IN HELICOBACTER PYLORI | 68 | 2 | 1 | 1 | 0 | 0 | 8 | 13 | 14 |
| GAP JUNCTION | 90 | 5 | 5 | 18 | 1 | 1 | 13 | 22 | 24 |
| HOMOLOGOUS RECOMBINATION | 28 | 1 | 1 | 0 | 0 | 0 | 12 | 14 | 14 |
| HUNTINGTONS DISEASE | 182 | 14 | 14 | 14 | 0 | 0 | 228 | 191 | 283 |
| JAK STAT SIGNALING PATHWAY | 155 | 2 | 2 | 0 | 0 | 0 | 6 | 6 | 6 |
| LYSOSOME | 121 | 5 | 5 | 2 | 0 | 0 | 22 | 30 | 30 |
| PORPHYRIN AND CHLOROPHYLL | 41 | 1 | 1 | 0 | 0 | 0 | 18 | 24 | 29 |
| RNA DEGRADATION | 59 | 1 | 1 | 1 | 0 | 0 | 15 | 19 | 19 |
| TGF BETA SIGNALING PATHWAY | 86 | 1 | 1 | 4 | 0 | 0 | 1 | 1 | 1 |
| VASOPRESSIN REGULATED WATER | 44 | 16 | 14 | 11 | 0 | 0 | 103 | 107 | 132 |

*Table 11: Summary table for KEGG pathways significantly enriched in protein altering variants*
**Number of genes studied in the dataset that are in the pathway*
***Number of genes with a CiliaCarta score ≥ 1 and/or established/candidate gene from J. Reiter's review*
****Number of genes with a CiliaCarta score ≥ 1 and/or established/candidate gene from J. Reiter's review + significant in the case-only burden gene test*
***** Number of genes studied in the dataset that are in the pathway + significant in the case-only burden gene test*

| KEGG Pathway | Number of genes | Number of ciliome genes* | Number of mutated ciliome genes | Number of ciliopathy genes** | Number of ciliopathy genes enriched in the gene test*** | Number of ciliome enriched genes in the gene test**** | Number of different variants | Number of mutated patients | Sum across patients |
|---|---|---|---|---|---|---|---|---|---|
| AMINOACYL TRNA BIOSYNTHESIS | 41 | 1 | 1 | 1 | 0 | 0 | 24 | 25 | 39 |
| AXON GUIDANCE | 129 | 6 | 4 | 3 | 0 | 0 | 20 | 30 | 30 |
| EPITHELIAL CELL SIGNALING IN HELICOBACTER PYLORI INFECTION | 68 | 2 | 1 | 1 | 0 | 0 | 8 | 13 | 14 |
| GAP JUNCTION | 90 | 5 | 5 | 18 | 1 | 1 | 13 | 22 | 24 |
| HEDGEHOG SIGNALING PATHWAY | 56 | 4 | 4 | 13 | 0 | 0 | 51 | 69 | 78 |
| HOMOLOGOUS RECOMBINATION | 28 | 1 | 1 | 0 | 0 | 0 | 15 | 17 | 19 |
| HUNTINGTONS DISEASE | 182 | 14 | 14 | 14 | 0 | 0 | 239 | 196 | 302 |
| JAK STAT SIGNALING PATHWAY | 155 | 2 | 2 | 0 | 0 | 0 | 10 | 8 | 12 |
| PORPHYRIN AND CHLOROPHYLL METABOLISM | 41 | 1 | 1 | 0 | 0 | 0 | 24 | 25 | 39 |
| PROTEASOME | 46 | 17 | 12 | 2 | 0 | 0 | 31 | 29 | 34 |
| PROTEIN EXPORT | 24 | 6 | 5 | 2 | 0 | 0 | 15 | 20 | 30 |
| VASOPRESSIN REGULATED WATER REABSORPTION | 44 | 16 | 16 | 11 | 0 | 0 | 127 | 125 | 182 |
| VIBRIO CHOLERAE INFECTION | 54 | 5 | 4 | 6 | 0 | 0 | 9 | 17 | 25 |

## 3.3 Discussion

In this work we implemented the COBT method, a statistical test for the identification of genes carrying an excess of rare variant burden in rare disease cohorts upon case-only experimental designs, *i.e.*, when genomic controls are not available. To that aim, we took advantage of publicly available large exome and genome sequencing projects, allowing to obtain estimates of expected gene variation rates on the general population. Comprehensive goodness-of-fit tests validated the Poisson-based statistical assumptions of the approach in terms of genetic counts dispersion, homogeneity and frequency of zeros. Furthermore, stringent filtering of genomic regions and qualifying variants together with genomic control of p-values showed COBT's ability to keep both inflation and false positives at low rates. Yet, COBT approach appeared as sensitive to stochastic sampling of the general population and to an uneven representation of specific subpopulations between the reference and the target cohorts, as illustrated through the analysis of a sample of non-finish Europeans from the general population. Finally, the application of COBT to the re-analysis of an inhouse cohort of ciliopathy patients genetically profiled through ciliome-sequencing showed the capacity of the COBT test to recapitulate previously reported causal genes in a statistically significant manner. In addition, clinical expert assessment of the gene hits identified on genetically unsolved patients allowed to propose novel candidate pathogenic variants at heterozygous state for 7 cases, for which clinical symptoms were found to be compatible with the gene-phenotype associations previously described in the literature.

Our work is in line with recent statistical genetics tests showing the possibility of identifying gene-phenotype associations on case-only design studies by making use of aggregated population frequencies from the general population[42,109,155]. Notwithstanding, as comprehensively reviewed in Povysil et al. (2019)[74] and Lee et al. (2014)[73], stringent filters accounting for sequencing factors (such as depth and coverage), hidden population structure, mappability of genomic regions and the definition of qualifying variants were needed to minimize spurious signals in case-only experimental designs. Our approach represents however a conceptual novelty in the field. Thus, current state-of-the-art case-only tests such as TRAPD[81] and CoCoRV[82] are based on the estimation of an excess of mutated individuals for a given gene. This is in contrast with COBT as well as with classical gene-based burden tests for case-control designs[73], where the co-occurrence of multiple variants targeting the same gene within an individual positively contributes to the excess of actual variant burden being evaluated. Moreover, from a practical perspective, when applied to a putatively healthy cohort randomly sampled from the general population and in contrast to COBT, both TRAPD and CoCoRV presented significant inflation rates in their p-value distributions and ultimately led to an unexpected excess of significant gene hits, which may compromise their utility for the analysis of patient cohorts.

Case-only burden tests were explored in this work using genes as the aggregation unit and based on their protein-coding sequence. Natural extensions of the COBT approach could be proposed for the analysis of gene non-coding regions (*e.g.*, introns, untranslated regions and promoters) either jointly or separately considered from the coding ones, as reviewed in Bocher & Génin (2020)[156]. To that aim, the rapid increase in the availability of whole genome-sequencing data from the general population may soon provide frequency estimates close to the ones previously generated on the basis of WES data, as it is now the case in the gnomAD database[44]. Similarly, the COBT approach could be readily extended to the analysis of gene-sets defined on the basis of a common molecular function or biological pathway. Thus, mutation frequencies may be aggregated to generate estimates about the number of variants that may be expected from random sampling on the general population. Gene-set based case-only burden tests could thus contribute to identify novel gene-phenotype associations that might be so far undetected due to low statistical power in small, rare disease cohorts or on diseases with a large genetic heterogeneity.

COBT is a burden test method that allows to identify gene-disease associations that is particularly useful when covariate analysis cannot be afforded because of technical limitations. COBT is thus an alternative to classical regression-based approaches that are more precise with covariates. However, COBT results must be interpreted with caution as statistical power might be an issue, especially for small heterogeneous cohorts. One of the main interests of the approach is that it may reveal gene-disease association for modifier variants or secondary hits that would not have been detected with a classical method. Nonetheless, expert knowledge is necessary to examine the functional relevance of such variants.

# CHAPTER 4. CLINICAL SIMILARITY FOR RARE VARIANTS PRIORITISATION

The aim of this part of my thesis is to develop a method to identify candidate variants in heterogeneous developmental disorders by making use of both genotypes and phenotypes. To achieve this, I proposed to systematically construct groups of similar patients using the Human Phenotype Ontology to prioritise exome variants. Such a strategy assumes that phenotypically similar patients assessed through HPO are caused by shared genetic features. To test this hypothesis, I first computed the semantic similarities between patients within a reference cohort (DDD) based on their HPO terms as a proxy of their phenotypic similarity. Then, I evaluated whether this proximity shared genotypic characteristics identified on the patients' exomes. As a positive control for this evaluation, I used a list of DNMs that were considered by the DDD consortium as likely pathogenic.

As HPO database is a directed *acyclic* graph, numerous similarity computation approaches have already been developed. However, most of such approaches have been applied to the assessment of the similarity of the HPO profiles associated to a given patient and those associated to disease genes (HPO-gene-disease links are provided and regularly updated in the database). There are numerous ways to compute a score of similarity between 2 individual HPO terms. In addition, different ways exist to aggregate such pairwise similarities between individual terms in order to assess the similarity between 2 lists of HPO terms, which seems to be a highly discriminating factor. Therefore, I have tried to answer two questions:

- Are there differences between methods of semantic similarity and/or a better method of aggregating scores of semantic similarity?
- Can semantic similarity be used to identify shared causal genetic mechanisms between phenotypically similar patients?

## 4.1 MATERIAL AND METHODS

### 4.1.1 Deciphering Developmental Disorders cohort

The DDD consortium provided access to the cohort from the 2017 DDD study consisting of the exomes of 4,290 samples from patients together with from their parents ($n = 4287$ trios, $n = 3$ duos), as well as the lists of HPO terms describing their clinical symptoms[48]. The VCF files of the 4,290 samples were downloaded through European Genome-phenome Archive (EGA) as well as their associated HPO annotations. In the following analysis, the dataset was evaluated considering four different subsets:

- The whole dataset (4,286 patients)
- The fraction of patients from the whole dataset with at least 2 HPO terms (3,955 patients)
- Patients with DNMs (2,620 patients)

- The fraction of patients with DNMs and at least 2 HPO terms (2,426 patients)

### 4.1.2 Bioinformatic pipeline for *de novo* variants detection and filtering

Exome data were recorded by trio through multiple VCF files. Out of the 4,290 probands, 4,287 had a complete trio and 4,286 of them had HPO annotations. The VCFs were corrupted: VEP annotations were removed, and the files were corrected with GATK's FixVcfHeader tool: 4,282 VCF were recovered and reannotated. The calling of *de novo* variants was carried out with *hail*'s *de novo* caller. A total of 127,271 *de novo* variants were found among 3,789 samples. McRae and colleagues' study guidelines were followed to filter the transcripts with some minor modifications:

1. The transcript with the most severe consequence was kept following these VEP categories:
    1. PTV: splice donor, splice acceptor, stop gained, frameshift, initiator codon and conserved exon terminus variant
    2. Missense: missense, stop lost, inframe deletion, inframe insertion, coding sequence and protein altering variant
    3. Silent variant: synonymous
    4. Other (not retained afterwards)
2. If there were several transcripts with the same level of consequence, the APPRIS principal transcript was kept
3. If there were no APPRIS principal transcript, the canonical transcript was kept
4. If there were none or several, the longest transcript was kept

If none of these conditions could be assessed (*e.g.*, transcript ID not found in bioMart Ensembl release 108, GRCh37.p13), the variant was removed. Including variants categorized as "other", 126,847 *de novo* variants remained across 3,779 samples.

For variant filtering, the coverage study guidelines could not be directly followed as GVCFs were not available. Instead of filtering variants on the Minor Allele Frequency (MAF), I used a proxy: variants with an allele count AC < 76 were kept ( $\frac{3789\times2}{100} = 75.78$ ; *i.e.*, 3,798 corresponding to the number of samples in which *de novo* variants were initially found). Then, only PTVs, missense or silent variants were kept and different filters depending on the confidence of variants detection (*i.e.*, computed with phred-scaled genotype likelihoods in the VCF) were applied:

- For high quality SNVs:
    - $P_{de\ novo} > 0.99$ and $AB > 0.3$ and $AC = 1$
    - $P_{de\ novo} > 0.99$ and $AB > 0.3$ and $DR > 0.2$
    - $P_{de\ novo} > 0.5$ and $AB > 0.3$ and $AC < 10$ and $DP > 10$

- For high quality indels: $P_{de\ novo} > 0.99$ and $AB > 0.3$ and $AC = 1$
- For medium quality SNVs:
    - $P_{de\ novo} > 0.5$ and $AB > 0.3$
    - $AC = 1$
- For medium quality indels: $P_{de\ novo} > 0.5$ and $AC < 10$ and $AB > 0.3$
- For low quality SNVs and indels: $AB > 0.2$

where:

- AB = read allele balance of the proband (number of alternate reads divided by total reads)
- DP = read depth (DP field) of the proband
- AC = count of alternate alleles across all individuals in the dataset at the site
- DR = ratio of the read depth in the proband to the combined read depth in the parents
- $P_{de\ novo}$ is the posterior probability of a DNM; $P_{de\ novo} = \frac{P(d|x)}{P(d|x)+P(m|x)}$ with $x$ being the event of existence of a DNM, $d$ the event of the accurate occurrence of a DNM in the proband and $m$ the event that at least one parental allele is heterozygous with the proband call being accurate (more details can be found in the hail *de novo* caller function: https://hail.is/docs/0.2/methods/genetics.html#hail.methods.de_novo)

Finally, only 6,438 variants collectively involving 2,620 samples remained in the dataset.

### 4.1.3 Information Content computation

HPO data was processed with the ontologySimilarity and ontologyIndex packages that belong to the ontologyX R suite[157]. Information Content was computed as the negative logarithm of the frequency of each term in three different corpora, leading to three alternative definitions:

- **IC_f** = IC based on the frequency of terms as ancestors within the whole HPO corpus, $IC(x) = -\log_{10}\left(\frac{number\ of\ child\ terms\ for\ x}{total\ number\ of\ unique\ terms\ \ in\ HPO}\right)$ (*i.e.*, frequency with which an HPO term is an ancestor of other terms: this comprises 16,801 unique terms) computed with the `descendants_IC()` function from ontologyIndex
- **IC_n** = IC based on the frequency of the terms within the cohort, $IC(x) = -\log_{10}\left(\frac{number\ of\ patients\ in\ the\ cohort\ annotated\ with\ x}{total\ number\ of\ terms\ in\ the\ cohort}\right)$ (note that all HPO terms are "artificially extended" to include their ancestors even if they are not used in the cohort so that an IC is computed for the most common ancestor and thus, the more precise an HPO term, the rarer it is as compared to its ancestors and the more weight it has) resulting in 3,668 unique terms with an IC computed

- **HPOA_IC** = IC based on the frequency of the terms within OMIM thanks to the HPO Annotations[158] file (http://purl.obolibrary.org/obo/hp/hpoa/phenotype.hpoa), containing 8,389 unique OMIM diseases (HPO terms are also "artificially extended" to have the common ancestors of all terms), $IC(x) = -\log_{10}\left(\frac{number\ of\ diseases\ in\ OMIM\ annotated\ with\ x}{total\ number\ of\ terms\ in\ OMIM}\right)$ resulting in 9,804 unique terms with an IC computed

The IC_f is the most unbiased option. The rationale behind both IC_n and HPOA_IC is to give more weight to specific terms that are more frequently used in the dataset under investigation, or across OMIM disease genes, respectively.

### 4.1.4 HPO redundancy filtering

Raw HPO descriptions of patients made by physicians may contain some redundant terms (*i.e.*, HPO terms that are "relatives" in the ontology). Unless the relative information between terms is collapsed, the same information would be considered twice. The most informative term should be the only one kept for each clinical symptom description. Thus, for each patient, every HPO term that was an ancestor of another HPO term was removed from the list (Figure 18).

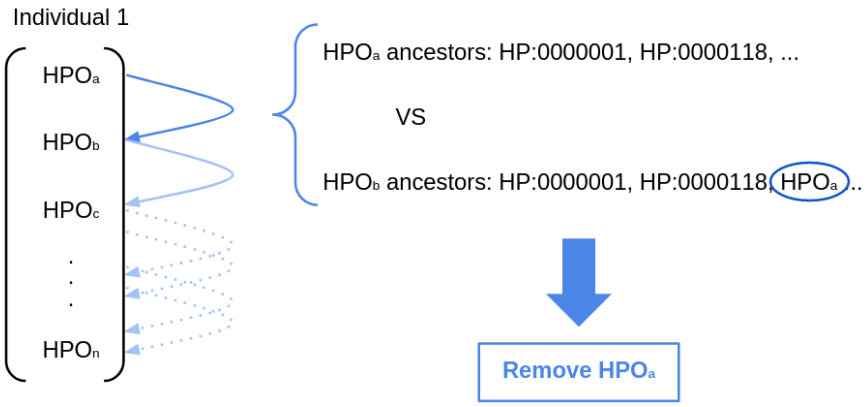



*Figure 18: Conceptual diagram of the HPO redundancy filtering*

### 4.1.5 Semantic similarity computation

Five semantic similarity scores were used to assess phenotypic similarity:

- Resnik (equation [15]):

$$sim_{Resnik}(HPO_1, HPO_2) = IC\left(MICA_{HPO_1,HPO_2}\right);\ sim_{Resnik} \in [0, +\infty[$$

- Lin (equation [16]):

$$sim_{Lin}(HPO_1, HPO_2) = \frac{2 \times IC(HPO_{MICA})}{IC(HPO_1) + IC(HPO_2)};\ sim_{Lin} \in [0,1]$$

- Jiang-Conrath (equation [17]):

$$sim_{JC}(HPO_1, HPO_2) = \frac{1}{1 + IC(HPO_1) + IC(HPO_2) - 2 \times IC(HPO_{MICA})};\ sim_{JC} \in [0,1]$$

- Information Coefficient [18]):

$$sim_{IC}(HPO_1, HPO_2) = sim_{Lin}(HPO_1, HPO_2) \times (1 - \frac{1}{1 + IC(HPO_{MICA})}) ; sim_{IC} \in [0,1]$$

- ERIC (equation [19]):

$$sim_{ERIC}(HPO_1, HPO_2) = \max [0; 2 \times IC(HPO_{MICA} - \min(IC(HPO_1), IC(HPO_2)] ;$$

$$sim_{ERIC} \in [0, +\infty[$$

Score aggregation was assessed with Best Match Average, Best Match Sum and maximum score methods. However, only BMA scores will be shown in the following analyses as the percentages of phenotypically close patients sharing DNMs were systematically higher as compared to BMS and maximum semantic similarities (*cf.*, Supplementary figure 6, Supplementary figure 7 and Supplementary figure 8). The semantic similarity between all couples of patients from the DDD cohort were stored in 5 symmetric matrices. The similarity score $C_{12}$ between sample 1 and 2 will be the same as $C_{21}$ the similarity between samples 2 and 1 (because of the BMA method which averages the best matches in both directions).

### 4.1.6 Phenotypic-genomic proximity evaluation: Monte Carlo simulations

A statistical measure was needed to define couples of samples significantly closer than expected by chance in terms of the semantic similarities described above. Hence, I used the p-values computation method based on Monte-Carlo simulations developed by Köhler *et al.* for the Phenomizer software. To that aim, 10,000 random lists of HPO terms of size $n$, for each n ranging from 1 to 30 HPO terms (minimum to maximum number of terms per DDD patient) were generated, resulting in 30 vectors of 10,000 random HPO lists (*i.e.*, $30 \times 10{,}000$ HPO terms lists). Lists were curated to remove redundancy (*cf.*, 4.1.4). For each of the 5 semantic similarity computation methods and each DDD patient ($n = 4{,}290$), clinical similarity scores of the samples against the $30 \times 10{,}000$ random HPO lists were computed. The p-values for each pair of samples A and B, with HPO lists $HPO_A$ and $HPO_B$, of size $n_A$ and $n_B$, respectively, were computed by assessing the ranking of its similarity score among the 10,000 similarity scores obtained between the $HPO_A$ list from patient A and the 10,000 random lists of size $n_B$ previously generated (*cf.*, Figure 19). Out of the 10,000 scores, the number of scores that are higher than the evaluated one are tallied, and the p-value is computed as follows:

$$p - value_{Monte-Carlo} = \frac{(n_{similarity\ scores\ higher\ than\ the\ couple's\ score} + 1)}{10{,}001} \quad [26]$$

P-values are stored into 5 square matrices for each semantic similarity computation method. They are no longer symmetric because p-value for $C_{12}$ will not necessarily be the same as p-value for $C_{21}$ as p-value computation depends on the number of HPO terms per patient which will not be the same for the 2 samples of the couple. P-values corresponds to the probability of having 2 patients closer than

expected by chance from the HPO database. The p-values for each couple of samples are corrected for multiple testing by Benjamini-Hochberg correction[124]. Each couple of samples will get 2 different p-values. Therefore, the couple of samples was considered significantly phenotypically close only if both p-values for $C_{12}$ and $C_{21}$ were significant after FDR (*cf.*, Figure 19).

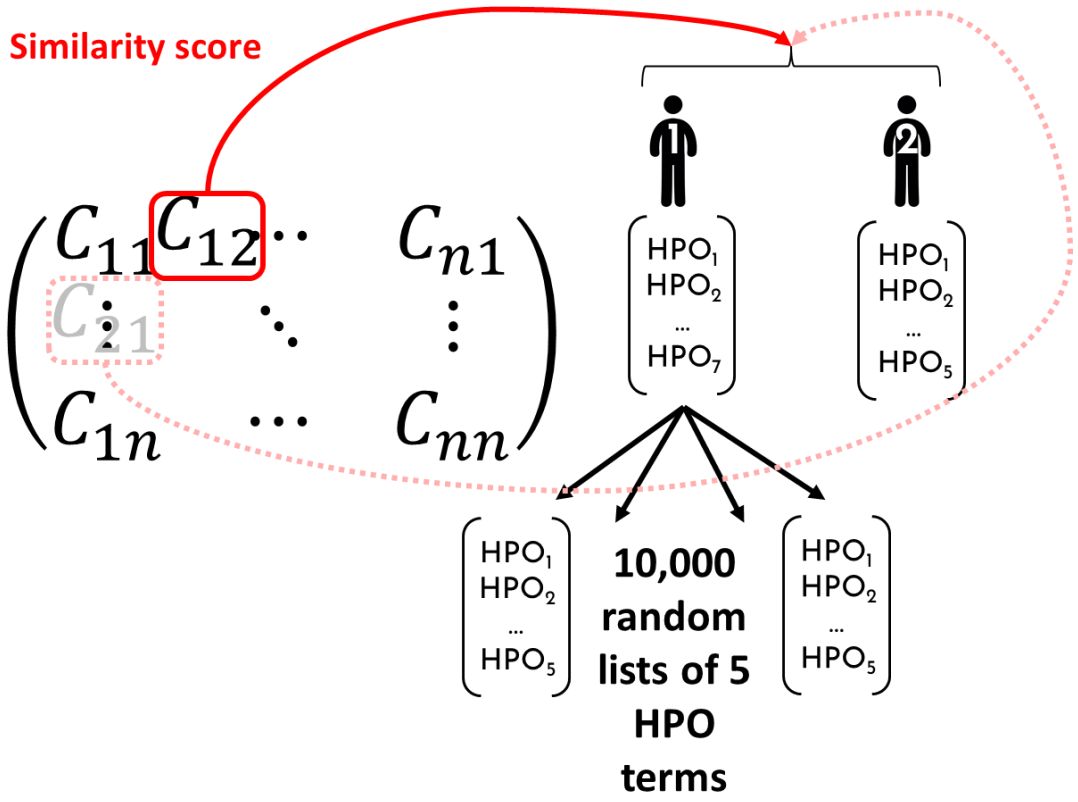



*Figure 19: Schematic representation of the p-value computation with Monte Carlo simulations*
*The semantic similarity score of couple of patients 1&2, $C_{12}$, is compared to the score of each of the 10,000 couples formed by sample 1 with a random list of 5 HPO terms, the same length of HPO terms as sample 2; the same is done in reverse for $C_{21}$*

### 4.1.7 Biological pathways

Three different sets extracted from the Broad Institute's Molecular Signatures Database (MSigDB, https://www.gsea-msigdb.org) used in the Gene Set Enrichment Analysis (GSEA) software were selected for variant collapsing in biological pathways:

- Hallmark, which summarizes and "represent specific well-defined biological states or processes and display coherent expression" and "were generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections", however it only comprises 50 gene sets that are very large)
- Reactome[159], which is a "manually curated and peer-reviewed pathway database" and comprises 1604 gene sets

- KEGG[160], which is also a manually curated pathway database and comprises 196 gene sets

## 4.2 RESULTS

### 4.2.1 DDD cohort studied through Cohort Analyzer

We first evaluated the overall quality of the clinical information reported in the DDD dataset on the basis of HPO terms per patient. A recent study developed *Cohort Analyzer*, a suite of descriptive methods to evaluate the quality of the phenotypic descriptions of clinical cohorts based on phenotypic ontologies[161]. Therein, authors highlighted the necessity of high-quality phenotyping for personalized medicine applications. In this next section, I applied the *Cohort Analyzer* method on the DDD cohort by using the ontologyX R suite.

#### *4.2.1.1 Main summary statistics*

In downstream analyses I evaluated both the whole DDD dataset and a subset of it with patients annotated with more than 2 HPO terms. In addition, such sets were inspected considering all patients, or only those with *de novo* variants. Main summary statistics were overall similar across the different settings (Table 12).

*Table 12 : Main summary statistics of DDD HPO terms*

| Dataset | Unique HPO terms Cohort | Cohort size | HPO terms per patient (average) | Number of HPO terms per patient at percentile 90 of the HPO number distribution | Percentage of HPO terms with more specific child terms |
|---|---|---|---|---|---|
| **Whole dataset** | 2,673 | 4,286 | 6.85 | 3 | 48.90% |
| **Whole dataset with more than 2 HPO terms** | 2,654 | 3,955 | 7.27 | 3 | 49.02% |
| **With de novo variants** | 2,202 | 2,620 | 6.90 | 3 | 49.14% |
| **With de novo variants & more than 2 HPO terms** | 2,185 | 2,426 | 7.31 | 4 | 49.20% |

The average number of HPO terms is close to 7, whether looking at samples having or not *de novo* variants and whether we keep samples with less than 2 HPO terms or not. In the 90$^{th}$ percentile of the distribution of number of HPO terms per patient, samples have more than 2 HPO terms in every considered dataset, indicating good data quality. However, across all the considered subsets**,** half of

the patients presented HPO terms with additional child terms, suggesting that their phenotypic characterization could have been done in a more precise manner. In these regards, the HPO terms most frequently found across the DDD dataset are rather general, such as “global developmental delay” or “specific learning disability” (Table 13). When comparing the number of terms with more specific child terms against the number of HPO terms annotated per sample, no strong correlation was found (Figure 20, Wilcoxon rank sum test $p - value \approx 1$).

*Table 13: Most frequent HPO terms in DDD dataset*

| Dataset | Top HPO | Whole dataset | Percentage of the samples | Whole dataset with more than 2 HPO terms | Percentage of the samples |
|---|---|---|---|---|---|
| **Whole dataset** | **1** | Global developmental delay | 53.42% | Global developmental delay | 50.57% |
| | **2** | Delayed speech and language development | 25.18% | Delayed speech and language development | 24.76% |
| | **3** | Microcephaly | 20.65% | Microcephaly | 20.31% |
| | **4** | Seizure | 16.84% | Seizure | 16.09% |
| | **5** | Specific learning disability | 11.37% | Specific learning disability | 11.15% |
| **With *de novo* variants** | **1** | Global developmental delay | 39.37% | Global developmental delay | 37.53% |
| | **2** | Delayed speech and language development | 18.62% | Delayed speech and language development | 18.31% |
| | **3** | Microcephaly | 15.76% | Microcephaly | 15.51% |
| | **4** | Seizure | 12.44% | Seizure | 11.95% |
| | **5** | Moderate global developmental delay | 8.95% | Moderate global developmental delay | 8.70% |

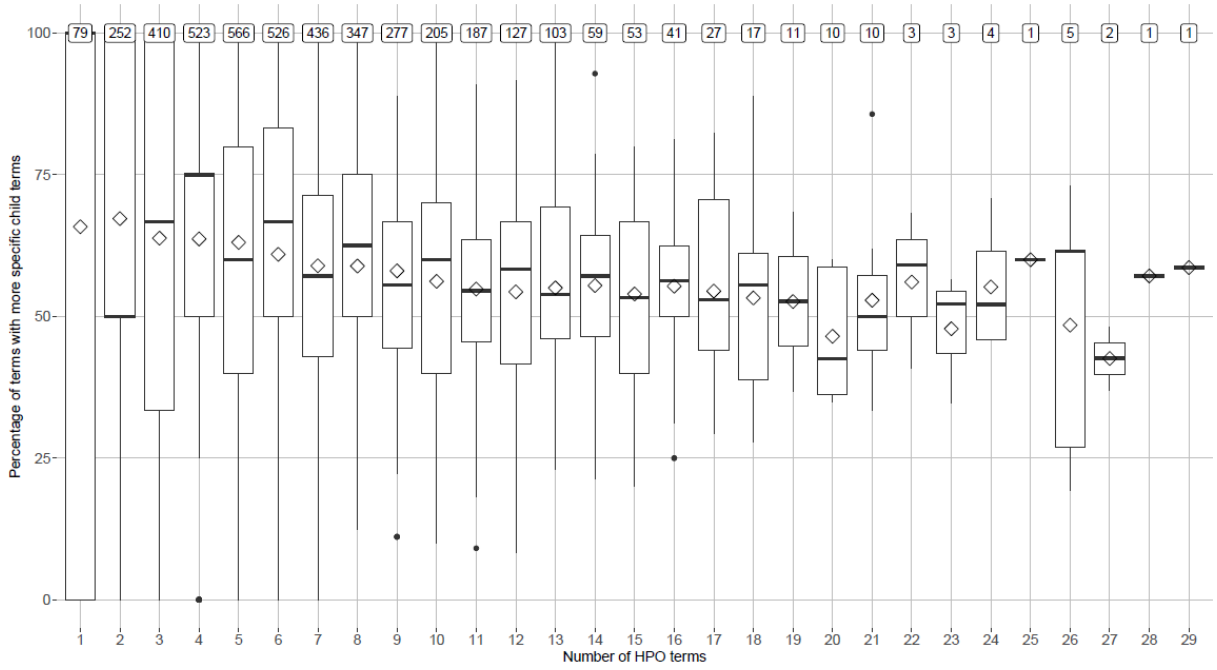



*Figure 20: Percentage HPO terms with more specific child terms per patient as a function of the number of HPO terms for the whole dataset*
*Boxplot displaying the distribution of percentages of HPO terms with more specific child terms reported for each patient as a function of their total number of HPO terms. The whole DDD dataset was considered in the figure. Diamonds represent the mean within each box and the number on top of each boxplot is the number of patients within each box.*

#### *4.2.1.2 Distributions of the HPO term levels*

Another interesting indicator proposed by Rojano and colleagues in *Cohort Analyzer* is to compare the "HPO levels" (*i.e.,* how deep the reported HPO terms are within the ontology) between those used in the cohort and those from the total HPO ontology. This can be interpreted as a proxy of the precision in the annotation process. Again, I analysed the whole DDD dataset as well as a subset considering only samples with DNMs, each with or without samples annotated with 2 or less HPO terms. In addition, I analysed the HPO levels weighting or not by the frequency at which terms occur within the cohort. Weighting allows to control for potential situations where a single highly-phenotyped patient would strongly contribute to the cohort's score. The distribution of HPO terms across HPO levels is highly similar within the full cohort as compared to the subset of patients with *de novo* variants, and this independently of whether patients with less than 2 HPO terms are considered or not (Figure 21).

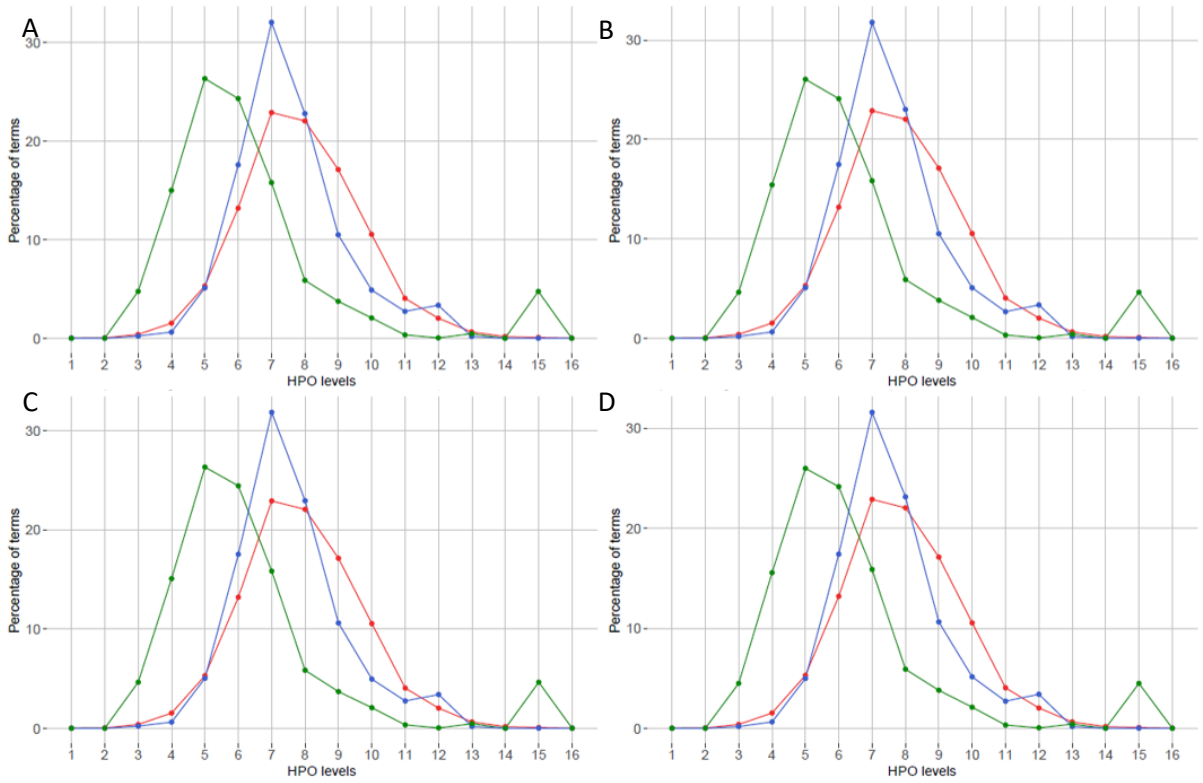



*Figure 21: Percentages of HPO terms in the dataset as a function of the HPO levels*
*Fraction of HPO terms as a function of the level in the HPO hierarchy in the whole cohort (A); the patients with de novo variants (B); the whole cohort patients with more than 2 HPO terms (C); the patients with* de novo *variants with more than 2 HPO terms (D); the red line represents the level in the ontology, the green line represents the unique terms in the cohort and the blue line represents the weighted cohort*

The unique terms curve (green curves in Figure 21) are skewed towards lower levels of HPO terms as compared to the ontology, suggesting that overall, less informative terms than average (and less than what would be expected by chance) were used. However, a small peak at HPO level 15 was observed, suggesting that very precise terms were also used more than what would be expected by chance. When correcting with the frequency in the cohort, the observed distribution got closer to the one representing the total ontology. The observed trends show that the DDD cohort is not annotated with less precision than what could be expected from the global HPO distribution levels. Nonetheless, it is possible that the HPO ontology is not well adapted for developmental disorders, in the sense that HPO terms used to describe them are generally of low level.

Rojano and colleagues also provided an index measuring "the extent to which a cohort has been phenotyped in terms of HPO depth" called DsI (for Dataset specificity Index), applicable "both for unique terms and considering term frequency"[161] (Table 14). To compute DsI, the HPO hierarchy is divided into two sections: a high specificity section and a low specificity section (respectively on the right and on the left of plots in Figure 21). For each level $L$ in the HPO, the difference score $d_L$ measuring the gap between the proportion of terms observed in the cohort and the proportion of terms in the ontology is computed: $d_L = P_{obsL} - P_{ontologyL}$.

A Low section Score ($LsS$) and a High section Score ($HsS$) are computed for the two previously described sections with $L_{max}$ being the level with the highest number of terms in the ontology (9 terms), $L_S$ the number of levels in the section and $L_0$ the highest level of the ontology (16 terms): $LsS = \frac{\sum_{L=1}^{Lmax} d_L \times (L_{max} - L + 1);\, if\ d_L > 0}{L_S}$ and $HsS = \frac{\sum_{L=L_{max}+1}^{L_0} d_L \times (L_0 - L);\, if\ d_L > 0}{L_S}$. DsI is the ratio of the two: $DsI = \frac{HsS}{LsS}$.

*Table 14: Dataset specificity Index for the weighted cohort or unique terms of the DDD dataset*

| Frequency | Whole dataset | With de novo variants | Whole dataset & more than 2 HPO terms | With de novo variants & more than 2 HPO terms |
|---|---|---|---|---|
| **Weighted** | 0.487 | 0.586 | 0.550 | 0.649 |
| **Unique** | 0.023 | 0.022 | 0.022 | 0.021 |

DsI values of 1 indicate that the contribution to the dataset of left and right sections of Figure 21 are identical (*i.e.*, this would imply that the green or blue curve from Figure 21 are symmetric with a peak at 9). A value of 0 means there is a strong imbalance, usually towards lower section which contributes more. On the one hand, when taking unique terms, the score is close to zero, which is not surprising as few terms are seen above level 9. The same message can be conveyed for the four datasets. On the other hand, with weighted HPO terms, the score is closer to 0.5 or even above with *de novo* variants. This suggests that some high Information Content terms might be a bit more represented in the cohort than low-level terms. This score is however hard to interpret without multiple reference cohorts and values.

### 4.2.2 Phenotypic-Genomic proximity evaluation

In order to prioritise variants based on phenotypic clustering, the hypothesis that clinical similarity assessed through HPO semantic similarity correlates with shared genomic features needs to be evaluated first. Therefore, I have compared the semantic similarity scores between pairs of patients against the number of genes and pathways presenting *de novo* variants that they share. The hypothesis is that phenotypically close patients should share DNMs in common genes and/or pathways more often than phenotypically dissimilar patients.

#### *4.2.2.1 Phenotypically closest couples of samples*

To compare phenotypic proximity, I computed 5 symmetric matrices of patients' phenotypic similarity with the 5 previously described method of semantic similarity based on the non-redundant HPO terms lists. The first question that can be asked is: how often the two phenotypically closest patients share a DNM in the same gene? A straightforward way to answer this question is to compare the number of shared mutated genes between the closest couples of patients as compared to random couples of patients from the dataset. This can be done by counting the number of couples sharing at least one *de*

*novo* mutated gene (regardless of whether *de novo* variant is the same or not). This was assessed with the set of 2,620 patients with detected DNMs. This analysis was repeated in five different fashions (each time with the 5 similarity metrics), comparing:

- the phenotypically closest pairs of samples to 1,000 randomly selected pairs among the 2,620 samples
- the top 5 phenotypically closest pairs of patients to 1,000 randomly selected sets of 5 pairs of patients among the 2,620 samples (*i.e.*, at least one of the 5 pairs need to share a common gene with a DNM)

For this analysis, it was assumed that patients sharing a *de novo* mutation on the same gene, even with a different mutation, would suffer from a phenotypically similar disease. This seems to be a reasonable assumption as DNMs are rare events and patients from the DDD cohort are sporadic cases. When comparing the closest couples for each of the 2,620 patients to 1,000 randomly selected couples, a clear signal was found: closest samples shared more mutated genes than random, and this independently of the semantic similarity used (Figure 22). Therefore, it can be assumed there is a correlation between phenotypic proximity assessed with HPO and genomic similarity. However, out of 2,620 couples of maximum phenotypic similarity, only between 12 and 20 shared a common mutated gene ($0.5 \sim 0.8\%$), which is extremely low. If the analysis is pushed to the top 5 closest pairs of individuals for each patient, the signal is slightly improved, but it remains extremely low with $1.2 \sim 1.5\%$ of pairs of patients with shared mutated genes (Figure 23).

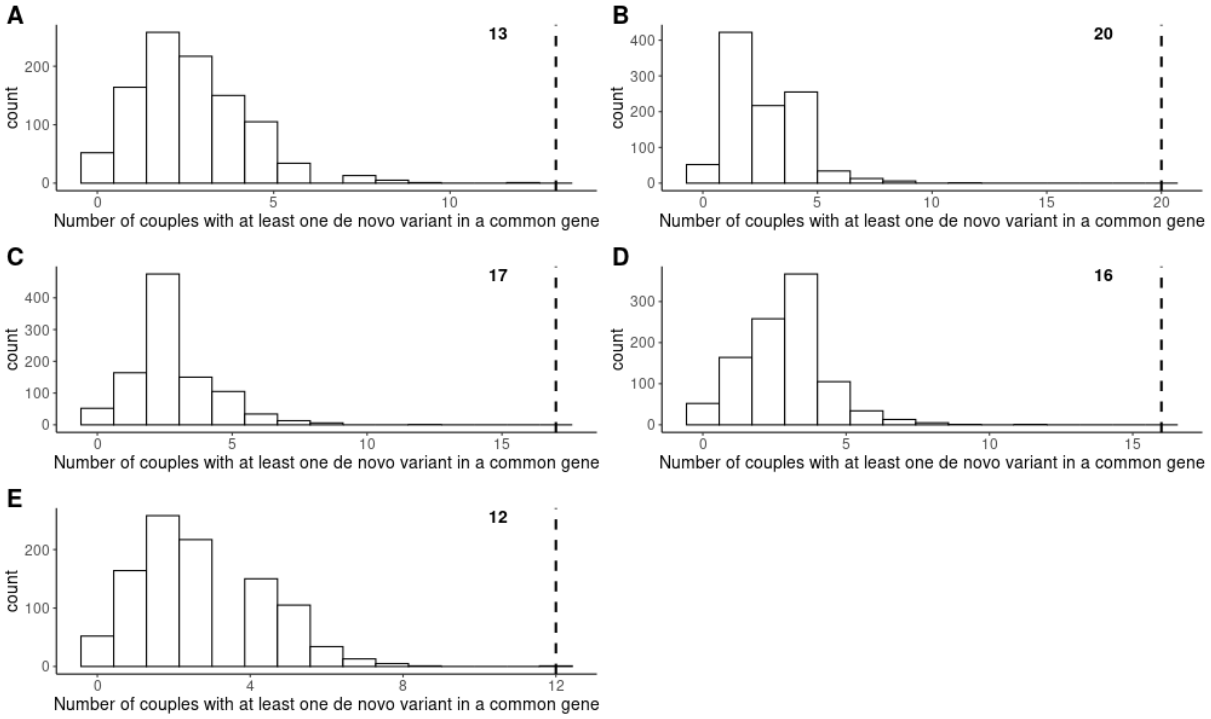



*Figure 22: Distribution of the number of couples with at least one* de novo *variant in a common gene*
*Distribution of the number of couples with at least one* de novo *variant in a common gene in 1,000 randomly selected couples of patients and number of phenotypically closest couples with at least one de novo variant in a common gene (dotted bar) computed with the 5 semantic similarity measures: Resnik (A); Lin (B); Information Coefficient (C); Jiang-Conrath (D); ERIC (E)*

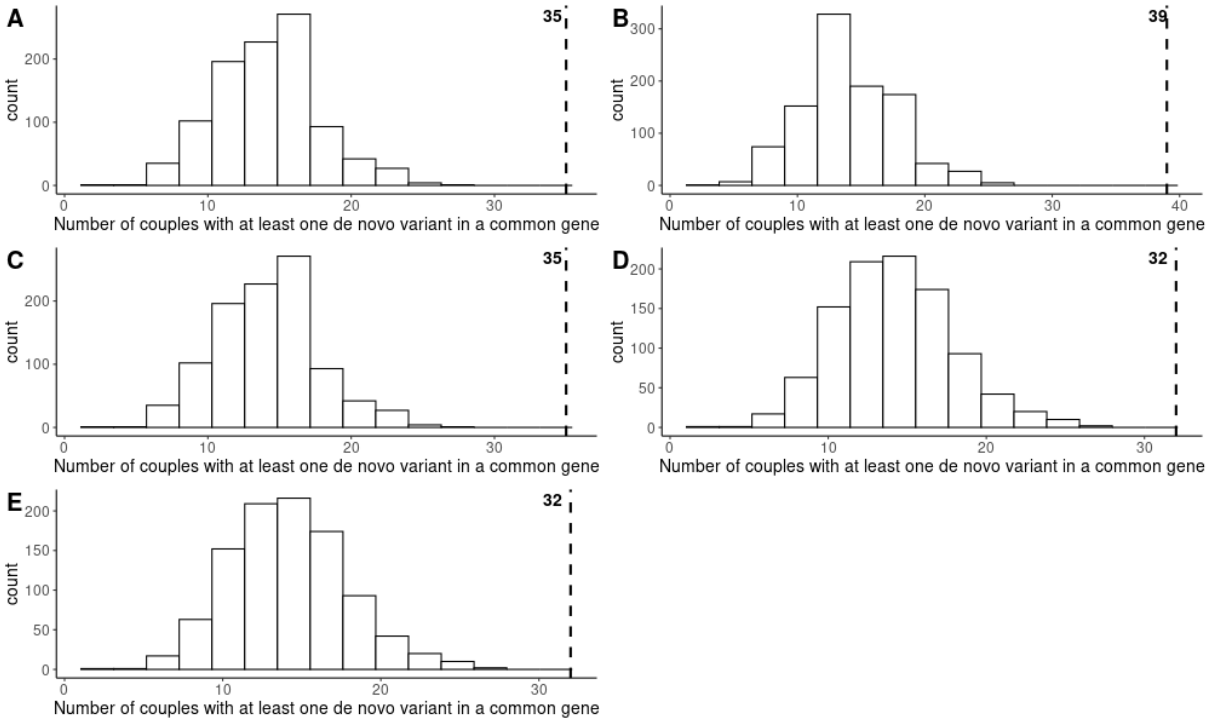



*Figure 23 : Distribution of the number of couples with at least one* de novo *variant in a common gene for the top 5 closest couples*
*Distribution of the number of couples with at least one* de novo *variant in a common gene in 1,000 randomly selected sets of 5 couples of patients and number of top 5 phenotypically closest couples with at least one de novo variant in a common gene (dotted bar) computed with the 5 semantic similarity measures: Resnik (A); Lin (B); Information Coefficient (C); Jiang-Conrath (D); ERIC (E)*

We then evaluated whether the clinically closest patients shared DNMs in genes involved in common biological processes or pathways. To that aim we used 3 different databases: Hallmark, Reactome and KEGG. Only Reactome database will be presented here as the other two databases were constituted of gene sets of too large size, which introduced too much noise (*cf.*, Supplementary figure 2, Supplementary figure 3, Supplementary figure 4, Supplementary figure 5). DNMs in the DDD cohort collectively involved 1,455 Reactome pathways across 2,086 samples. Analogous analyses as in the previous section were then performed.

Thus, for each patient, we identified the closest patient in the cohort in terms of clinical similarity and assessed the number of such pairs that shared a DNMs affecting the same pathway. Here, between $16.9{\sim}18.5\%$ of pairs shared at least a common pathway, clearly deviating from a background distribution derived from random pairs (Figure 24). Such observations were consistently observed independently of the semantic similarity used. When reproducing the analysis with top 5 closest samples (Figure 25), this percentage of pairs of patients sharing pathways targeted by at least one DNM raised to $46.9{\sim}51.1\%$. The previous results suggest that a pathway-based approach provides higher statistical power to identify common genetic factors between pairs of patients showing the highest clinical similarity.

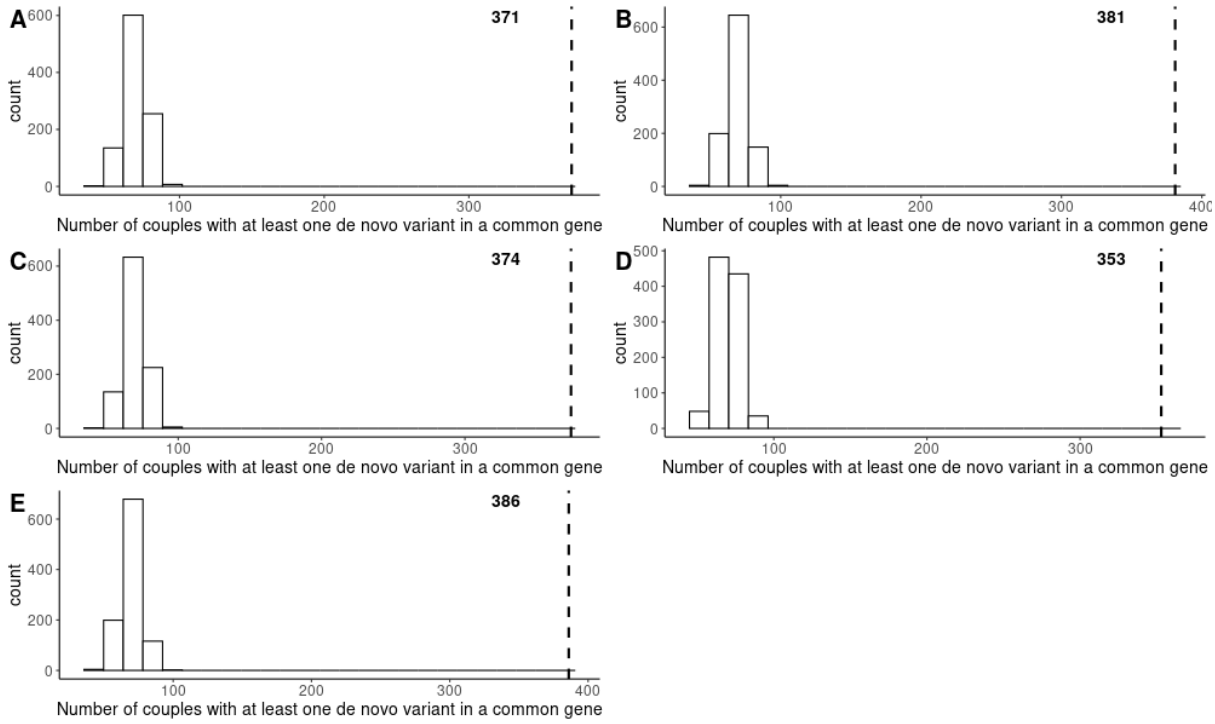



*Figure 24: Distribution of the number of couples with at least one* de novo *variant in a common Reactome pathway*
*Distribution of the number of couples with at least one* de novo *variant in a common Reactome pathway in 1,000 randomly selected couples of patients and number of phenotypically closest couples with at least one de novo variant in a common pathway (dotted bar) computed with the 5 semantic similarity measures: Resnik (A); Lin (B); Information Coefficient (C); Jiang-Conrath (D); ERIC (E)*

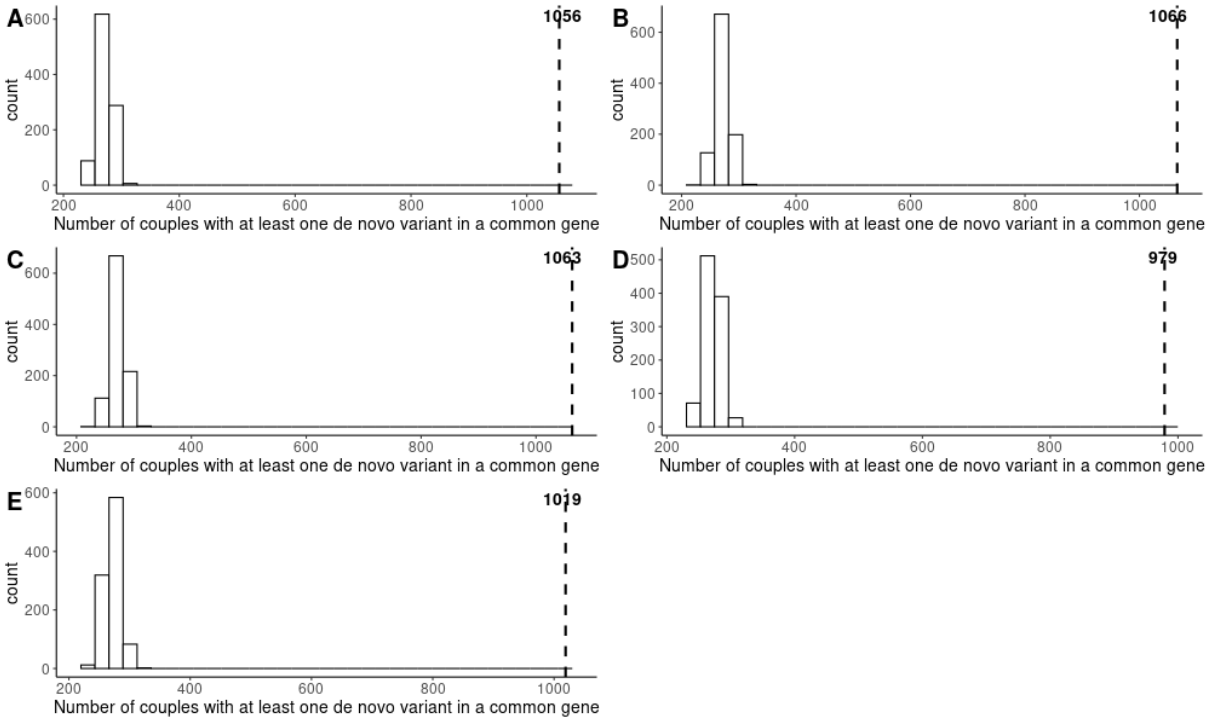



*Figure 25: Distribution of the number of couples with at least one* de novo *variant in a common Reactome pathway for the top 5 closest couples*
*Distribution of the number of couples with at least one* de novo *variant in a common Reactome pathway in 1,000 randomly selected sets of 5 couples of patients and number of top 5 phenotypically closest couples with at least one de novo variant in a common pathway (dotted bar) computed with the 5 semantic similarity measures: Resnik (A); Lin (B); Information Coefficient (C); Jiang-Conrath (D); ERIC (E)*

Overall, the previous results show that phenotypically closest couples of samples do share DNMs in genes and pathways more than expected by chance. In other words, clinical proximity assessed by semantic similarity computation of HPO terms reflects a genomic proximity for the closest patients. Gathering DNMs by pathways rather than genes provided a higher statistical power in this assessment.

#### *4.2.2.2 Systematic evaluation*

In the previous section we restricted the analysis to the top-1 or top-5 most clinically similar patients and showed that such pairs of individual shared DNMs in genes and pathways more often than expected by chance. Here we further evaluated whether such conclusions could be extended to the total set of individuals for which their clinical similarity with a query patient was significantly higher than random. To that aim, p-values were assigned to semantic similarities using Monte-Carlo simulations, as described in section 4.1.6. This consists in checking which couples of patients are phenotypically closer than expected by chance and then count how many of them do share common *de novo* mutated genes or pathways.

The proportion of significant couples sharing a DNM, out of all possible pairs $(N \times (N-1)/2)$ was then assessed. Independently of the similarity measure or the IC, the percentages of significantly close couples of patients sharing a *de novo* mutated gene was still very low, below 1% (Table 15).

*Table 15: Percentages of significantly and non-significantly close DDD couples sharing a* de novo *mutated gene*

| | | Resnik | Lin | IC | JC | ERIC |
|---|---|---|---|---|---|---|
| **IC_f** | **Significant** | 0,24 % | 0,24 % | 0,24 % | 0,25 % | 0,25 % |
| | **Non-significant** | 0,20 % | 0,19 % | 0,19 % | 0,18 % | 0,18 % |
| **IC_n** | **Significant** | 0,27 % | 0,26 % | 0,27 % | 0,27 % | 0,28 % |
| | **Non-significant** | 0,22 % | 0,18 % | 0,19 % | 0,19 % | 0,21 % |

However, if those percentages are compared to the numbers of non-significant couples, the percentages appeared to be similar. Thus, less than 1% of pairs of samples that appeared as significantly close in terms of phenotypic profiles shared a DNM on the same gene. Similar figures were obtained, however, for pairs of samples not being significantly close regarding their phenotypes. For the five semantic similarity metrics, the number of significant couples per sample is high ($mean \in [1{,}264; 1556]$ out of 2,620 samples for the 5 metrics, Figure 26). The Monte-Carlo method to assess p-values might be too sensitive, or the semantic similarity computation itself could also be too sensitive on a complete cohort to assess patients' similarity significancy.

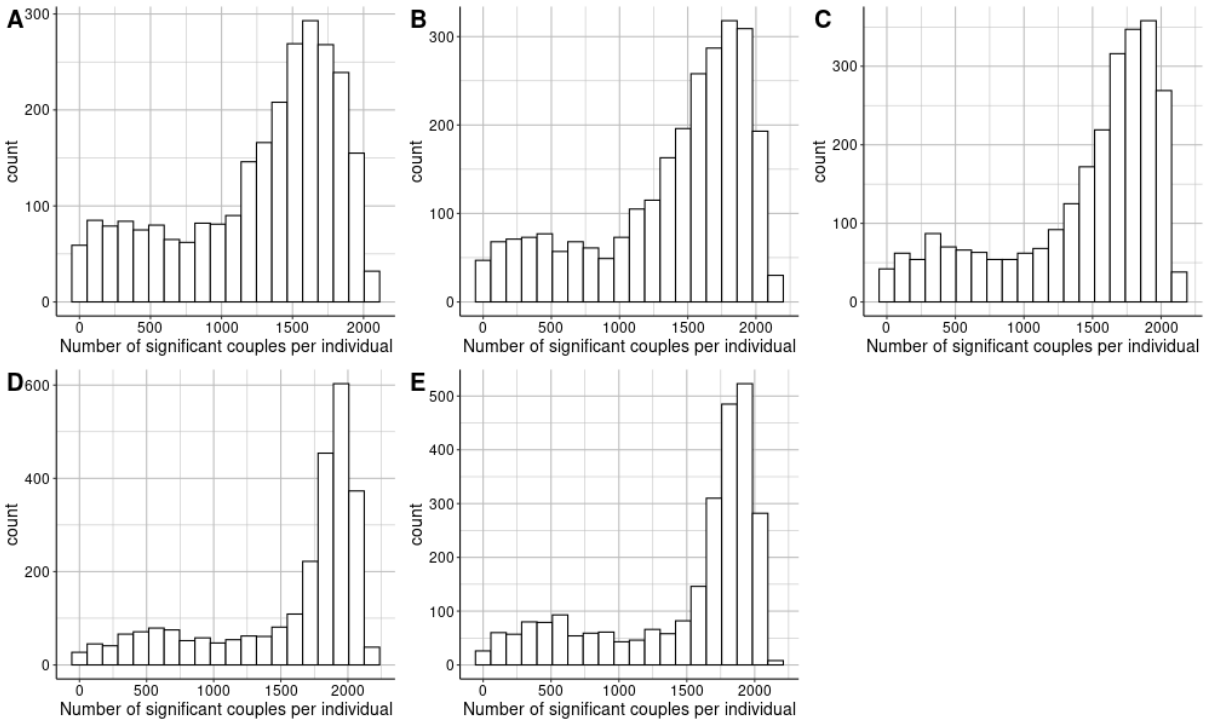



*Figure 26: Distribution of the number of significantly close couples per sample computed with the 5 semantic similarity measures*
*Resnik (A); Lin (B); Information Coefficient (C); Jiang-Conrath (D); ERIC (E)*

In order to gain in power, I gathered DNMs in pathways rather than genes, as done in the previous section. Here I computed the percentages of couples of samples sharing a mutated pathway using the *de novo* tables produces previously and collapsing the variants per pathways rather than per genes. On the one hand, some variants can thus be in several pathways as genes exist in multiple pathways

but on the other hand, we may lose variants which target a gene that is absent from any pathway. Because of these last considerations, I reproduced the same analysis for 3 pathway datasets: GSEA Hallmark, Reactome and KEGG (*cf.*, 4.1.7).

*Table 16: Percentages of significantly and non-significantly close DDD couples sharing a de novo mutated pathway*

| | | | Resnik | Lin | IC | JC | ERIC |
|---|---|---|---|---|---|---|---|
| **Hallmark** | **IC_f** | **Significant** | 11,46 % | 11,40 % | 11,44 % | 11,54 % | 11,56 % |
| | | **Non-significant** | 10,90 % | 10,90 % | 10,85 % | 10,70 % | 10,70 % |
| | **IC_n** | **Significant** | 11,57 % | 11,49 % | 11,59 % | 11,54 % | 11,74 % |
| | | **Non-significant** | 11,13 % | 10,84 % | 10,94 % | 10,89 % | 11,10 % |
| **Reactome** | **IC_f** | **Significant** | 35,72 % | 35,83 % | 35,82 % | 36,14 % | 36,02 % |
| | | **Non-significant** | 34,79 % | 34,62 % | 34,57 % | 34,12 % | 34,30 % |
| | **IC_n** | **Significant** | 35,46 % | 36,32 % | 36,31 % | 36,30 % | 34,54 % |
| | | **Non-significant** | 35,18 % | 34,24 % | 34,70 % | 34,52 % | 35,21 % |
| **KEGG** | **IC_f** | **Significant** | 12,26 % | 12,16 % | 12,13 % | 12,16 % | 12,20 % |
| | | **Non-significant** | 10,31 % | 10,24 % | 10,19 % | 10,02 % | 10,04 % |
| | **IC_n** | **Significant** | 12,17 % | 12,14 % | 12,25 % | 12,01 % | 12,53 % |
| | | **Non-significant** | 11,05 % | 10,27 % | 10,60 % | 10,55 % | 11,01 % |

For all pathway databases considered, the percentages of shared DNMs for couples of significantly clinically close patients are higher than for genes (Table 16). In the case of Reactome, they are above 35% for both IC types of computation. Nonetheless, just like for genes, percentages are equivalent for significantly and non-significantly close couples. The previous results suggest that being significantly similar from a phenotypic point of view is not a sufficient condition to share a genetic component, either at the gene or at the pathway levels. Such common genetic factors only appeared for the top-most similar pairs.

## 4.3 Discussion and conclusions

In this work, I have evaluated the capacity of 5 semantic similarity metrics on the HPO to assess the clinical proximity of patients suffering from developmental disorders and see how well it reflects their shared genetic factors in terms of DNM in common genes or pathways. I have analysed the HPO phenotyping of DDD patients with Cohort Analyzer and have observed that the cohort is almost as deeply covered as the ontology itself allows. My results showed that the DDD cohort is not poorly annotated as compared to the expectations draws from the global ontology.

Throughout the analyses, I used 5 widely used semantic similarity measures, which led to highly similar results. However, the alternative ways of computing Information Content of HPO terms had a higher impact in downstream analyses than semantic similarity computation. However, both methods (IC_f

and IC_n) performed similarly for capturing shared genetic factors, as observed through Monte-Carlo simulations (Table 16).

The clinically closest patients assessed through HPO semantic similarity carry DNMs in common genes or pathways. So, HPO semantic similarity could be used to drive a variant or gene-prioritising method for very close individuals. For instance, unsolved cases within a cohort could benefit from the clinical proximity with another patient to uncover variants that might have been missed by classical filtering and prioritising methods. The Monte-Carlo analysis also showed that semantic similarity computation does not correlate well with genomic proximity of patients on the whole cohort. This probably cannot be explained by an excessively poor HPO annotation of the cohort because the cohort IC depth is not dramatically different from the whole ontology. Either the ontology is underpowered to reveal the genomic architecture of all diseases (whether it is because the cohort is too heterogeneous, because the ontology is not precise enough or because terms used in DD are not precise enough), or semantic similarity is too sensitive to assess significancy between patients' similarity in a whole cohort with heterogeneous annotation of patients. Another factor to consider is that *de novo* variants are strong indicators but also very rare and thus not the only option to justify complex genomic architectures of a disease (*i.e.*, they might not be sufficient).

Therefore, the clinical proximity assessed with HPO semantic similarity computation for variant prioritisation seems to be a robust methodology for well-annotated close patients. However, the genetic signal only appears for the most significant pairs of individuals. The methodology cannot be used in a systematic manner the same way Phenomizer did. More work needs to be done to achieve a systematic method for phenotypically-guided prioritisation of variants. Several solutions can be considered to improve the accuracy of semantic similarity to show genetic proximity. First, only working with deeply phenotyped patients that have at least $n$ HPO terms of the highest possible level. Additionally, the use of methods to extend the phenotyping of patients could improve this step. For example, methods exist to automatically extract symptoms from electronic health records and convert them into HPO terms[162,163]. Finally, the use of other semantic similarity approaches that are not based on IC computation could provide more accurate results, for instance, the use network-based methods[164].

# CHAPTER 5. DISCUSSION

During my thesis, I have had the chance to work on two very different methodologies within the clinical bioinformatics field. I experienced two different scenarios as the use of HPO for gene and variant prioritisation is an extensively studied topic whereas case-only burden test is a niche within the field. To my knowledge only two teams claimed to have developed a case-only burden test while hundreds of papers have been released on the use of HPO for clinical genetics during my thesis. On the one hand, I had the chance to be part of a collective effort to improve rare disease diagnosis by using clinical data in the form of HPO and on the other hand, I felt lucky to work on a limited topic that can be relevant to other researchers and physicians. For instance, as I write, CoCoRV was released only two months ago and many research teams are still publishing new burden test frameworks in recent years[156,165–167].

This chapter will be dedicated to a brief discussion of the results of my work and its limits as well as future developments and perspectives that could be brought to COBT and the use of HPO for gene prioritisation. Finally, I will describe prospects that can be contemplated and challenges that will be addressed for rare disease association methods in clinical bioinformatics.

## 5.1 DISCUSSION ON COBT AND THE USE OF HPO SEMANTIC SIMILARITY FOR GENE-PRIORITISATION

### 5.1.1 HPO semantic similarity for gene prioritisation

At the very beginning of my thesis, when discussing the project with my supervisor, I was very excited about the idea to cluster patients in order to define homogeneous groups of patients and prioritise variants for rare disease analysis. I wanted to create a method to automatically cluster patients within a heterogeneous cohort to build sub-groups of similar patients. As HPO terms of patients were collected based on their clinical records, the fact that HPO terms would show the genetic proximity of patients was closer to a certainty than a hypothesis to me. Furthermore, hundreds of papers of methods using HPO for gene-prioritisation already existed. Unfortunately, I spent all my thesis trying to gather evidence that in a large cohort phenotypically similar patients would share a common genetic information. Perhaps, other researchers have come up to the same conclusion and this is why the vast majority of papers use clinical similarity to prioritise genes with existing knowledge (i.e., OMIM for most of them), nonetheless, I never read a paper showing that. The initial idea remains very interesting but some more work on the HPO needs to be done to apply such methodology with information content and semantic similarity systematically for variant and gene prioritising.

Natural Language Processing (NLP) is a field within artificial intelligence that aims at "understanding" human language. Several methods have been developed to interpret electronic health records in order to summarize the information[162,163,168]. There has been evidence that such method, namely EHR-Phenolyzer, helps improving the diagnosis rate[162]. However, the accuracy of such method is far from being as good as one would expect. Even with a deep phenotyping of EHR to get HPO and prioritise genes, EHR-Phenloyzer manages to prioritise the right gene among the top 100 genes for 16 out of 28 patients. Of course, applying NLP method on top of already phenotyped patients to get more HPO terms is different, but the results might not be fundamentally better than the one presented in this work to get clusters of similar patients and prioritise genes.

We could not demonstrate that clinical semantic similarity can help differentiate genetically similar patients within a full cohort. As stated in the results section, this does not mean that the hypothesis is wrong because it holds true for the top closest pairs of patients. On the one hand, there could be a lack of power and increasing the number of HPO and their accuracy (*i.e.*, level in the ontology) could solve the issue, or, on the other hand, perhaps that semantic similarity assessed with information content is not powerful enough to show genetic proximity. Using the HPO as a simple graph without taking advantage of the hierarchy (*i.e.*, without using semantic similarity) could improve the power of the associations. Peng and colleagues proposed to measure the phenotype-to-phenotype similarity by incorporating biological network information to assess the similarity[164]. Their method, PhenoNet, has not been tested on clinical dataset of patients. However, as compared ontology-based method (Resnik with information content), the correlation between terms proved to be more than twice as much for PhenoNet. The problem with such method is that it lies on supplemental data which can thus not be accessible for all terms.

### 5.1.2 Modifier variants for ciliopathies

COBT framework has been applied to a cohort of patients suffering from various ciliopathies with the hope of discovering new associations and causal genes but also potential modifier variants. Most ciliopathies are recessive, meaning homozygous variants need to be found by clinicians to diagnose the condition. When I presented my work to some researchers specialized in ciliopathies, they were skeptical. However, even among patients with the same symptoms, there can be discrepancies in the sets of mutated genes. Several conditions among ciliopathies are already known for having large phenotypic heterogeneity and locus heterogeneity such as Bardet–Biedl syndrome for which oligogenic modifiers have been identified[169]. A recent study of 99 Bardet–Biedl syndrome patients have been deeply analysed looking for mutational load in 39 Bardet–Biedl syndrome associated genes[170]. The team showed with the help of *in silico* tools and clinical annotation that several genes associated with the syndrome were correlated with more severe phenotypes. This validates our hypothesis that

in rare autosomal recessive ciliopathies, mutational load can alter the phenotype. The case presented in this study is exactly the type of study in which COBT could be useful: the cohort is composed of cases with heterogeneous phenotypes but suffering from the same rare disorder, however, no common pattern has been found. Furthermore, here, as compared to our ciliome study, patients were sequenced in WES or with more recent kits than the ones used in our study, suggesting that the overall quality might be better. Hence, it would be very interesting to re-analyse their data with COBT or eventually to merge their dataset with ours to apply our pipeline. With more recent sequencing data that supposedly have a better coverage, QQ-plots might not be over-inflated.

### 5.1.3 QQ-plots inflation

Whether using CoCoRV, TRAPD or our method, the QQ-plot showed an inflation of p-values. The inflation of p-values means that they are collectively lower than expected. This could be due to a filtering that is too soft. However, in this work, I have followed the same filters as TRAPD, CoCoRV and state-of-the-art recommendations for exome studies and burden tests. For the 1000 Genomes study, I even had a harder filter for coverage. Thus, it is unexpected to observe such an inflation that is not happening in CoCoRV's nor TRAPD's paper. I contacted the author of TRAPD to get help in understanding the behaviour of the data, but I never received an answer. If I had access to the filtered dataset used in TRAPD or CoCoRV's paper, I would be able to have an even better comparison of our methods. It would be very interesting to apply the three methods on their data and see how COBT performs in comparison to the others. My expectation is that COBT should be more stringent than CoCoRV and TRAPD, thus having less false positive. Nevertheless, COBT might have more false negatives than CoCoRV and overall, not perform as good as CoCoRV for specific genes under balancing selection that carry few variants but that are shared among different ethnicities. CoCoRV with its FET test would easily consider different ethnicities and count the number of mutated controls while COBT could fail to detect it if few sites exist.

### 5.1.4 Technical considerations in COBT pipeline

Initially, just like in Samocha's initial framework, I used the trinucleotidic context to assess the probability of mutation of genes, corrected by the ratio of mutations from gnomAD as a proxy to represent the general population. This was extremely complex in term of code and hugely longer to run. I checked the differences between using this complex strategy and simply using the ratio of number of mutations from gnomAD as parameter of the test and the results were similar. Therefore, I integrated only the number of variants from gnomAD into the test. This makes sense that results are not so different: the number of variants per gene is highly correlated with gene length[42] and in gnomAD, with such a high number of samples, the number of variants per gene also represents the variability of mutations available. The correlation between the probabilities computed with the two

methods was high. Unfortunately, I do not have this data as I did not keep track of it and as the framework has dramatically changed since then (notably, the test filtering is stricter and I converted my pipeline to hail, so, it would be very long to re-adapt the previous method).

One topic that seems overlooked in exome and targeted sequencing studies is the transcript selection. The strategy is often not even indicated in the papers while it could be of high relevance in some organ-specific related disorders in which the variant consequence on different isoforms of the protein can be highly relevant. By looking into details into the code, TRAPD is for example selecting all the canonical transcripts of the qualifying variants and keeps the first one (https://github.com/mhguo1/TRAPD/blob/master/code/make_snp_file.py). We chose to use APPRIS database over canonical transcripts because in GRCh37, only the longest CDS is used to define canonical transcripts while APPRIS uses functional annotation and cross-species conservation. I believe it is important to have a detailed explanation on the choice for the transcript selection in GRCh37. Nonetheless, in GRCh38 the canonical transcript is also based on APPRIS (among other resources), hence, in future studies using GRCh38, using the canonical transcript will probably become the gold-standard.

### 5.1.5 Perspectives for COBT

COBT has been used on genes and on biological pathways with the hope to gain in power. However, the opposite strategy of assessing the counts per exon rather than per gene is also a method that could have been interesting to test, not for power reasons but for biological relevance. The rationale for this strategy is that particular variants (even synonymous mutations) can affect specific exons by altering the splicing[171,172]. The case of two children suffering from hepatorenal fibrocystic disease (a ciliopathy) was solved through re-analysis and curation of WES data combined with homozygosity mapping; the authors discovered a synonymous variant in NPHP3 for the two patients and further patient-derived RNA analysis showed an activation of a mid-exon splice donor leading to frameshift[173]. This proves the existence of exon-specific mechanisms even with synonymous variants that could be detected with a more fine-tuned approach. For that, statistical power needs to be maximized with WES data of better quality and more patients than in our study. There have been examples of such approaches with more classical burden test framework (SKAT) using protein domains and family coordinates but without strong evidence of association[174].

As will be discussed deeply in the next section, recent work gave insightful and broader description of the non-coding genome which represent more than 98% of the whole genome. This could be a chance to adapt COBT framework to new genomic regions that are not necessarily bound to gene sequences but rather broader biologically relevant units such as Topologically Associated Domains (TADs). TADs

are 3D regions that include genes as well as their promoter and regulatory elements (silencers and enhancers) interacting between each other much more frequently than regions outside of the TAD[175]. TADs are bounded by insulator proteins that supposedly stop interactions between promoters and regulatory elements[176].

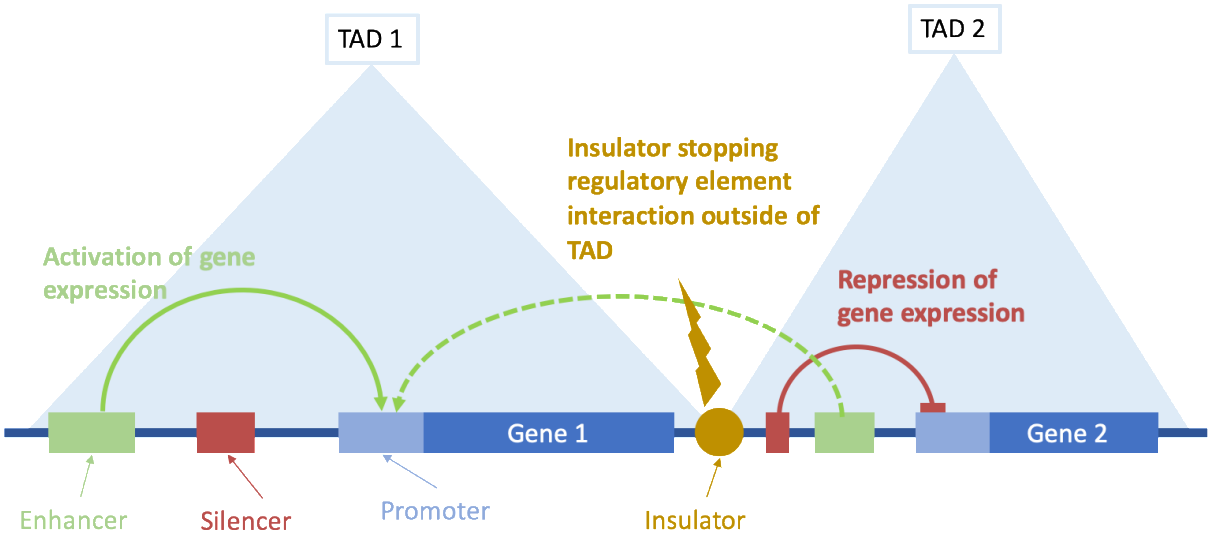



*Figure 27: Schematic functional representation of a TAD organization with regulatory gene expression (adapted from Bocher and Génin[156])*

Enhancers and silencers respectively activate and repress the promoter of a gene in specific interactions implying transcription factors and other proteins thanks to close contact in the 3D conformation of the DNA strand. These interactions are studied through Hi-C sequencing and are hugely complex[176]. A disruption of a regulatory element would interfere with the gene regulation, hence, protein production that might lead to a disease-phenotype. In an exome study, such effect would be invisible. Thus, adapting COBT to TADs could increase its applicability by allowing users to look for potential non-coding disease-related rare variants that make sense regarding the 3D structure of DNA. Non-coding rare variant methodologies, including burden testing, have already been published recently[165].

## 5.2 Prospects and challenges

### 5.2.1 A note on computational speed and user-friendliness of bioinformatics methods

CoCoRV is very long to run and unusable for researchers or clinicians not familiar with computational biology. The whole TRAPD pipeline has run in 4 to 6 hours depending on the dataset (synonymous or missense for the 1000 Genomes Project). COBT runs in 2 to 3 hours (filtering process included). To optimize CoCoRV, I have splitted the pipeline by chromosome and specifically for missense chromosome 2 has been running for 6 hours on a 640Go server with 72 threads. On the same note,

CoCoRV's pipeline is well-documented, but the program itself is not user-friendly. TRAPD was easier to use and the documentation clearer. A lot of bioinformatics methods are not designed to be applied by user who are not trained to bioinformatics which dramatically reduces their utility. Moreover, each method is coded using different languages. From what I have seen, at least in France, most biologists and clinicians who are comfortable with R for biostatistics and computational biology. I believe most methods should thus at least apply the statistical part (outside of the filtering) in R so that users can try to optimize or adapt the script to their use. One of the major problems I encountered while trying to run TRAPD and CoCoRV's pipelines is the installation of all required libraries. For the sake of reproducibility, versions of each library were provided, however, they were not always easily accessible nor compatible, especially for CoCoRV.

To avoid this limitation, COBT is provided with a Conda environment (https://docs.conda.io) so that the user can directly install the exact same software and libraries versions as the one described in my results. Conda is an open-source package management system initially built for Python programs that works on all classical operating systems (*e.g.*, Windows, MacOS, Linux) and allows users to quickly and easily install, run, update packages as well as all dependencies. Users can work on separate environments, allowing them to have several versions of the same programs without conflicts. For bioinformatics, a Conda channel called Bioconda has been created, gathering more than 8,000 tools[177]. More and more bioinformatics programs propose a Bioconda environment to install required package dependencies or directly install the program, such as Burrows-Wheeler Aligner[14], Ensembl VEP annotation program[61] or CADD[59] among hundreds of others. The advantage is that Conda is very simple to use and allows user to manage different projects in several environments to avoid conflictual dependencies. Others choose to provide a docker container images (https://www.docker.com). Docker utilise OS-level virtualisation to provide the user with containers that are lighter virtual machines. Thus, Docker simulates a whole OS rather than just an environment, hence it is a more stable solution than Conda but not as easy to deploy in my opinion. Several bioinformatics software programs provide a docker container image, such as APPRIS[65] or Ensembl VEP annotation program but especially machine learning programs including Xrare[118] and many deep learning programs[178].

### 5.2.2 A note on reproducibility

According to several studies[179–183], there is a reproducibility issue in science and especially in biomedical research where up to 85% of the research investment is "wasted" (*i.e.*, 200 billion dollars in 2010)[184]. This reproducibility issue is a reality but 90% of 1,500 scientists interviewed in 2016 shared the feeling that there was indeed a "reproducibility crisis"[181]. There are different threats to the reproducibility in research: cognitive biases, transparency of the methodology, quick and poor peer-

review, conflicts of interest etc. Several solutions have been proposed to address these threats, such as encouraging or rewarding transparency and open-science to improve reproducibility as well as diversifying peer-review thanks to preprints for example. Nonetheless, some biases are harder to counter such as the tendency to not publish negative results, leading other researchers to reproduce the same mistakes and even worse to sometimes make-up negative results in order to get published. This is a vicious circle because as journals are in competition to publish exciting papers, they tend to accept only ground-breaking results, hence, negative results are not published or published in a low impact journal. The last decade has seen the creation of several journals aiming at publishing only negative results, such as the open-access Negative Results (https://www.negative-results.org/) and Journal of Negative Results in Biomedicine for example.

Even if data sharing seems like a mandatory aspect of scientific publication, for privacy protection it is not always so easy in the biomedical field to make personal data available, especially genomic data. This of course makes the reproducibility of the research very complicated. For bioinformatics methods, many studies provide the simulated data on which they tested their framework. However, the behaviour of a method on simulated data and on clinical data might be very different. With the ever-growing publicly available genomic databases, if most researchers were to test their pipeline on the same deeply studied data, this would hugely affect the reproducibility of all studies by more often pointing out mistakes and avoid data fabrication.

In the bioinformatics community several initiatives have been conducted to improve the reliability of their studies. Sandve and colleagues proposed 10 simple rules for the reproducibility of computational research[185]. Some of these pieces of advice might seem logical but are not always so easy to apply on a daily basis, such as recording all intermediate results or avoiding manual data manipulation steps. For some other rules, informatic solutions have been developed to help bioinformaticians in their task, for instance version control system programs, Git being the most well-known, that is also widely used to store programs and open-source code. Most bioinformatics programs have a dedicated GitHub or GitLab webpage. The Bioconductor project[186] and Galaxy[187] are initiatives that fulfil most rules described in *Sandve et al.*'s paper. If they follow their rules, researchers can distribute their own methods via these cloud solutions. However, they can be a bit rigid and not be suited for particular pipelines. In such cases Conda or Docker might be better solutions. Kulkarni and colleagues proposed a solution to distribute docker applications and an easy framework for R users in the bioinformatics community called the Reproducible Bioinformatics Project[183]. The project had the great ambition to gather simple docker implementations of bioinformatics software as well as more complex pipelines. However, the website does not seem to be updated (http://www.reproducible-bioinformatics.org/).

With the number of available solutions, it is probably too much of a constraint to adapt one's pipeline to a standard. Therefore, it appears simpler for bioinformatic software developers to use workflow managers that will simplify pipelines and keep track of software installation and versions. Galaxy has already been mentioned but other alternatives implemented as domain-specific languages (*i.e.*, programming languages developed for particular application domains) are more flexible. Nextflow[188] and Snakemake[189], that is based on python programming language, are probably the most famous among bioinformaticians because of their simplicity. The strength of such tools is that they ensure portability of the pipelines while increasing their reliability and reducing their complexity for users.

At some point in my thesis, I wanted to convert all my pipeline to Snakemake as I was familiar with python language. However, I attended a Nextflow course and realized neither Netflow nor Snakemake would be compatible with hail that uses lazy programming and parallelization. Therefore, I chose to simply use Conda workflow manager to keep track of package versions and for ease of deployment. I believe the bioinformatics community would gain a lot in terms of reproducibility and clarity if more software programs, especially the ones supposed to be used for clinical research, were deployed with workflow manager environments and not only version control system programs.

### 5.2.3 Lifespan of a clinical bioinformatics tool

Since the release of HPO, a multitude of gene-prioritising methods have been published. In Table 17, I showed the number of Google scholar citations a handful of the methods I am aware of. Many methods were published between 2018 and 2019 (roughly at the beginning of my thesis). Most of them compare their results with Exomiser, PHIVE or eXtasy. Hence, unsurprisingly they are the most cited papers. Among the recent papers, several are almost not cited at all such as GPSim[190] and Phenogenon[191]. Among the most cited papers since 2019, there is ClinPhen[192] which is a method that automatically transforms patients' records to prioritised HPO terms. This method is very interesting because it can not only speed up the gene-prioritisation process, but also replace the tedious step of symptom HPO matching for clinicians. Surprisingly, many articles citing ClinPhen are not clinical studies but rather other bioinformatics methodology studies, for example:

- Deep phenotyping: embracing complexity and temporality—towards scalability, portability, and interoperability (https://doi.org/10.1016/j.jbi.2020.103433)
- Doc2Hpo: a web application for efficient and accurate HPO concept curation (https://doi.org/10.1093/nar/gkz386)
- Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease (https://doi.org/10.1093/jamia/ocz179)

- Natural language processing to facilitate breast cancer research and management (https://doi.org/10.1111/tbj.13718)

The 2 other most cited methods of 2019 (Table 17) are Xrare[118] and DeepPVP[193] which are respectively a machine learning (XGBoost) and a deep learning method. These two studies are cumulating both the use of HPO for gene-prioritising and machine/deep learning methods which are two hot topics in the field. Just as for ClinPhen, many articles citing those two studies in Google scholar are related to bioinformatics method development. The factual use of all these methods in a clinical setting is hard to quantify and looking only at the number of citations in Google scholar is very disputable (looking at the altmetrics would have probably been better but not all papers had this index).

*Table 17: Number of Google scholar citations of a collection of HPO gene-prioritising methods*

| Method | Year | Number of Google scholar citations | Benchmark data | Benchmark methods |
|---|---|---|---|---|
| **Phenotype–disease matching tool for rare genetic diseases**[194] | 2018 | 10 | Simulated | Phenomizer, BOQA |
| **Xrare**[118] | 2019 | 40 | Simulated & clinical | Exomiser, eXtasy, CADD, REVEL, MutationTester |
| **HANRD**[195] | 2018 | 32 | Clinical | Phenomizer, BiRW |
| **eXtasy**[126] | 2013 | 163 | Simulated | PolyPhen, MutationTaster, SIFT, CAROL |
| **PhenIX**[128] | 2014 | 232 | - | - |
| **Phen-Gen**[196] | 2014 | 143 | Simulated | VAAST, PHEVOR, eXtasy |
| **Exomiser/PHIVE** [127] | 2014 | 313 | - | - |
| **Phen2Gen**[197] | 2019 | 30 | Clinical | Phenolyzer |
| **PhenoRank**[197] | 2018 | 30 | Simulated | Exomiser, PRINCE, DADA |
| **RelativeBestPair**[121] | 2018 | 14 | Simulated | Different semantic similarity computation (Resnik, IC, JC, Lin, Wang, RBP, GraphIC) |

| | | | | |
|---|---|---|---|---|
| **GPSim**[190] | 2019 | 6 | Simulated | Resnik, Zhang, BOG, SemFunSim |
| **ClinPhen**[192] | 2019 | 54 | Clinical | cTAKES, MetaMap |
| **DeepPVP**[193] | 2019 | 45 | Simulated | Exomiser |
| **Phenogenon**[191] | 2020 | 6 | - | - |

Nonetheless, all methods present in Table 17 are only a subset of all the methods I am aware of and there are probably much more methods than the ones developed in the articles I have read. But most of them are not very cited, in addition, they are usually cited in other bioinformatics method development papers. A legitimate question to ask is: do these methods really have a clinical usefulness? And will they all be maintained? Furthermore, most methods provide a benchmark in which their method usually perform best. Exomiser is usually considered the gold-standard method, thus included in the benchmark, but other selected methods are often not the same. Obviously not all methods are comparable, but many benchmarks are applied on simulated data and not all studies include a clinical analysis benchmark. To my knowledge there are only 2 studies comparing most available HPO gene-prioritising software methods. Pengelly and colleagues published a benchmark in 2017[198] comparing most available methods including PhenIX, Phen-Gen, hiPHIVE and eXtasy. They reanalysed 21 clinical exomes and tried to rank the right causal genes. Their study showed that hiPHIVE (Exomiser) performed best. Earlier this year, a new benchmark has been published by Jacobsen and colleagues with researchers from the HPO consortium, namely Robinson and Smedley[199]. They reanalysed 4,877 solved cases from the UK 100,000 Genomes Project[200] thanks to phenotype-driven approaches. They first selected a collection of peer-reviewed freely available software programs for HPO-based prioritisation of variants through VCF format including Exomiser, Xrare, DeepPVP, Phen-Gene, eXtasy and others that were not mentioned previously. However, they chose to only compare Exomiser and LIRICAL[201] because they were the only tools running on both GRCh37 and GRCh38 with local installation and programming queries. The study underlines the importance of developing tools for noncoding variants as 4% of 100,000 Genomes Project pilot study molecular diagnoses are involving non-coding rare disease variants. Only Genomiser[202] (Exomiser framework adapted for non-coding genome), AnnotSV[203], SvAnna[204] and Phen-Gen handle non-coding variant rare disease analysis. On the same note, Exomiser and Phen-Gen are the only tools able to consider incomplete penetrance. The authors raise the problem of the lack of uniformity in the methods' input now that the HPO framework has become a standard in hospitals for deep phenotyping. Even if the study as a whole is disputable as the authors are the same researchers who developed the two benchmarked tools, this questions the

utility of the other non-selected tools in a clinical setting, especially because of the lack of upgrade to GRCh38.

The same issue can be raised with COBT which at the moment only works for GRCh37. The method would still be valid for GRCh38 but it would require to modify part of the pipeline. More importantly, gnomAD v3.1 is now providing individual genotypes for a subset of WGS data since 2021[205]. Illumina revealed in September of 2022 the soon to come release of NovaSeq X, their newest WGS machine that will produce WGS for 200$ per run while being twice as quick and three times more accurate[206]. Future studies will most probably not need case-only burden tests anymore if clinicians and researchers can sequence controls at low cost. The lifespan of bioinformatics tools is highly dependent on the available technology which evolves very fast. Whilst the clinical utility and impact of a bioinformatics tool is difficult to ensure, each technical development is a new step towards a better accuracy and hopefully better care.

### 5.2.4 Future challenges and developments for variant, gene-prioritisation methods and other genome analysis

#### *5.2.4.1 Non-coding genome*

The vast majority of the genome is non-coding and there has been evidence of disease-associated variants in enhancer elements, DNase hypersensitivity regions or chromatin marks identified through GWAS[60]. Non-coding disease-associated variants were identified in 5' or 3' UTRs of genes, but others were sometimes associated with non-coding genes[207,208] (*e.g.*, microRNAs and lncRNAs). Variants identified by GWAS are then examined in high-coverage resequencing studies to assess their disease association[209].

Interpretation of functional non-coding variants' regulatory role in gene expression has also progressed dramatically thanks to quantitative trait loci mapping approaches[210]. Consequently, bioinformatics methods have been developed to predict non-coding variants' roles in human disease. Such tools cannot rely on the same methodology as tools designed for coding variants because most of them are based on the quantification of the variant's impact on the protein sequence. Alternative methodologies have been developed, such as Genome-Wide Annotation of VAriants (GWAVA)[211] which allows the prioritisation of non-coding variants using genomic and epigenomic annotations trained on HGMD mutations. Several methods have been published to predict the molecular functional impact of non-coding variants, such as FATHMM-indel which can predict the pathogenicity of non-coding indels[212].

Identification of disease-associated variants for rare diseases cannot rely on the same techniques as for complex diseases because the latter are uncovered by GWAS. However, they cannot be identified using functional prediction neither as in exome studies because no predicted impact on the protein is available. New non-coding specific rare variant association tests have to be designed. Bocher and colleagues proposed a burden test method based on CADD scores to define CADD regions[165]. The strength of the method is that it provides a filtering of the variant that is also based on CADD to prioritise variants based on deleteriousness that outperformed other filtering strategies in the simulations.

Machine learning and deep learning methods are another promising way to establish the potential pathogenicity of non-coding variants. NCBoost for example is a tool using supervised machine learning (based on XGBoost method) to score non-coding variants and propose rare disease candidate variants[213].

#### *5.2.4.2 Artificial intelligence methods*

Artificial intelligence methods refer to sometimes machine learning but more often to deep learning methods. They have become extremely popular among bioinformaticians in the last decade. Many methods have been developed for various functions such as functional prediction as already mentioned, but also cell classification and single-cell data categorisation[214,215]. Artificial neural networks are a collection of connected units called nodes (usually numbers between 0 and 1) that can be trained with data to learn rules. Deep neural networks are artificial neural networks containing multiple layers. Several methods of deep neural networks have been developed including convolutional neural networks (*i.e.*, initial layers of convolutional nodes fully connected to later layers) and recurrent neural networks (*i.e.*, nodes arranged in the form of a chain). Deep learning methods need colossal amount of data to learn from and predict accurate result. Deep learning models can now predict splicing sites[216], prioritise non-coding sequences based on functional annotation and genomic sequence[217]. To many biologists deep learning methods are perceived as a "black box", sometimes rightly as some the rules can be invisible to the user in hidden layers with tens of millions of parameters learnt in the process. I will not go into details here because it is a very complicated topic that I am not extremely familiar with, but one thing I know is the "garbage in/garbage out" theory stipulating that if the data used for training is of poor quality, then the model will perform poorly because the rules drawn from it will be skewed[218]. While reading AI-related papers, despite the complexity and the black box issue, biologists need to carefully pay attention to the data used for training. A sadly notorious example of training set bias is the arXiv paper by Wu and Zhang that claimed to have created an AI capable of detecting criminals only using face images[219]. The reality is that the classifier learnt to recognize convicted criminals because of a selection bias: all photographs from

convicted criminals were official ID photographs whereas non-criminals' pictures were professional headshots in which most of them were smiling. The authors claimed their algorithm had a 90 percent accuracy and was using facial features of the photographs including lip curvature, eye inner corner distance, and nose-mouth angle. Fundamentally, their method was capable to discriminate smiling persons versus non-smiling persons. This turns out to be almost funny, but a bias that ends up in such interpretations could have devastating social consequences. To come back to the medical field, more and more papers about deep learning image classifiers are published for cancer recognition for instance. Such methods could someday become widespread in hospitals and an error could be fatal. In genomics and bioinformatics deep learning have also become extremely trendy these last few years[178,214,215].

Nonetheless, during my thesis I have seen many bioinformatics deep learning methods published every week. I believe more methods will continue to be released and bioinformaticians will need to get familiar with these methods, study their strengths and weaknesses to understand future publications or even start developing new methods themselves.

#### *5.2.4.3 Polygenic Risk Scores*

Polygenic Risk Score (PRS) is another very popular topic among bioinformaticians. Their aim is to predict disease risks by making use of the huge amount of GWAS results produced during the last two decades. PRS were introduced shortly after the first GWAS publications in 2007 by Wray and colleagues[220]. As the method is bounded by the heritability of a trait, this parameter is essential. The computation is quite simple: it is a sum of the product of each SNP's effect size to the genotype of the sample. Several improvements have been proposed such as the use of ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO) method or considering LD in the computation to measure the contribution of each genomic region[221]. By using of the tremendous number of variant-to-phenotype associations discovered through GWAS, the initial hope of PRS was to identify individuals with a disease-risks at low cost as only genotyping data is required. However, several hurdles to the use of PRS in human disease prediction have been raised. First, some assumptions of the model are wrong. PRS were developed following the same principles as Estimated Breeding Values (EBV) used in agriculture for quantitative phenotypes, for cattle reproduction for example[222] (*e.g.*, size of the offspring for meat production). Yet, PRS are meant to be used for qualitative binary phenotypes (sick versus not sick) which implies that an unobserved quantitative trait is transmitted over generations of patients. However, such model supposes that environmental factors are negligible, and this is not true in diabetes for example[223]. The other issue with PRS is population stratification that influences the

differences between cases and controls from the initial GWAS in which SNPs were identified (an issue that is considered in EBV).

While new technical improvements are still published regularly regarding PRS computation[224], PRS are under strong debate regarding their clinical use[221]. The use of PRS raises both ethical and technical issues. For instance, industrial applications have already been developed for prenatal diagnosis while the efficacy of such models is still highly debated. As previously stated, the method is simple and even though ultimately most bioinformatics methods in clinical research aim at providing clinical solutions, this should not be at the cost of scientific rigor. When discovering PRS during my thesis I thought it was an exciting and promising topic, however, I have learnt to be cautious with strong promises from overly simple models. Companies selling genetic tests such as 23andMe are using simple models to provide ethnical and geographical information to customers. Nonetheless, they are not harmful whereas companies selling genetic tests to predict disease-risk for prenatal diagnosis can have terrible consequences. GWAS themselves have been criticized for not providing what the scientific community was hoping for (*i.e.*, missing heritability), so, in my opinion, although there might be a true potential in PRS research it should be followed with care, especially outside of the academic area.

#### *5.2.4.4 The added value of multi-omic analysis to rare disease research*

Most of the topics I have discussed in this manuscript are related to single-omic methods, namely genomics. The aim of the HPO clinical data analysis was to integrate phenotypic and genomic data. I believe that data integration can result in a better understanding of rare disease mechanisms and an improvement in diagnosis rate. The use of phenomics to characterize the clinical symptoms, genomics and epigenomics to look for potentially harmful variants and finally transcriptomics, proteomics and metabolomics to assess the functional impact of mutations could unveil complex biological mechanisms of rare disorders (Figure 28).

As described previously, phenomics can be used together with genomics for gene prioritisation (*e.g.*, Exomiser etc), but transcriptomics can also be used to functionally validate candidate genes or variants. For instance, Zhao and colleagues performed RNA-seq on a child suffering from primary endocardial fibroelastosis and confirmed the role of a variant in ALMS1 gene (a known ciliopathy-associated gene) identified thanks to WES[225]. The authors referred to their approach as “integrative genomic analysis” in the sense that they used both WES and RNA-seq analysis to confirm the functional impact of the LoF variant.

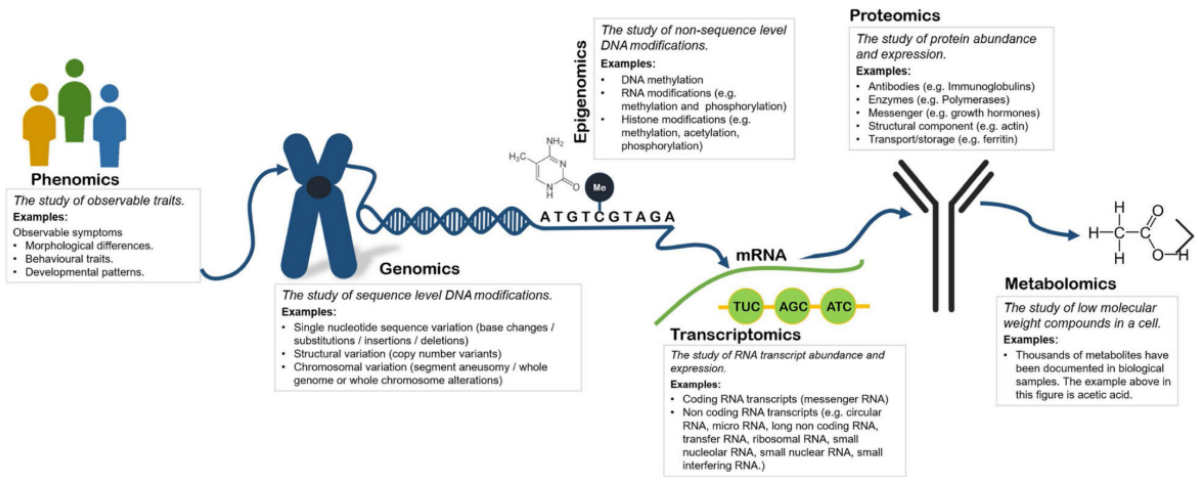



*Figure 28: Diagram representing a complete multi-omic data integration*
*Source: Kerr* et al. *(2020)*[226]

Kerr and colleagues' review on multi-omic approaches for rare diseases showed that only 11% of the articles considered were using multi-omics data integration specific software programs. Apart from phenotypic/genomic data integration methods that have been extensively discussed, more and more bioinformatics software programs developed to handle multi-omics data integration are created, such as PARADIGM which uses genomics data and biological pathways graphs[227] or Wang and colleagues' similarity network fusion method to integrate mRNA expression, DNA methylation and microRNA data for cancer[228]. There is a lot of effort and hope infused in graph theory and artificial intelligence to integrate data and eventually provide the community with new bioinformatics methods for disease prediction and diagnosis.

As highlighted in Kerr *et al.*'s review as well as in other studies discussed in this manuscript, multi-omics data integration can be beneficial in identifying markers for prognosis and help solving undiagnosed rare disease phenotypes[225,226]. Especially in the case of diseases with a high heterogeneity such as Bardet–Biedl syndrome, multi-omics integration could prove their usefulness to narrow down the analysis and eventually speed-up the diagnosis process. Developing robust workflows to handle multi-omics could be fruitful for rare diseases research and clinical care. However, this implies that clinicians can perform WES/WGS and transcriptomics on top of filling EHRs with HPO for example. As high throughput data is less and less expensive, the limiting factor might be time.

# CHAPTER 6. CONCLUSIONS

Considering the hypothesis and objectives initially formulated and in light of the experimental results presented, the conclusions of this thesis are:

I. The COBT method was developed, implementing a genomic processing pipeline and a Poisson-based per-gene burden test for case-only cohorts using the observed aggregated frequencies of rare genetic variants in a reference population.

II. When applied on a target cohort used as negative control, the method proved to be valid regarding their statistical assumptions as well as capable of controlling for genomic inflation and false positive rates on real genome sequencing data.

III. The application of the COBT approach on a clinically heterogeneous cohort of ciliopathy patients profiled through ciliary exome-targeted sequencing demonstrated its capacity to recapitulate already known gene-phenotype associations in a statistically significant manner as well as proposing novel candidate genetic variants that require further genetic, experimental, and clinical validation.

IV. A computational pipeline was implemented to evaluate clinical similarity among rare disease patients based on the Human Phenotype Ontology and making use of Resnik, Lin, Jiang-Conrath, Information Coefficient and ERIC semantic similarities.

V. When applied to a cohort of sporadic developmental disorder patients profiled through whole exome sequencing in trios and phenotyped using the Human Phenotype Ontology, patients with a high clinical similarity presented *de novo* protein-altering variants on the same genes or on the functionally-related gene set at rates significantly higher than what would be expected from random sampling.

VI. Further investigations are needed to systematically apply a phenotypically-guided variant prioritisation using only phenotype-to-phenotype proximity.

# CHAPTER 7. BIBLIOGRAPHY


1. Discovery of DNA Double Helix: Watson and Crick | Learn Science at Scitable. https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/.

2. Lander, E. S. & Schork, N. J. Genetic Dissection of Complex Traits. *Science (1979)* **265**, 2037–2048 (1994).

3. Teare, M. D. & Koref, M. F. S. Linkage analysis and the study of Mendelian disease in the era of whole exome and genome sequencing. *Brief Funct Genomics* **13**, 378–383 (2014).

4. Rahman, N. *et al.* Ehlers-Danlos syndrome with severe early-onset periodontal disease (EDS-VIII) is a distinct, heterogeneous disorder with one predisposition gene at chromosome 12p13. *Am J Hum Genet* **73**, 198–204 (2003).

5. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467 (1977).

6. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).

7. Luckey, J. A. *et al.* High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res* **18**, 4417 (1990).

8. Khan, A. R., Pervez, M. T., Babar, M. E., Naveed, N. & Shoaib, M. A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. *Evol Bioinform Online* **14**, (2018).

9. Lander, E. S. *et al.* Correction: Initial sequencing and analysis of the human genome. *Nature 2001 412:6846* **412**, 565–566 (2001).

10. Craig Venter, J. *et al.* The sequence of the human genome. *Science (1979)* **291**, 1304–1351 (2001).

11. Abdellah, Z. *et al.* Finishing the euchromatic sequence of the human genome. *Nature 2004 431:7011* **431**, 931–945 (2004).

12. Collins, F. S., Morgan, M. & Patrinos, A. The Human Genome Project: Lessons from Large-Scale Biology. *Science (1979)* **300**, 286–290 (2003).

13. Homo sapiens genome assembly GRCh38 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000001405.26/.

14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754 (2009).

15. Moorthie, S., Mattocks, C. J. & Wright, C. F. Review of massively parallel DNA sequencing technologies. *HUGO Journal* **5**, 1–12 (2011).

16. European Commission. *Useful Information on Rare Diseases from an EU Perspective*. www.orpha.net.

17. Simone Baldovino, Antoni Montserrat Moliner, Domenica Taruscio, Erica Daina, D. R. Rare Diseases in Europe: from a Wide to a Local Perspective. *The Israel Medicine Association Journal* **18**, (2016).

18. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Res* **47**, D1038–D1043 (2019).

19. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**, D789–D798 (2015).

20. Rath, A. *et al.* Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Hum Mutat* **33**, 803–808 (2012).

21. Reiter, J. F. & Leroux, M. R. Genes and molecular pathways underpinning ciliopathies. *Nature Reviews Molecular Cell Biology 2017 18:9* **18**, 533–547 (2017).

22. Cornillie, F. J., Lauweryns, J. M. & Corbeel, L. Atypical bronchial cilia in children with recurrent respiratory tract infections. A comparative ultrastructural study. *Pathol Res Pract* **178**, 595–604 (1984).

23. Schmidts, M. *et al.* Exome sequencing identifies DYNC2H1 mutations as a common cause of asphyxiating thoracic dystrophy (Jeune syndrome) without major polydactyly, renal or retinal involvement. *J Med Genet* **50**, 309–323 (2013).

24. Wallmeier, J. *et al.* Motile ciliopathies. *Nature Reviews Disease Primers 2020 6:1* **6**, 1–29 (2020).

25. Gleeson, J. G. *et al.* Molar tooth sign of the midbrain-hindbrain junction: occurrence in multiple distinct syndromes. *Am J Med Genet A* **125A**, 125–134 (2004).

26. Orphanet: Nephronophthisis. https://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=en&Expert=655.

27. van Dam, T. J. P. *et al.* CiliaCarta: An integrated and validated compendium of ciliary genes. *PLoS One* **14**, e0216705 (2019).

28. Arnaiz, O., Cohen, J., Tassin, A.-M. & Koll, F. Remodeling Cildb, a popular database for cilia and links for ciliopathies. *Cilia 2014 3:1* **3**, 1–10 (2014).

29. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet* **17**, 502–510 (2001).

30. Kimura, M. Evolutionary Rate at the Molecular Level. *Nature 1968 217:5129* **217**, 624–626 (1968).

31. Iyengar, S. K. & Elston, R. C. The genetic basis of complex traits: rare variants or 'common gene, common disease'? *Methods Mol Biol* **376**, 71–84 (2007).

32. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am J Hum Genet* **90**, 7–24 (2012).

33. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature 2009 461:7265* **461**, 747–753 (2009).

34. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* **11**, 773–785 (2010).

35. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *New England Journal of Medicine* **372**, 2235–2242 (2015).

36. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062 (2018).

37. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature 2015 526:7571* **526**, 75–81 (2015).

38. Clarke, L. *et al.* The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Res* **45**, D854–D859 (2017).

39. Auer, P. L. *et al.* Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. *The American Journal of Human Genetics* **99**, 791–801 (2016).

40. Dorschner, M. O. *et al.* Actionable, Pathogenic Incidental Findings in 1,000 Participants' Exomes. *The American Journal of Human Genetics* **93**, 631–640 (2013).

41. Amendola, L. M. *et al.* Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Res* **25**, 305–315 (2015).

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

43. Karczewski, K. J. *et al.* The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* **45**, D840 (2017).

44. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature 2020 581:7809* **581**, 434–443 (2020).

45. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biology 2016 17:1* **17**, 1–19 (2016).

46. Auton, A. *et al.* A global reference for human genetic variation. *Nature 2015 526:7571* **526**, 68–74 (2015).

47. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).

48. Study, D. D. D. *et al.* Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).

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

50. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature Genetics 2011 43:7* **43**, 712–714 (2011).

51. Michaelson, J. J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431–1442 (2012).

52. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics 2010 42:9* **42**, 790–793 (2010).

53. Hoischen, A. *et al.* De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nature Genetics 2011 43:8* **43**, 729–731 (2011).

54. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nature Genetics 2010 42:6* **42**, 483–485 (2010).

55. Posey, J. E. *et al.* Molecular diagnostic experience of whole-exome sequencing in adult patients. *Genet Med* **18**, 678–685 (2016).

56. Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**, 1870–1879 (2014).

57. Mueller, W. F., Larsen, L. S. Z., Garibaldi, A., Hatfield, G. W. & Hertel, K. J. The Silent Sway of Splicing by Synonymous Substitutions. *J Biol Chem* **290**, 27700–27711 (2015).

58. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877 (2016).

59. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886–D894 (2019).

60. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102–R110 (2015).

61. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 1–14 (2016).

62. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164–e164 (2010).

63. Cingolani, P. Variant Annotation and Functional Prediction: SnpEff. 289–314 (2022) doi:10.1007/978-1-0716-2293-3_19.

64. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760–1774 (2012).

65. Rodriguez, J. M. *et al.* APPRIS 2017: Principal isoforms for multiple gene sets. *Nucleic Acids Res* **46**, D213–D217 (2018).

66. Miko, I. Phenotype Variability: Penetrance and Expressivity. *Nature Education* **1**, 137 (2008).

67. The BRCA1 and BRCA2 Genes | CDC. https://www.cdc.gov/genomics/disease/breast_ovarian_cancer/genes_hboc.htm.

68. Chen, S. & Parmigiani, G. Meta-Analysis of BRCA1 and BRCA2 Penetrance. *J Clin Oncol* **25**, 1329 (2007).

69. Kim, S., Morris, N. J., Won, S. & Elston, R. C. Single-Marker and Two-Marker Association Tests for Unphased Case-Control Genotype Data, with a Power Comparison. *Genet Epidemiol* **34**, 67 (2010).

70. Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. *Annu Rev Genet* **44**, 293–308 (2010).

71. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311–321 (2008).

72. Hong, E. P. & Park, J. W. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics Inform* **10**, 117–122 (2012).

73. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95**, 5–23 (2014).

74. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature Reviews Genetics* vol. 20 747–759 Preprint at https://doi.org/10.1038/s41576-019-0177-4 (2019).

75. Guo, M. H. H. *et al.* Determinants of Power in Gene-Based Burden Testing for Monogenic Disorders. *Am J Hum Genet* **99**, 527–539 (2016).

76. Derkach, A. *et al.* Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. *Bioinformatics* **30**, 2179 (2014).

77. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nature Medicine 2022 28:2* **28**, 243–250 (2022).

78. Petrovski, S. & Goldstein, D. B. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol* **17**, 1–3 (2016).

79. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–909 (2006).

80. Hu, Y.-J. *et al.* Testing Rare-Variant Association without Calling Genotypes Allows for Systematic Differences in Sequencing between Cases and Controls. *PLoS Genet* **12**, e1006040 (2016).

81. Guo, M. H., Plummer, L., Chan, Y. M., Hirschhorn, J. N. & Lippincott, M. F. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *Am J Hum Genet* **103**, 522–534 (2018).

82. Chen, W. *et al.* A rare variant analysis framework using public genotype summary counts to prioritize disease-predisposition genes. *Nature Communications 2022 13:1* **13**, 1–18 (2022).

83. Cirulli, E. T. *et al.* Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science* **347**, 1436–1441 (2015).

84. Raghavan, N. S. *et al.* Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer's disease. *Ann Clin Transl Neurol* **5**, 832 (2018).

85. Lapidus, A. L. Genome Sequence Databases: Sequencing and Assembly. *Encyclopedia of Microbiology* 196–210 (2009) doi:10.1016/B978-012373944-5.00028-6.

86. Panoutsopoulou, K. & Walter, K. Quality control of common and rare variants. *Methods in Molecular Biology* **1793**, 25–36 (2018).

87. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901 (2005).

88. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034–1050 (2005).

89. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110–121 (2010).

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

91. Sim, N. L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* **40**, W452 (2012).

92. Davydov, E. v. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, (2010).

93. Ganna, A. *et al.* Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. *Am J Hum Genet* **102**, 1204–1211 (2018).

94. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* **615**, 28–56 (2007).

95. Madsen, B. E. & Browning, S. R. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet* **5**, e1000384 (2009).

96. Curtis, D. A weighted burden test using logistic regression for integrated analysis of sequence variants, copy number variants and polygenic risk score. *European Journal of Human Genetics* **27**, 114 (2019).

97. Han, F. & Pan, W. A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants. *Hum Hered* **70**, 42 (2010).

98. Hoffmann, T. J., Marini, N. J. & Witte, J. S. Comprehensive Approach to Analyzing Rare Genetic Variants. *PLoS One* **5**, e13584 (2010).

99. Lin, D. Y. & Tang, Z. Z. A General Framework for Detecting Disease Associations with Rare Variants in Sequencing Studies. *Am J Hum Genet* **89**, 354 (2011).

100. Liu, D. J. & Leal, S. M. 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 Genet* **6**, e1001156 (2010).

101. Neale, B. M. *et al.* Testing for an Unusual Distribution of Rare Variants. *PLoS Genet* **7**, e1001322 (2011).

102. Pan, W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol* **33**, 497–507 (2009).

103. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am J Hum Genet* **89**, 82 (2011).

104. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).

105. Hansen, M. C., Haferlach, T. & Nyvold, C. G. A decade with whole exome sequencing in haematology. *Br J Haematol* **188**, 367–382 (2020).

106. Krawczak, M., Ball, E. v. & Cooper, D. N. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am J Hum Genet* **63**, 474–488 (1998).

107. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nature Genetics 2009 41:4* **41**, 393–395 (2009).

108. Vigeland, M. D., Gjøtterud, K. S. & Selmer, K. K. FILTUS: a desktop GUI for fast and efficient detection of disease-causing variants, including a novel autozygosity detector. *Bioinformatics* **32**, 1592–1594 (2016).

109. Zhi, D. & Chen, R. Statistical guidance for experimental design and data analysis of mutation detection in rare monogenic Mendelian diseases by exome sequencing. *PLoS One* **7**, 31358 (2012).

110. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Res* **47**, D1038–D1043 (2019).

111. Robinson, P. N. *et al.* The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *The American Journal of Human Genetics* **83**, 610–615 (2008).

112. Ashburner, M. *et al.* Creating the gene ontology resource: design and implementation. *Genome Res* **11**, 1425–1433 (2001).

113. Human Phenotype Ontology. https://hpo.jax.org/app/.

114. Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. (1995).

115. Lin, D. An Information-Theoretic Definition of Similarity. Preprint at (1998).

116. Jiang, J. J. & Conrath, D. W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of the 10th Research on Computational Linguistics International Conference, ROCLING 1997* 19–33 (1997) doi:10.48550/arxiv.cmp-lg/9709008.

117. Li, B., Wang, J. Z., Feltus, F. A., Zhou, J. & Luo, F. Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins. (2010).

118. Li, Q., Zhao, K., Bustamante, C. D., Ma, X. & Wong, W. H. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Genetics in Medicine* **21**, 2126–2134 (2019).

119. Pesquita, C. *et al.* Metrics for GO based protein semantic similarity: A systematic evaluation. *BMC Bioinformatics* **9**, 1–16 (2008).

120. Li, J. *et al.* A Comprehensive Evaluation of Disease Phenotype Networks for Gene Prioritization. *PLoS One* **11**, e0159457 (2016).

121. Gong, X., Jiang, J., Duan, Z. & Lu, H. A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology. *BMC Bioinformatics* **19**, (2018).

122. Kulmanov, M. & Hoehndorf, R. Evaluating the effect of annotation size on measures of semantic similarity. *J Biomed Semantics* **8**, (2017).

123. Köhler, S. *et al.* Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet* **85**, 457–64 (2009).

124. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

125. Rae, W. *et al.* Clinical efficacy of a next-generation sequencing gene panel for primary immunodeficiency diagnostics. *Clin Genet* **93**, 647–655 (2018).

126. Sifrim, A. *et al.* eXtasy: variant prioritization by genomic data fusion. *Nat Methods* **10**, 1083–1084 (2013).

127. Robinson, P. N. *et al.* Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res* **24**, 340–8 (2014).

128. Zemojtel, T. *et al.* Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med* **6**, 252ra123 (2014).

129. Ayadi, A. *et al.* Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mamm Genome* **23**, 600–610 (2012).

130. Schriml, L. M. *et al.* Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res* **40**, (2012).

131. Singleton, M. V. *et al.* Phevor Combines Multiple Biomedical Ontologies for Accurate Identification of Disease-Causing Alleles in Single Individuals and Small Nuclear Families. *The American Journal of Human Genetics* **94**, 599–610 (2014).

132. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* doi:10.1145/2939672.

133. Yoo, B., Birgmeier, J., Bernstein, J. A. & Bejerano, G. InpherNet accelerates monogenic disease diagnosis using patients' candidate genes' neighbors. *Genetics in Medicine 2021* 1–9 (2021) doi:10.1038/s41436-021-01238-2.

134. Bainbridge, M. N. *et al.* Whole-genome sequencing for optimized patient management. *Sci Transl Med* **3**, 87re3-87re3 (2011).

135. Ensembl 2014 | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/42/D1/D749/1059722.

136. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4204768/.

137. Failler, M. *et al.* Mutations of CEP83 Cause Infantile Nephronophthisis and Intellectual Disability. *Am J Hum Genet* **94**, 905 (2014).

138. Thomas, S. *et al.* TCTN3 Mutations Cause Mohr-Majewski Syndrome. *The American Journal of Human Genetics* **91**, 372–378 (2012).

139. Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* **2011**, (2011).

140. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports 2019 9:1* **9**, 1–5 (2019).

141. Fuentes Fajardo, K. v. *et al.* Detecting false-positive signals in exome sequencing. *Hum Mutat* **33**, 609–613 (2012).

142. The Hail Team. GitHub - hail-is/hail: Scalable genomic data analysis. Preprint at https://github.com/hail-is/hail (2008).

143. Rao, C. R. & Chakravarti, I. M. Some Small Sample Tests of Significance for a Poisson Distribution. *Biometrics* **12**, 264 (1956).

144. Fernández-Fontelo, A., Puig, P., Ainsbury, E. A. & Higueras, M. An exact goodness-of-fit test based on the occupancy problems to study zero-inflation and zero-deflation in biological dosimetry data. *Radiat Prot Dosimetry* **179**, 317–326 (2018).

145. Fisher, R. A. The Significance of Deviations from Expectation in a Poisson Series. *Biometrics* **6**, 17 (1950).

146. Higueras, M., González, J. E., di Giorgio, M. & Barquinero, J. F. A note on Poisson goodness-of-fit tests for ionizing radiation induced chromosomal aberration samples. *https://doi.org/10.1080/09553002.2018.1478012* **94**, 656–663 (2018).

147. Devlin, B., Roeder, K. & Bacanu, S. A. Unbiased methods for population-based association studies. *Genet Epidemiol* **21**, 273–284 (2001).

148. Stenson, P. D. *et al.* The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum Genet* **139**, 1197 (2020).

149. Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**, 1–9 (2014).

150. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics 2014 46:3* **46**, 310–315 (2014).

151. Breheny, P., Stromberg, A. & Lambert, J. P-Value histograms: Inference and diagnostics. *High Throughput* **7**, 23 (2018).

152. Döhler, S., Durand, G. & Roquain, E. New FDR bounds for discrete and heterogeneous tests. *Electron J Stat* **12**, 1867–1900 (2018).

153. Devlin, B., Bacanu, S. A. & Roeder, K. Genomic Control to the extreme. *Nature Genetics 2004 36:11* **36**, 1129–1130 (2004).

154. Richards, 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* **17**, 405 (2015).

155. Hendricks, A. E. *et al.* ProxECAT: Proxy External Controls Association Test. A new case-control gene region association test using allele frequencies from public controls. *PLoS Genet* **14**, e1007591 (2018).

156. Bocher, O. & Génin, E. Rare variant association testing in the non-coding genome. *Human Genetics 2020 139:11* **139**, 1345–1362 (2020).

157. Greene, D., Richardson, S. & Turro, E. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics* btw763 (2017) doi:10.1093/bioinformatics/btw763.

158. Köhler, S. *et al.* Encoding Clinical Data with the Human Phenotype Ontology for Computational Differential Diagnostics. *Curr Protoc Hum Genet* **103**, (2019).

159. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481–D487 (2016).

160. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).

161. Rojano, E. *et al.* Evaluating, Filtering and Clustering Genetic Disease Cohorts Based on Human Phenotype Ontology Data with Cohort Analyzer. *Journal of Personalized Medicine 2021, Vol. 11, Page 730* **11**, 730 (2021).

162. Son, J. H. *et al.* Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes. *Am J Hum Genet* **103**, 58–73 (2018).

163. Clark, M. M. *et al.* Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci Transl Med* **11**, (2019).

164. Peng, J., Hui, W. & Shang, X. Measuring phenotype-phenotype similarity through the interactome. *BMC Bioinformatics* **19**, 114 (2018).

165. Bocher, O. I. *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 Genet* **18**, e1009923 (2022).

166. Marenne, G. *et al.* RAVAQ: An integrative pipeline from quality control to region-based rare variant association analysis. *Genet Epidemiol* **46**, 256–265 (2022).

167. Dutta, D., Scott, L., Boehnke, M. & Lee, S. Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. *Genet Epidemiol* **43**, 4–23 (2019).

168. Garcelon, N. *et al.* Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. *J Biomed Inform* **73**, 51–61 (2017).

169. Katsanis, N. The oligogenic properties of Bardet-Biedl syndrome. *Hum Mol Genet* **13 Spec No 1**, (2004).

170. Perea-Romero, I. *et al.* Allelic overload and its clinical modifier effect in Bardet-Biedl syndrome. *npj Genomic Medicine 2022 7:1* **7**, 1–7 (2022).

171. Parmley, J. L., Chamary, J. v. & Hurst, L. D. Evidence for Purifying Selection Against Synonymous Mutations in Mammalian Exonic Splicing Enhancers. *Mol Biol Evol* **23**, 301–309 (2006).

172. Pagani, F., Raponi, M. & Baralle, F. E. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proceedings of the National Academy of Sciences* **102**, 6368–6372 (2005).

173. Olinger, E. *et al.* A discarded synonymous variant in NPHP3 explains nephronophthisis and congenital hepatic fibrosis in several families. *Hum Mutat* **42**, 1221–1228 (2021).

174. Richardson, T. G. *et al.* A protein domain and family based approach to rare variant association analysis. *PLoS One* **11**, e0153803 (2016).

175. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

176. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nature Reviews Genetics* vol. 17 661–678 Preprint at https://doi.org/10.1038/nrg.2016.112 (2016).

177. Dale, R. *et al.* Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nat Methods* **15**, 475–476 (2018).

178. Zou, J. *et al.* A primer on deep learning in genomics. *Nat Genet* **51**, 12–18 (2019).

179. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nature Human Behaviour 2017 1:1* **1**, 1–9 (2017).

180. Phillips, P., Lithgow, G. J. & Driscoll, M. A long journey toward reproducible results. *Nature* **548**, 387 (2017).

181. Baker, M. & Penny, D. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).

182. Wratten, L., Wilm, A. & Göke, J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods 2021 18:10* **18**, 1161–1168 (2021).

183. Kulkarni, N. *et al.* Reproducible bioinformatics project: A community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics* **19**, 5–13 (2018).

184. Macleod, M. R. *et al.* Biomedical research: Increasing value, reducing waste. *The Lancet* **383**, 101–104 (2014).

185. Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. Ten Simple Rules for Reproducible Computational Research. *PLoS Comput Biol* **9**, e1003285 (2013).

186. Gentleman, R. C. *et al.* Open Access Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).

187. Digan, W. *et al.* An architecture for genomics analysis in a clinical setting using Galaxy and Docker. *Gigascience* **6**, (2017).

188. Köster, J. *et al.* Sustainable data analysis with Snakemake. *F1000Res* **10**, (2021).

189. di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature Biotechnology* vol. 35 316–319 Preprint at https://doi.org/10.1038/nbt.3820 (2017).

190. Su, S., Zhang, L. & Liu, J. An Effective Method to Measure Disease Similarity Using Gene and Phenotype Associations. *Front Genet* **10**, 466 (2019).

191. Pontikos, N. *et al.* Phenogenon: Gene to phenotype associations for rare genetic diseases. *PLoS One* **15**, (2020).

192. Deisseroth, C. A. *et al.* ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genetics in Medicine* **21**, 1585–1593 (2019).

193. Boudellioua, I., Kulmanov, M., Schofield, P. N., Gkoutos, G. v. & Hoehndorf, R. DeepPVP: Phenotype-based prioritization of causative variants using deep learning. *BMC Bioinformatics* **20**, (2019).

194. Chen, J. *et al.* Novel phenotype–disease matching tool for rare genetic diseases. *Genetics in Medicine* **21**, 339–346 (2018).

195. Rao, A. *et al.* Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks. *BMC Med Genomics* **11**, 57 (2018).

196. Javed, A., Agrawal, S. & Ng, P. C. Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat Methods* **11**, 935–937 (2014).

197. Zhao, M. *et al.* Phen2Gene: Rapid Phenotype-Driven Gene Prioritization for Rare Diseases. *bioRxiv* 870527 (2019) doi:10.1101/870527.

198. Pengelly, R. J. *et al.* Evaluating phenotype-driven approaches for genetic diagnoses from exomes in a clinical setting. *Sci Rep* **7**, (2017).

199. Jacobsen, J. O. B. *et al.* Phenotype-driven approaches to enhance variant prioritization and diagnosis of rare disease. *Hum Mutat* (2022) doi:10.1002/HUMU.24380.

200. Turnbull, C. *et al.* The 100 000 Genomes Project: Bringing whole genome sequencing to the NHS. *BMJ (Online)* **361**, (2018).

201. Robinson, P. N. *et al.* Interpretable Clinical Genomics with a Likelihood Ratio Paradigm. *Am J Hum Genet* **107**, 403–417 (2020).

202. Smedley, D. *et al.* A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *Am J Hum Genet* **99**, 595 (2016).

203. Geoffroy, V. *et al.* AnnotSV and knotAnnotSV: A web server for human structural variations annotations, ranking and analysis. *Nucleic Acids Res* **49**, W21–W28 (2021).

204. Danis, D. *et al.* Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *Am J Hum Genet* **108**, 1564–1577 (2021).

205. gnomAD Production team. gnomAD v3.1 | gnomAD browser. https://gnomad.broadinstitute.org/news/2020-10-gnomad-v3-1/ (2020).

206. illumina, I. Press Release. https://www.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=8d04df3f-d9c1-4c85-8177-6ea604627ccd (2022).

207. Bhartiya, D., Jalali, S., Ghosh, S. & Scaria, V. Distinct patterns of genetic variations in potential functional elements in long noncoding RNAs. *Hum Mutat* **35**, 192–201 (2014).

208. Duan, J. *et al.* A rare functional noncoding variant at the GWAS-Implicated MIR137/MIR2682 locus might confer risk to schizophrenia and bipolar disorder. *Am J Hum Genet* **95**, 744–753 (2014).

209. Kapoor, A. *et al.* An enhancer polymorphism at the cardiomyocyte intercalated disc protein NOS1AP locus is a major regulator of the QT interval. *Am J Hum Genet* **94**, 854–869 (2014).

210. Zhang, X. *et al.* Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat Genet* **47**, 345–352 (2015).

211. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294–296 (2014).

212. Ferlaino, M. *et al.* An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome. *BMC Bioinformatics* **18**, 1–8 (2017).

213. 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 Biol* **20**, 32 (2019).

214. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics* vol. 20 389–403 Preprint at https://doi.org/10.1038/s41576-019-0122-6 (2019).

215. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol Syst Biol* **12**, 878 (2016).

216. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).

217. Vitsios, D., Dhindsa, R. S., Middleton, L., Gussow, A. B. & Petrovski, S. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. *Nat Commun* **12**, (2021).

218. Bergstrom, C. T. & West, J. D. (Jevin D. Calling bullshit : the art of skepticism in a data-driven world. 318.

219. Wu, X. & Zhang, X. Automated Inference on Criminality using Face Images. (2016) doi:10.48550/arxiv.1611.04135.

220. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res* **17**, 1520–1528 (2007).

221. Herzig, A. F., Clerget-Darpoux, F. & Génin, E. The False Dawn of Polygenic Risk Scores for Human Disease Prediction. *Journal of Personalized Medicine 2022, Vol. 12, Page 1266* **12**, 1266 (2022).

222. van Vleck, L. D. *et al.* Estimated breeding values for meat characteristics of crossbred cattle with an animal model. *J Anim Sci* **70**, 363–371 (1992).

223. Génin, E. & Clerget-Darpoux, F. Revisiting the Polygenic Additive Liability Model through the Example of Diabetes Mellitus. *Human Heredity* vol. 80 171–177 Preprint at https://doi.org/10.1159/000447683 (2016).

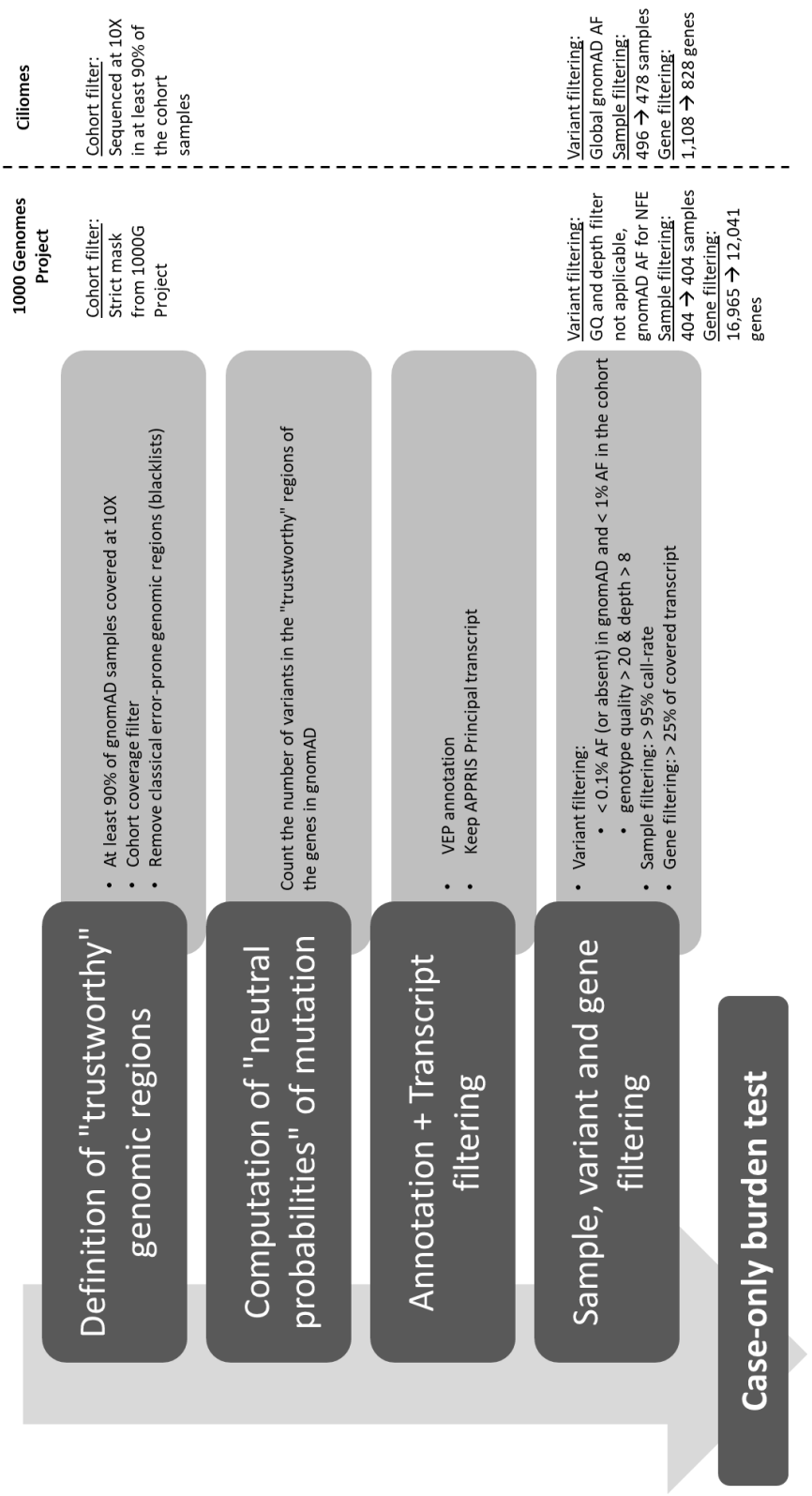
224. Ding, Y. *et al.* Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. *Nat Genet* **54**, 30–39 (2022).

225. Zhao, Y. *et al.* Recessive ciliopathy mutations in primary endocardial fibroelastosis: a rare neonatal cardiomyopathy in a case of Alstrom syndrome. *J Mol Med (Berl)* **99**, 1623–1638 (2021).

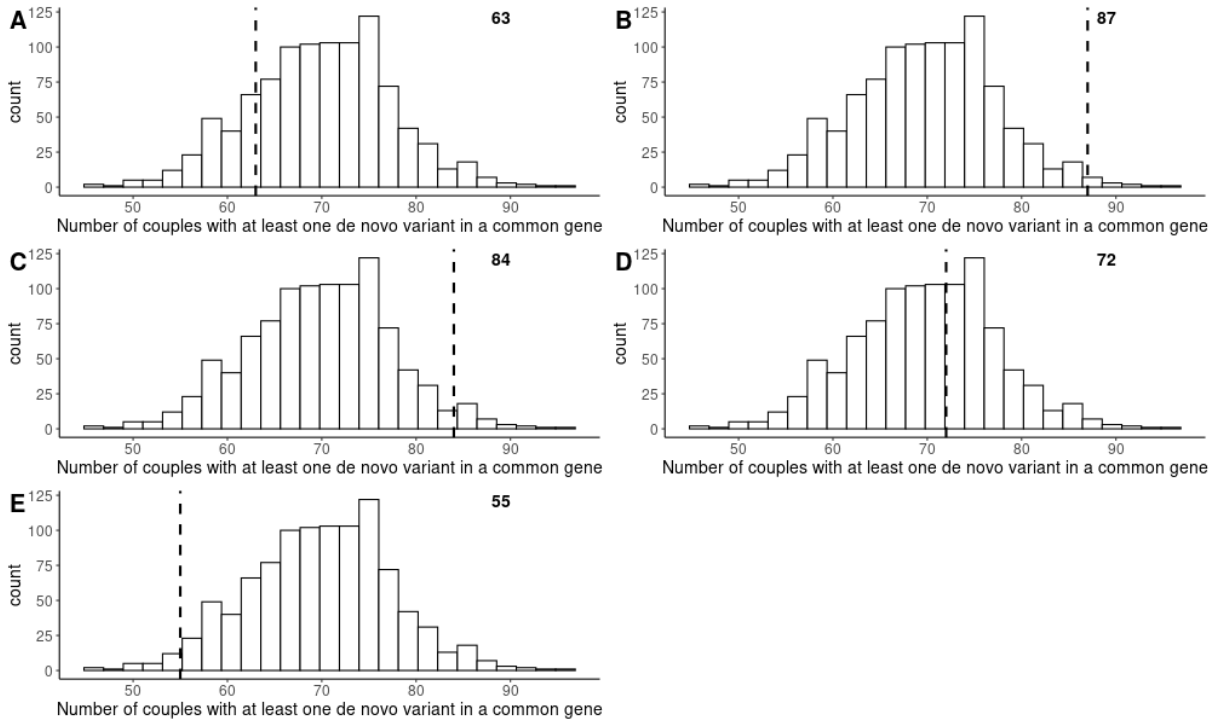
226. Kerr, K. *et al.* A scoping review and proposed workflow for multi-omic rare disease research. *Orphanet J Rare Dis* **15**, (2020).

227. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, (2010).

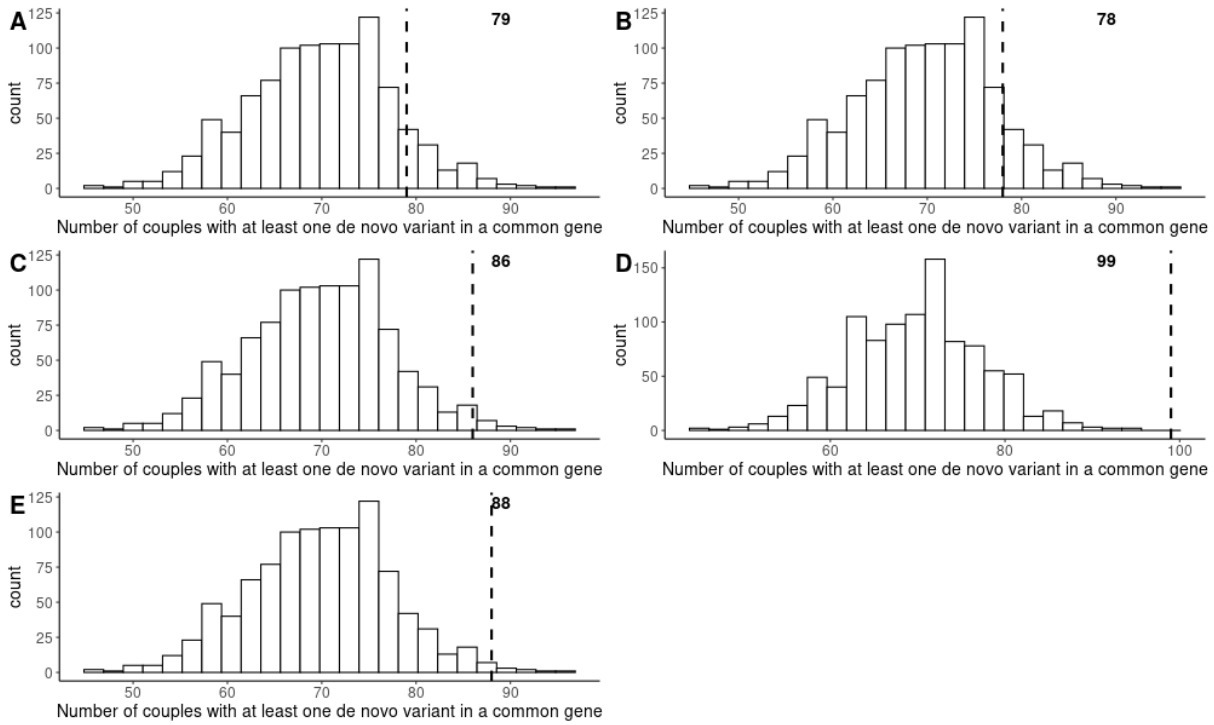
228. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* **11**, 333–337 (2014).



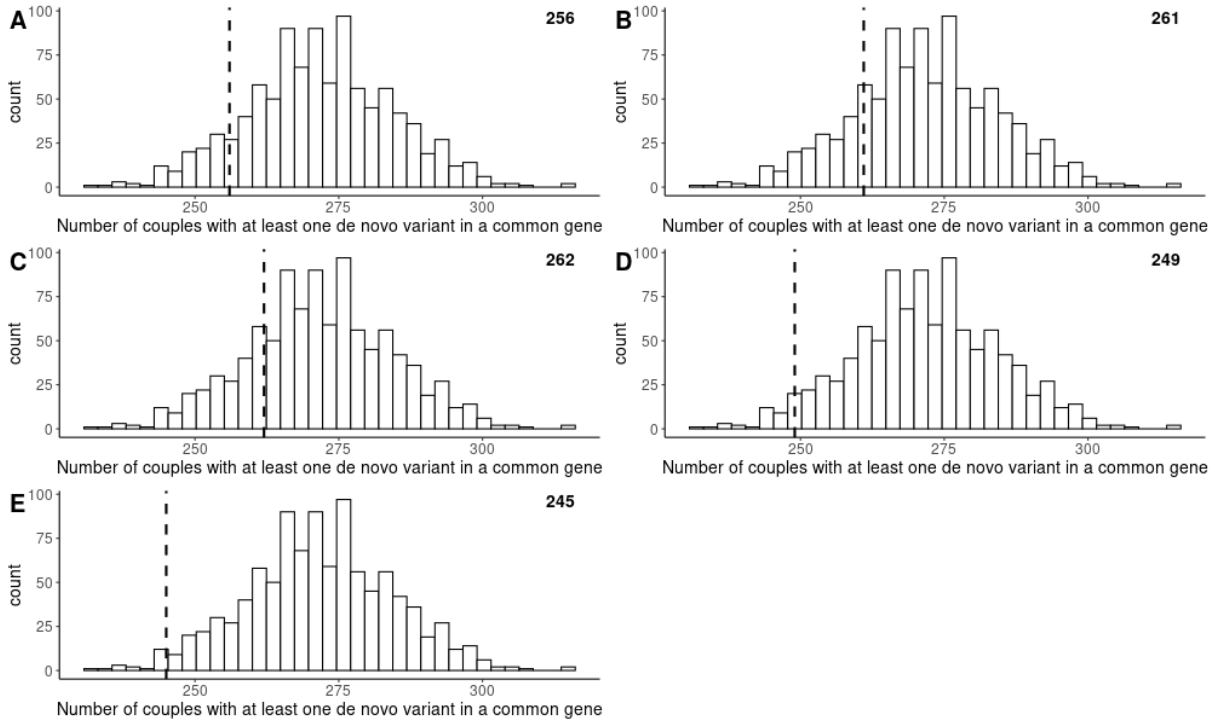

*Supplementary figure 1: Schematic representation of the pipeline for COBT*



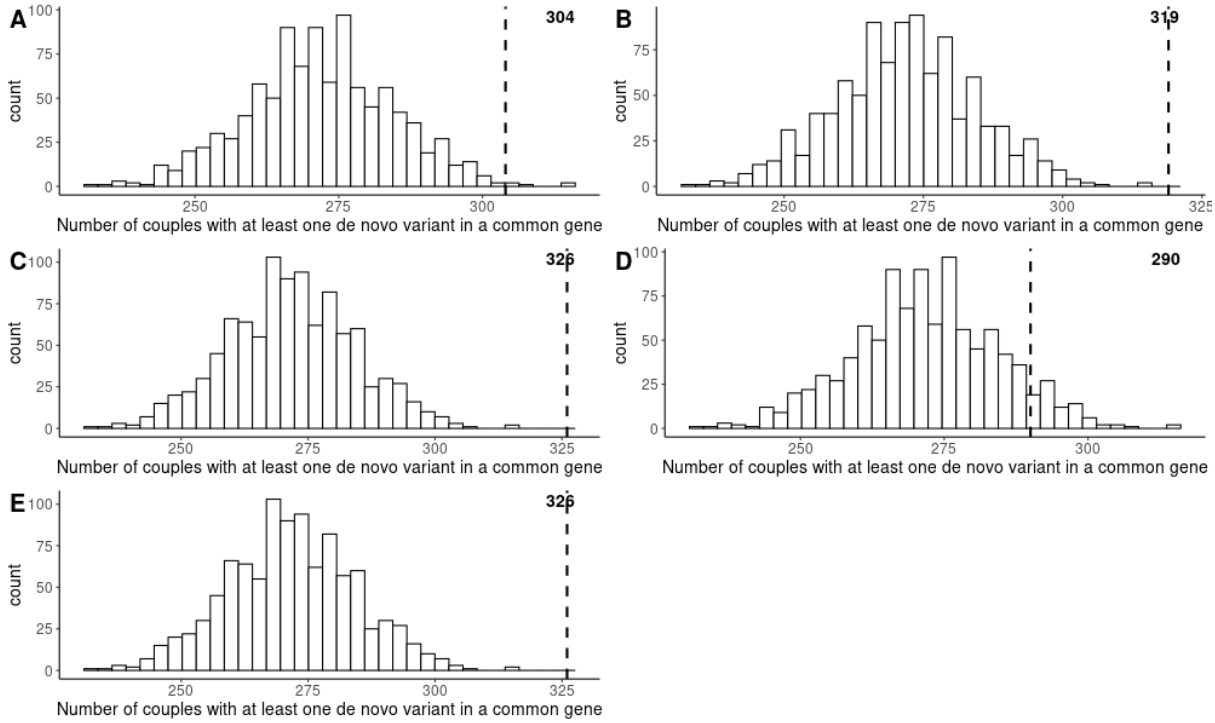

*Supplementary figure 2: Distribution of the number of couples with at least one* de novo *variant in a common Hallmark pathway*
*Distribution of the number of couples with at least one* de novo *variant in a common Hallmark pathway in 1,000 randomly selected couples of patients and number of phenotypically closest couples with at least one de novo variant in a common pathway (dotted bar) computed with the 5 semantic similarity measures: Resnik (A); Lin (B); Information Coefficient (C); Jiang-Conrath (D); ERIC (E)*




*Supplementary figure 3: Distribution of the number of couples with at least one* de novo *variant in a common KEGG pathway*
*Distribution of the number of couples with at least one* de novo *variant in a common KEGG pathway in 1,000 randomly selected couples of patients and number of phenotypically closest couples with at least one de novo variant in a common pathway (dotted bar) computed with the 5 semantic similarity measures: Resnik (A); Lin (B); Information Coefficient (C); Jiang-Conrath (D); ERIC (E)*



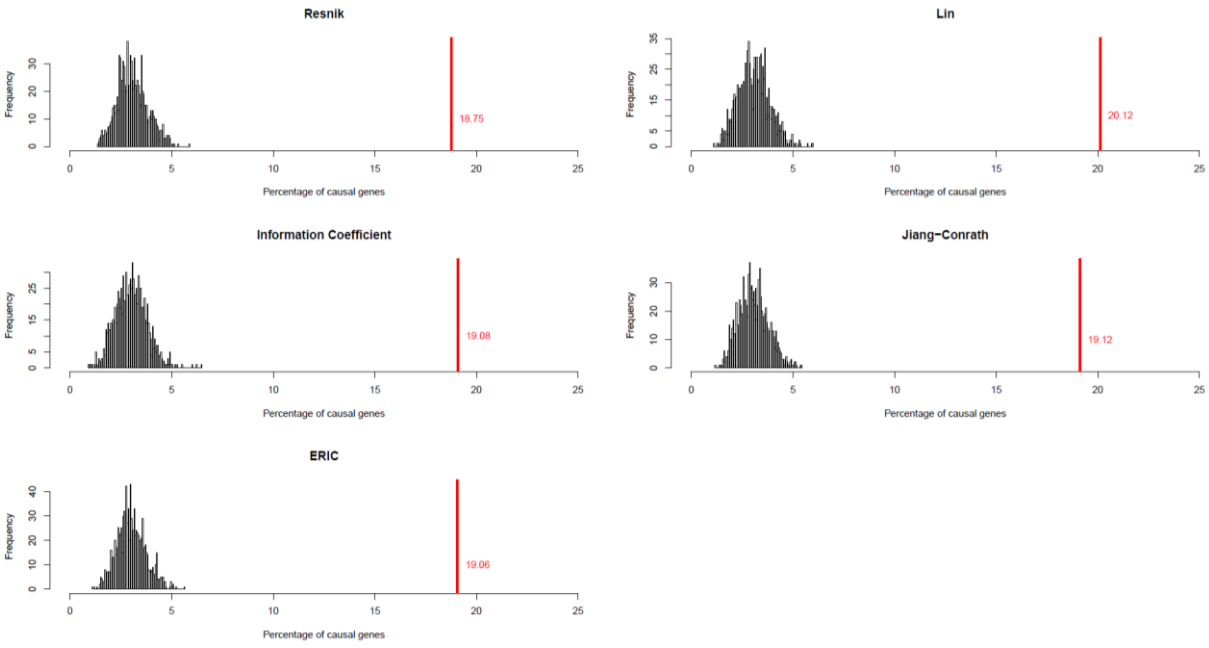

*Supplementary figure 4: Distribution of the number of couples with at least one* de novo *variant in a common Hallmark pathway for the top 5 closest couples*
*Distribution of the number of couples with at least one* de novo *variant in a common Hallmark pathway in 1,000 randomly selected sets of 5 couples of patients and number of top 5 phenotypically closest couples with at least one de novo variant in a common pathway (dotted bar) computed with the 5 semantic similarity measures: Resnik (A); Lin (B); Information Coefficient (C); Jiang-Conrath (D); ERIC (E)*



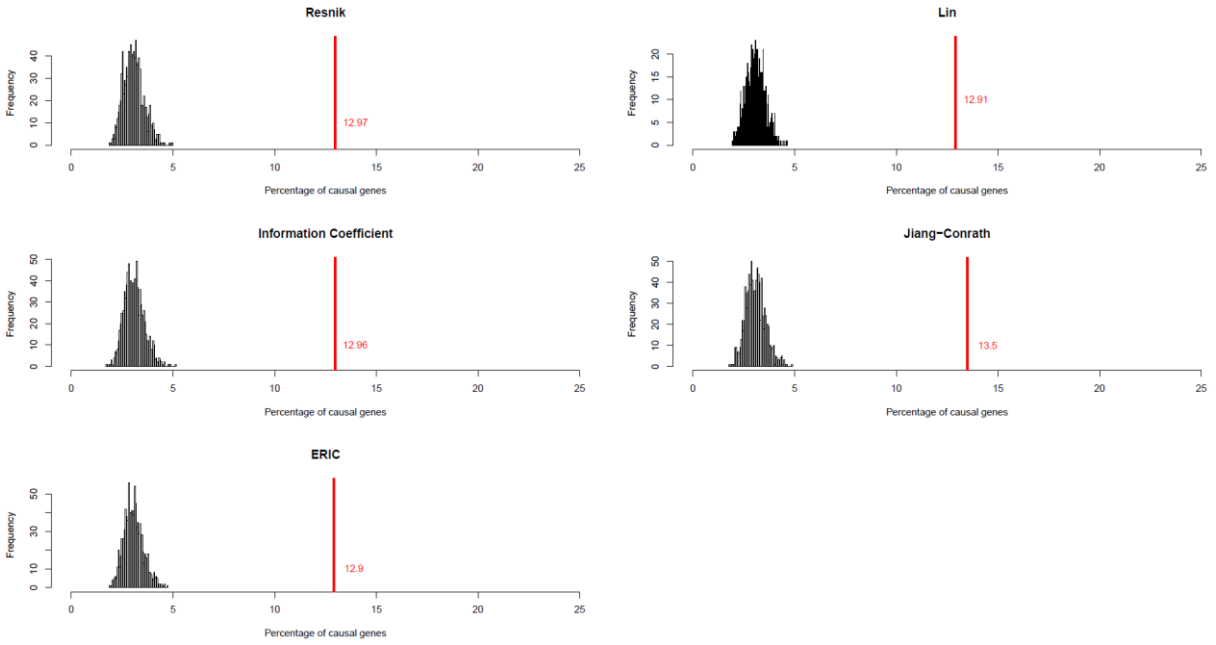

*Supplementary figure 5: Distribution of the number of couples with at least one* de novo *variant in a common KEGG pathway for the top 5 closest couples*
*Distribution of the number of couples with at least one* de novo *variant in a common KEGG pathway in 1,000 randomly selected sets of 5 couples of patients and number of top 5 phenotypically closest couples with at least one de novo variant in a common pathway (dotted bar) computed with the 5 semantic similarity measures: Resnik (A); Lin (B); Information Coefficient (C); Jiang-Conrath (D); ERIC (E)*

*Supplementary information:*

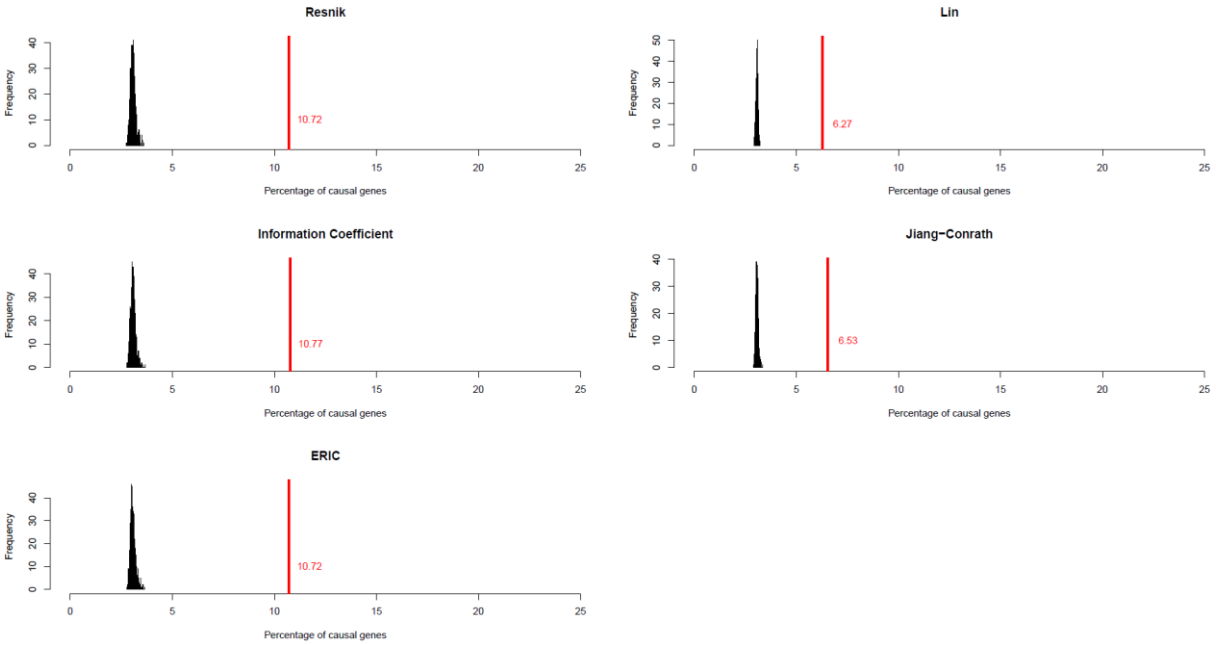
Before studying the DDD cohort, 632 HPO annotated ciliopathy patients from the Cil'lico cohort were used to assess the difference between BMA, BMS and maximum scores for the 5 semantic similarity measures. No HPO redundancy filtering was applied on this cohort. The following supplementary figures refer to this dataset. In order to measure the differences between the aggregation methods, the semantic similarity matrices were constructed as presented in 4.1.5 but instead of computing the number of couples with a common mutated gene in 1,000 random sampling as presented in 4.2.2.1, I randomly selected 1,000 times 632 pairs of patients and counted how many pairs had the same causal genes. I repeated the procedure for the 5 semantic similarity computation methods and the 3 aggregation methods. Finally, I compared those distributions to the pairs of actual pairs of closest patients. The aim was to assess the frequency with which couples of patients with maximum similarity are indeed genetically close (by sharing the same causal gene) as compared to random couples for the 5 semantic similarity metrics and the 3 aggregation methods. While no dramatic difference is observed between the semantic similarity metrics, huge differences can be seen between the three aggregation methods. BMA performs better in displaying genetic resemblance.




*Supplementary figure 6: Distribution of the percentage of pairs of patients sharing identical causal gene for 1,000 randomly selected couples of individuals for 5 semantic similarity computation methods compared to the couples with maximum semantic similarity (red bar) assessed with Best Match Average aggregation method*




*Supplementary figure 7: Distribution of the percentage of pairs of patients sharing identical causal gene for 1,000 randomly selected couples of individuals for 5 semantic similarity computation methods compared to the couples with maximum semantic similarity (red bar) assessed with Best Match Sum aggregation method*




*Supplementary figure 8: Distribution of the percentage of pairs of patients sharing identical causal gene for 1,000 randomly selected couples of individuals for 5 semantic similarity computation methods compared to the couples with maximum semantic similarity (red bar) assessed with maximum aggregation method*