Analysis and functional validation of candidate variants arising from human whole genome sequencing

Lukas Lange

Supervisors: Prof Jenny Taylor

Prof Sir John Bell

Dr David Sims

Wellcome Centre Human Genetics, Nuffield Department of Medicine

University of Oxford

This dissertation is submitted for the degree of

Doctor of Philosophy

Magdalen College July 2022

Abstract

Analysis and functional validation of candidate variants arising from human whole genome sequencing

Lukas Lange, Magdalen College, Michaelmas term 2020. Submitted as part of the requirements for the degree of Doctor of Philosophy

Whole genome sequencing (WGS) is increasingly used to diagnose rare genetic diseases (RGD). WGS data contains approximately 5,000,000 variants per patient, out of which often one variant causes disease. Variant prioritisation algorithms (VPA) help identify disease-causing variants and functional studies can provide evidence to confirm a variant's effect on a gene or gene product. In this thesis, I test and develop VPA on simulated and real patient data and conduct a functional study to contribute to the validation of *HDLBP* as a novel disease gene for Fine-Lubinsky syndrome (FLS).

Two established analysis frameworks, Exomiser and VAAST+Phevor, contain VPA that only use genotypic data (GA) and VPA that use genotypic and phenotypic data (GPA). The GA and GPA performance is benchmarked on eleven real WGS patient cases for which disease-causing variants were previously identified in known and novel genes. The GPA performed better than the GA, ranking more benchmark variants first (Exomiser: 4 vs. 5; VAAST+Phevor: 1 vs. 8), whilst reducing the percentage of variants requiring further analysis (Exomiser: from 32.3% to 2.2%; VAAST+Phevor: from 25.4% to 11.2%).

Identifying disease-causing variants in genes that are novel for a specific phenotype is challenging for GPA. A VPA called GPET is developed for disease-causing variant identification in novel genes based on genotypic, phenotypic, and tissue-specific expression data. GPET is benchmarked against Exomiser on a simulated dataset of disease-causing and non-disease-causing variants with imperfect phenotypic annotations, achieving an Area Under the Curve of 0.95, compared to 0.91 for Exomiser. GPET performs worse than Exomiser on the eleven RGD cases for known disease genes and better for all novel disease genes without phenotypic annotations.

Patients for one of the RGD cases are affected by FLS and carry a candidate variant in *HDLBP*. The variant is suspected to cause FLS by decreasing the RNA-binding activity of vigilin, the protein encoded by *HDLBP*. A functional study based on oligo(dT) capture is conducted, confirming reduced RNA-binding activity of vigilin due to the candidate variant.

The investigations in this thesis demonstrate that phenotypic and tissue-specific expression data can improve VPA performance for novel disease genes and provide evidence to support *HDLBP* as a disease gene for FLS.

Acknowledgements

First and foremost, I want to thank my supervisors, Prof Jenny Taylor, Prof Sir John Bell, and Dr David Sims. Jenny forever changed my life by accepting me into her group. I entered the Taylor Group as a chemical engineer and am leaving knowing that I will spend my career working in genetics. Despite my lack of experience at the start of my DPhil, Jenny not only gave me freedom to explore research areas I was interested in, she also provided resources to support my academic growth, including a bioinformatics fellowship that introduced me to David. David helped me develop my ideas for novel bioinformatics algorithms into functioning pipelines that were instrumental for much of the variant ranking work I conducted. He also provided invaluable guidance for my thesis and served as a diligent sparring partner. I would have never met David or Jenny without John. From my first interaction with John, he has been a mentor that provided direction not only for my academic development, which included introducing me to Jenny, but also for my personal development by supporting projects like the Rhodes Healthcare Forum. In addition to thanking my supervisors for their mentorship, I owe them my deepest gratitude for their patience and for letting me take advantage of opportunities for my professional development.

A very special thank you goes to all members of the Taylor Group, who taught me much of what I have learned in Oxford. In particular, I would like to thank Dr Alistair Pagnamenta for introducing me to every aspect of working with whole genome sequencing data. I could not have asked for a better teacher. Additionally, I want to

thank Dr Niko Popitsch for turning a curious engineer into an aspiring bioinformatician. I would also like to thank Dr Alfredo Castello and my collaborators in the Castello Lab, without whom I would not have been able to conduct my functional studies for *HDLBP*.

None of my work would have been possible without the members of the HICF2 consortium and the patient families that contributed their data to the HICF2 study, often knowing that a treatment might not yet exist for their disease. I will forever be in awe of the patient families' courage, wisdom, and dedication to a cause beyond their own fortune.

In addition to the many people that enabled my academic development, I would like to thank the Rhodes Trust for the financial support of my studies and for fostering a community that made me feel at home and provided infinite opportunities for learning. I would also like to thank Charles Conn for his continued mentorship in Oxford and beyond.

Finally, I would like to thank my family and friends. My parents, Angelika and Joachim, for having given me everything I have in life; my siblings, Karsten and Sophia, for never letting me lose my sense of humour; my best friends from Brussels, Michael and Sebastian, for constantly reminding me of what is truly important in life; my Oxford housemates - Hamish, Adam, Josh, and James - for endless fun and deep connection; my fellow DPhil sufferers - Bogdan, Virginia, Alex, (Nivi!), and Jonas - for never letting me forget that everyone had to go through the same experience; my German gang, Luca and Fritzi, for always making me laugh; and Harley, for teaching me more about myself than I ever thought I could know.

Declaration

I confirm that the work contained in this thesis is entirely my own, except where clearly stated otherwise.

Lukas Lange

Magdalen College

Related publications

- **Lukas Lange**, Alistair T Pagnamenta, David Sims, Jenny C Taylor, and the Health Innovation Challenge Fund 2 project team. A novel machine learning algorithm for variant prioritisation using genotype, phenotype, and expression data. *American Society for Human Genetics Conference*, 2018.

- **Lukas Lange**, Alistair T Pagnamenta, Niko Popitsch, Health Innovation Challenge Fund 2 project team, and Jenny C Taylor. Clinical usefulness of genotype-driven and genotype-and-phenotype-driven variant prioritisation algorithms (GAs and GPAs) for whole genome sequencing (WGS) analysis. *European Society for Human Genetics Conference*, 2017.

- **Lukas Lange**, Alistair T Pagnamenta, Stefano Lise, Susan Clasper, Helen Stewart, Elham S Akha, Samantha J L Knight, David A Keays, Jenny C Taylor, and Usha Kini. A *de novo* frameshift in *HNRNPK* causing a Kabuki-like syndrome with nodular heterotopia. *Clinical Genetics*, 90(3):258-262, 2016.

List of Figures

List of Tables

Nomenclature

Acronyms / Abbreviations

COL1A1 Collagen Type I Alpha 1 Chain

COL1A2 Collagen Type I Alpha 2 Chain

ACTC1 Actin, Alpha, Cardiac Muscle 1

CACNA1E calcium voltage-gated channel subunit alpha1 E

DCXR Dicarbonyl And L-Xylulose Reductase

DOCK11 Dedicator Of Cytokinesis 11

HDAC4 Histone Deacetylase 4

HNRNPQ Heterogeneous Nuclear Ribonucleoprotein Q

LoF Loss of function

POR Cytochrome P450 Oxidoreductase

PSTPIP1 proline-serine-threonine phosphatase interacting protein 1

PTCH1 Patched 1

RBPJ Recombination Signal Binding Protein For Immunoglobulin Kappa J Region

SAMD9L Sterile Alpha Motif Domain Containing 9 Like

SLC30A10 solute carrier family 30 member 10

TNNI2 Troponin I2, Fast Skeletal Type

TRMT1 TRNA Methyltransferase 1

WWOX WW Domain Containing Oxidoreductase

AAS Amino Acid Substitution

aCGH array comparative genomic hybridisation

AUC Area under the curve

BAF B-allele frequency

BDMR Brachydactyly mental retardation syndrome

BTL Binary Tissue Label

BWA Burrows-Wheeler Alignment tool

CADD Combined Annotation-Dependent Depletion

CASM Conservation-controlled amino acid substitution matrix

CDD Conserved Domain Database

CF Cystic fibrosis

CGH comparative genomic hybridisation

CNV copy number variant

dbSNP Single Nucleotide Polymorphism database

DCM Dilated cardiomyopathy

DDD Deciphering Developmental Disorders study

dNTPs deoxyribonucleotide triphosphates

DO Disease Ontology

eQTL expression Quantitative Trait Loci

EVS Exome Variant Server

ExAC Exome Aggregation Consortium

FLS Fine-Lubinsky syndrome

FMD Familial Meniere's disease

GA Genotype-based variant prioritisation algorithm

GADO GeneNetwork Assisted Diagnostic Optimization

GeneTIER Gene Tissue Expression Ranker

gnomAD Genome Aggregation Database

GO Gene Ontology

GPA Genotype-and-phenotype-based variant prioritisation algorithm

GTEx Genotype-Tissue Expression project

HGMD Human Gene Mutation Database

HGVS Human Genome Variation Society

HPO Human Phenotype Ontology

ICD-11 International Classification of Diseases and Related Health Problems, 11th Revision

IGV Integrative Genome Viewer

iPS cell Induced pluripotent stem cell

KTS Klippel-Trenaunay syndrome

MAF Minor allele frequency

MPO Mammalian Phenotype Ontology

NCBI National Center for Biotechnology Information

NGS Next generation sequencing

ONT Oxford Nanopore Technologies

OVA Online Variant Analysis tool

PAVAR Pathogenic Variant score

PCA Principal component analysis

PCR Polymerase chain reaction

PED Pedigree file used as input for Exomiser's inheritance pattern filter

PhenoDigm Phenotype comparisons for Disease Genes and Models

PolyPhen-2 Polymorphism Phenotyping v2 algorithm

PPI Protein-protein interaction

RBP RNA-binding protein

REVEL rare exome variant ensemble learner

ROC Receiver-Operator Characteristic

RPKM Reads Per Kilobase Per Million Mapped

RT-PCR Reverse transcription polymerase chain reaction

SBS Sequencing by synthesis

SIFT Sorting Intolerant From Tolerant

SNV single nucleotide variant

UMLS Unified Medical Language System

UniProtKB UniProt Knowledgebase

VCF Variant call format

VEP Variant Effect Predictor

VPA variant prioritisation algorithms

WES Whole exome sequencing

WGS Whole genome sequencing

WGS500 Whole Genome Sequencing 500 study

WOREE *WWOX*-related epileptic encephalopathy

Table of Contents

Chapter 1

Introduction

Afia, like her sister and three of her cousins, was born with dysmorphic facial features, skeletal abnormalities, severe developmental delay, brain abnormalities and large cornea. Afia, as well as her sister and cousins, has Fine-Lubinsky syndrome (FLS, OMIM:601353), a rare disease. While rare diseases are, as the name suggests, individually rare, Afia is not alone: globally, rare genetic diseases affect 400 million people, or approximately 5% of the population [1], making rare diseases a significant public health challenge. Fortunately, Afia and her relatives were referred to a clinician that had treated FLS cases before and recognised Afia's condition. Despite the gravity of her condition, Afia was lucky to receive a diagnosis at birth: the average rare disease patient in the UK waits approximately six years between the manifestation of first symptoms and an eventual diagnosis [2], a period so long that clinicians have coined the phrase 'diagnostic odyssey' for it.

While the clinician was quick to conclude that the collection of Afia's symptoms, or 'phenotypic features', suggested she was a FLS patient, this did not explain what caused the disease. FLS is a severe genetic condition and, as for 95% of all rare genetic diseases [3], no known cure exists. Yet, finding the genetic cause of the disease is crucially important for patients and their families. Questions like 'Did we

do something wrong?', 'If we had another child, would it also be affected?', 'Are there other patients like me out there?', 'How will my disease develop?' and 'Who is researching what I have?' plague affected individuals and their relatives and can be answered by identifying the genetic cause of the disease.

Some tests for single genes or even multi-gene panels are already part of clinical practice, e.g. *FBN1* and *TGFBR1* for Marfan syndrome [4], but they do not always provide the molecular diagnosis. There are more than 7,000 rare genetic diseases and molecular root causes for approximately 4,000, or 60% of them have been identified [5]. The genetic cause of FLS is not known.

Genetic testing methods have evolved over the past few decades to accomplish three things:

1. To give patients a diagnosis that carry variants in genes known for the patient's phenotype

2. To identify additional genes for diseases with already existing molecular characterisations

3. To identify new disease-causing genes for diseases with no existing molecular characterisations

The following literature review will elucidate how DNA sequencing (see Section 1.1) and bioinformatics have evolved to accomplish those goals, with a focus on variant prioritisation algorithms (VPA) (see Section 1.2) based on genotype (see Section 1.2.1), genotype and phenotype (see Section 1.2.2) and genotype, phenotype and expression data (see Section 1.2.3). Thereafter, I discuss the importance of functional validation of results from whole genome sequencing analysis using the example of FLS (see Section 1.3). Finally, an overview of the thesis is provided (see Section 1.4).

1.1 DNA sequencing for human disease

Genetic diseases can be caused by different types of genetic variants, and different methods are used to detect specific variants. In this section, I first describe the different types of genetic variants (see Section 1.1.1). Thereafter, I give an overview of technologies used to detect specific types of variants, from Sanger sequencing (see Section 1.1.2) to Next Generation Sequencing (NGS) (see Section 1.1.3), including targeted sequencing approaches (see Section 1.1.3.2) and whole genome sequencing (see Section 1.1.3.3).

1.1.1 Types of genetic variation

The Human Genome Variation Society (HGVS) maintains the Sequence Variant Nomenclature [6], a resource used to standardise how genetic variants are described. The Sequence Variant Nomenclature classifies variants into the following categories:

Substitution: a variant where one nucleotide is replaced by one other nucleotide, compared to a reference sequence.

Deletion: a variant where one or more nucleotides are not present, compared to a reference sequence.

Duplication: a variant where a copy of one or more nucleotides is inserted directly at the 3' end of the original sequence copy, compared to a reference sequence.

Insertion: a variant where one or more nucleotides are inserted at a specific position if the inserted sequence is not a direct copy of the sequence 5' of the insertion site, compared to a reference sequence.

Inversion: a variant where a sequence of multiple nucleotides replacing the original sequence is the exact reverse complement of the replaced sequence, compared to a reference sequence.

Conversion: a variant where a nucleotide sequence at one position of the genome is replaced by a nucleotide sequence from another position in the genome, compared to a reference sequence.

Deletion-insertion: a variant where one or more nucleotides are replaced by one or more nucleotides if the variation is not a substitution, inversion, or conversion, compared to a reference sequence.

Repeated sequences: a variant where a section of one or more nucleotides, the repeat unit, is repeated multiple times, compared to a reference sequence.

Complex: variants that cannot be described by the other listed variant types.

Other commonly used terms exist to describe variants, including single nucleotide variants (SNV) and copy number variants (CNV). SNV are substitutions, deletions, duplications, or insertions that only affect one nucleotide. CNV are larger variants that change the number of copies of a specific sequence present in a sample, including deletions and duplications.

Numerous technologies were developed to detect various types of genetic variants, as explained in the following section.

1.1.2 Sanger sequencing, linkage analysis, and comparative genomic hybridisation

The introduction of Sanger sequencing in the 1970s [7, 8] made possible high accuracy readings of defined regions of the genome. Sanger sequencing is primarily used to detect SNV and other smaller substitutions, deletions, duplications, and insertions.

Without prior knowledge of regions in the genome that likely harbour disease-causing variants however, Sanger sequencing is ineffective for the identification of novel disease genes. Technologies such as linkage analysis and comparative genomic

hybridisation (CGH) can be used to identify regions that might harbour disease-causing variants for Sanger sequencing.

Linkage analysis is used to identify the chromosomal location of disease genes. Linkage analysis relies on the observation that genes that are located close to each other on the chromosome remain linked together during meiosis. For families with affected family members, scientists can read out a few positions of the genome for each family member and determine the likelihood that a given chromosomal location is shared across affected individuals, but not unaffected family members. The chromosomal region that is shared between affected individuals likely harbours the disease-causing gene. Once that chromosomal region is identified, Sanger sequencing can be used to sequence every gene in the region and identify the gene, or genes, that carry disease-causing variants [9].

CGH is a tool used to identify CNV, where larger DNA segments are inserted or deleted at specific positions in the genome. To identify novel disease-causing genes with CGH, CNV are identified in patient cohorts affected by the same disease. Disease-causing genes are expected to exist at loci where inserted or deleted regions overlap for several patients.

CGH uses two different samples, a test and a control, that are labeled with fluorophores of different colours and hybridised to metaphase chromosomes, where the labeled samples will bind at their origin loci. The intensity of the fluorescent signal of each labeled sample is compared to each other along the length of the chromosome. A higher intensity of the test sample's signal indicates a gain of test sample DNA at that locus, whereas a higher intensity of the control sample's signal indicates a loss of test DNA in that region [10]. CGH can resolve regions of 5 - 10 million nucleotides.

A newer version of CGH, so-called array comparative genomic hybridisation (aCGH) uses microarrays instead of metaphase chromosomes. Microarrays are slides covered with an array of smaller DNA fragments from specific loci in the genome that

the test and the control sample hybridise to. The DNA fragments can be as small as 25 - 85 nucleotides, resulting in a significantly higher resolution of aCGH compared to CGH. The *CHD7* gene, for example, variants in which cause CHARGE syndrome, was identified using a combination of aCGH and Sanger sequencing [11].

1.1.3 Next generation sequencing

The combination of linkage analysis, CGH, or aCGH with Sanger sequencing is labour-intensive. Next generation sequencing (NGS) made the efficient sequencing of large numbers of genes and even the entire genome possible. In this section, I describe a method called sequencing by synthesis (SBS), which is used to produce over 90% of all NGS data [12] (see Section 1.1.3.1). Furthermore, I explain how SBS can be used to either sequence targeted regions of the genome (see Section 1.1.3.2), or the whole genome (see Section 1.1.3.3).

1.1.3.1 Sequencing by synthesis

SBS is conducted in four steps: library preparation, cluster generation, sequencing, and data analysis [13].

During library preparation, the DNA sample is randomly fragmented and adapters are ligated to the ends of each fragment.

Next, the library is loaded into small fluidic channels on a glass slide, also called a 'flow cell', where the DNA fragments are hybridised to surface-bound oligos that are complimentary to the library adapters. A compliment of the hybridised fragment is created by a polymerase, the double-stranded molecule is denatured, and the original template is washed off. Each fragment is amplified into distinct clonal clusters through a process called bridge amplification.

Once the clusters are generated, sequencing is initiated. A primer molecule sits at the end of each hybridised DNA fragment in the cluster. The sequencing relies on the incorporation of fluorescently labeled deoxyribonucleotide triphosphates (dNTPs) into the DNA strand that is synthesised as a compliment to each immobilised DNA fragment. Only one dNTP molecule is incorporated at each position based on the sequence of the template. After the addition of each nucleotide, a light source is used to excite the cluster, leading to the emission of a fluorescent signal for each cluster that is characteristic for each of the four DNA bases. Thus, each signal read-out for a cluster represents one base in the DNA fragment strand. The number of times that reaction is repeated determines the length of the read. For a given cluster, all identical strands are read simultaneously. Millions of clusters are analysed in parallel, producing millions of reads per sample. Reads are captured in a so-called FASTQ file.

FASTQ files contain one entry per read, represented as a string of nucleotides. Since SBS produces millions of reads in parallel based on fragmented DNA, algorithms like the Burrows-Wheeler Alignment tool (BWA) [14] or Stampy [15] have to be used to determine where in the genome each read stems from. These algorithms 'map' (align) each read to a position in a reference genome such as the human reference genome. This process, called 'alignment' or 'mapping', converts FASTQ files into so-called BAM files.

To ensure sequencing errors are caught, each nucleotide can be covered by multiple reads. The number of reads covering an individual nucleotide is called the 'coverage'.

By analysing where aligned reads differ from the reference genome, variants can be identified. The process of variant identification is called 'variant calling'. Examples of variant calling algorithms include the Genome Analysis Toolkit's (GATK) variant caller [16] and Platypus [13], which convert BAM files into so-called 'variant call format' (VCF) files.

Each variant in a VCF is annotated with additional information, such as the gene containing the variant, the impact of a variant on the resulting amino acid and subsequently the protein, and the frequency of the variant in specific populations [17–21], often referred to as the minor allele frequency (MAF). A variety of annotation algorithms exist, including ANNOVAR [18] and the Ensembl Variant Effect Predictor [19]. Annotated VCFs are then analysed to determine which variant might cause disease in the patient (see Section 1.2).

1.1.3.2 Targeted sequencing for gene panels and whole exome sequencing

Depending on the library preparation, SBS can be used to sequence either specific targeted regions of the genome, or the entire genome.

To only sequence targeted regions of the genome instead of the entire genome, library preparation techniques like target enrichment or amplicon generation are used. Target enrichment captures specific regions of interest that are isolated from the rest of the sample, while amplicon generation amplifies and purifies regions of interest for sequencing [12].

Targeted sequencing can be used to sequence a list of genes of interest for a specific phenotype, a so-called 'gene panel'. Gene panels are widely used in medical practice and represent the first-line diagnostic for many suspected genetic diseases, for example to screen newborns for Phenylketonuria, Sickle cell anemia, and Mucopolysaccharidosis Type 1 [22], or to test patients with familial hypercholesterolemia for disease-causing variants in *APOB* and *LDLR* [23]. Yet, the clinical scope of targeted gene panels for diagnosis is limited, with a diagnostic yield of often <50% [24, 25]. Particularly for cases such as Afia, with a disease like FLS for which no disease-causing genes have been published, more comprehensive sequencing approaches are needed.

Targeted sequencing can also be used to sequence all protein-coding regions of the genome, an approach called 'whole exome sequencing' (WES). WES successfully detects SNV and some CNV. Large-scale studies such as the UK BioBank [26], encompassing 500,000 healthy participants [27], deploy WES to improve our understanding of the genetic causes of common diseases. Similarly, focused rare genetic disease programs like the Deciphering Developmental Disorders (DDD) study use WES to diagnose rare disease patients. As of 2018, a sub-cohort of the DDD study first published in 2014 has achieved a diagnostic yield of 40% using WES [28]. Furthermore, numerous disease-causing genes were discovered using WES, including *DEPDC5* as a common cause of familial focal epilepsies [29].

However, the exome only represents 2% of the human genome and ignores noncoding regions relevant for disease [30, 31].

While improvements are being made for the detection of CNV from WES data, existing methods still fall short of the performance of techniques like aCGH [32].

Additionally, approaches for the detection of variants causing so-called repeat expansion disorders from WES data are improving, but are yet to replace more reliable methods such as polymerase chain reaction (PCR)-based assays. Repeat expansion disorders are caused by variants that expand regions of the genome in which short nucleotide sequences, usually consisting of two to six nucleotides, are repeated several times. Known repeat expansion disorders include Friedreich ataxia, Huntington disease and fragile X syndrome [33].

WES is also limited in its capacity to detect insertions, translocations, inversions, and variants in complex regions of the genome [34].

Finally, the coverage of WES drops significantly towards the edges of the protein-coding regions of genes, also called exons. Lower coverage at the edges of exons poses a risk to miss as much as 15% of disease-causing variants [35], for example

so-called splice-site variants. The analysis of Afia's genome presented in Section 1.3 uncovered a deleterious splice-site variant at the edge of an exon, hypothesised to be linked to FLS. With WES, this variant could have gone unnoticed.

Fortunately, most of the limitations of WES can be overcome with WGS, as described in the following section.

1.1.3.3 Whole genome sequencing

Instead of isolating a subset of genes or regions using targeted sequencing, SBS can be used to sequence the whole genome. Whole genome sequencing (WGS) overcomes many shortcomings of WES and has the potential to detect all classes of genetic variation, including SNV, CNV, translocations, inversions, and variants in noncoding regions and thus improve upon the diagnostic yield of WES [36–38].

As described in Section 1.1.3.2, WES coverage drops at exon borders, leading to disease-causing variants potentially not being detected [35]. Since WGS consistently covers exon borders, variants missed by WES can now be identified.

Furthermore, WGS includes noncoding regions of the genome, while WES only covers exons. WGS has already led to the discovery of disease-causing variants in noncoding regions that were missed by WES, such as variants in the 5' untranslated region of *GLS* causing glutaminase deficiency [34].

CNV can be called more reliably from WGS data than WES data, in part due to the consistent sequencing depth of WGS throughout the genome, and the fact that WGS also covers noncoding regions [39].

Methods for the detection of repeat expansions, however, are still improving. The previously mentioned variants causing glutaminase deficiency are trinucleotide repeat expansions that were detected with WGS [34]. Similar to repeat expansion detection via WES, however, the gold standard remain PCR-based assays [33].

A whole genome can now be sequenced for $1,000 [12], enabling large cohort studies such as WGS500, an early study of 500 genomes which evaluated the clinical utility of WGS [35]. All cases that were recruited into WGS500 had previously been screened with single gene Sanger sequencing or multi-gene panels of known genes relevant for the patient's phenotype, failing to deliver a diagnosis. The WGS500 study identified disease-causing variants in 21% of previously unsuccessfully genetically screened inherited cases overall [35], rising to 57% if the patient and their parents were sequenced, thereby demonstrating the importance of WGS for clinical settings. WGS500 paved the way for "Health Innovation Challenge Fund (HICF2)" funding to establish a clinical genome sequencing facility in Oxford, from which all of the WGS data analysed in this thesis stem, and for national genomics programmes such as the '100,000 Genomes Project' [40]. Participants of the 100,000 Genomes Project are allocated to disease-specific domains. The Rare Diseases Pilot study of the 100,000 Genomes Project achieved diagnostic yields ranging from 1.6 to 53.8%, depending on the disease domain [41].

Moreover, countries like Iceland run national sequencing projects producing population-scale genetic insights [42]. In addition to public projects, private companies such as Human Longevity are building up large WGS databases, selling tests both directly to consumers and to healthcare systems [43]. These developments are driven by the steep continued drop in sequencing costs, outpacing the computer industry's 'Moore's law' [44]. In light of this trend, it is no surprise that the computing resources needed to store and process genomic data are now rivalling the computing challenges of the traditional 'big data' producers YouTube and Twitter [45].

Given the demonstrated clinical value of WGS and the pace at which the amount of available data is increasing, the rate-limiting step for delivering valuable patient diagnoses is no longer sequencing, it is the fast and reliable interpretation of WGS data.

Despite the aforementioned advances, producing a molecular diagnosis for patients like Afia is challenging. It is becoming increasingly difficult to solve remaining rare genetic disease cases. Moreover, existing methods for WGS analysis are time-intensive and require expert analysts, adding complexity to the translation of these approaches into healthcare systems. Additionally, to make a definitive diagnosis for a patient, a potentially disease-causing variant has to be validated, either by identifying unrelated patients with a pathogenic variant in the same gene and the same phenotype, or functionally by definitively demonstrating the molecular impact of a variant *in vitro* or *in vivo*. To address these challenges, **the goal of my thesis is to identify the limitations of a subset of existing bioinformatic WGS analysis approaches (see Chapter 3), improve upon them (see Chapters 4 and 5), and functionally validate results for the patients in Afia's family (see Chapter 6)**.

1.2 Analysis of rare disease whole genome sequencing data with variant prioritisation algorithms

The advent of WGS and the translation of this technology into clinical settings is a catalyst for developing fast and accurate bioinformatics tools to analyse WGS data. The goal is to identify disease-causing variants among the approximately 5 million candidates called in human genomes. Filtering strategies can be used to reduce the number of variants that have to be manually analysed [46–48] and leverage different types of information annotated to variants in the VCF file after variant calling (see Section 1.1.3.1). Information used for filtering includes certain quality metrics, the MAF of specific variants [49–52], the likelihood of a variant to be disease-causing as predicted by specific algorithms, also called the 'pathogenicity', the inheritance mode, and evidence from literature on genes that have been linked to specific diseases.

While filtering approaches offer full transparency and flexibility in variant classification, the interpretation is time-consuming and bears a risk of false negatives [53]. VPA[1] are an alternative to filtering, aiming to rank likely disease-causing variants at the top of a list of potentially pathogenic variants, which can save time during interpretation [54] and reduces the number of false negatives excluded by filtering. In this section, I explain the basic concepts of three types of VPA: VPA based on genotypic information (see Section 1.2.1), VPA based on genotypic and phenotypic information (see Section 1.2.2), and VPA based on genotypic and phenotypic information, as well as expression data (see Section 1.2.3).

1.2.1 Variant prioritisation algorithms based on genotype

A number of VPA primarily use genotype-based information to rank variants. These genotype-based VPA are hereafter referred to as 'GA'. In this section, I highlight key concepts for pathogenicity prediction (see Section 1.2.1.1), describe reference databases used as a basis for GA (see Section 1.2.1.2), and explain how widely used GA work (see Section 1.2.1.3).

1.2.1.1 Key concepts for pathogenicity prediction

Algorithms like VAAST use different aspects of genotypic information to rank variants [20, 21, 55–58].

Most algorithms use the MAF as an indicator of pathogenicity. In theory, the rarer an allele is, the more likely it is to be damaging because it is under negative selection pressure. To estimate the MAF, population variant databases (see Section 1.2.1.2) are used, which the scientific community is encouraged to contribute to with every paper being published.

[1] Involving no or limited filtering steps, e.g. for MAF

Another input factor for VPA is the evolutionary conservation of specific protein sequences across species. If an amino acid is highly conserved across many different species, an amino acid change at that position is predicted to be damaging for the protein function.

Furthermore, the type of genetic variation plays an important role and algorithms differentiate between SNV, CNV, and other variants for pathogenicity prediction (see Section 1.1.1 for a list of the different types of genetic variation).

Genetic variations can either be synonymous, when a change in the nucleic acid sequence does not result in an amino acid change, or non-synonymous. The impact of synonymous and non-synonymous variants on protein function varies. Synonymous SNV do not lead to an amino acid change and were once thought to not have a functional impact [59]. More recently, synonymous SNV were however shown to affect gene function by impacting several processes, including the regulation of splicing [60] and transcription factors [61], protein synthesis [62], microRNA binding and mRNA folding [63]. Today, synonymous SNV are known to be associated to over 400 human diseases [64]. A non-synonymous SNV resulting in an amino acid change is called a missense variant. The severity of the SNV for protein function depends on the produced amino acid change. If the SNV results in the removal or introduction of a splice-site or the premature introduction or removal of a stop codon, the impact on protein function will be more severe than other amino acid changes. Similarly, the effect of insertions and deletions varies. Depending on the size of the insertion of deletion, the reading frame can be changed, substantially changing the protein sequence. Algorithms handle different types of variants that impact the protein differently, but will usually assign a severity metric to variants that increases from missense variants through to frame-shift variants [65].

Some algorithms also take into account if the suspected inheritance pattern of a patient's disease matches the zygosity of candidate variants [66].

VPA are used to predict pathogenicity of variants, but predictions vary depending on the algorithm used. To standardise the way the clinical genetics community classifies variants, the American College of Medical Genetics and Genomics (ACMG) published a set of guidelines for the interpretation of sequence variants. The guidelines rely on commonly available types of variant information, including population frequency data, computational predictions, functional information, and segregation data, to assign variants to one of five categories: 'pathogenic', 'likely pathogenic', 'uncertain significance', 'likely benign', and 'benign' [67]. The ACMG framework has since been integrated into a number of analysis tools and has become the variant classification standard for many laboratories [68].

1.2.1.2 Variant databases for genotype-based pathogenicity prediction

A number of databases exist that serve as resources to estimate allele population frequencies, check if variants have previously been identified in patients, and verify if variants of interest are known to be linked to a specific phenotype. The most widely used SNV databases include the Single Nucleotide Polymorphism database (dbSNP) [52], ClinVar [69], the Human Gene Mutation Database (HGMD) [70], SwissVar [71], the 1,000 Genomes Project database [49], and the Genome Aggregation Database (gnomAD) [72]. In its first release, gnomAD was known as the Exome Aggregation Consortium (ExAC) and contained exclusively exome data [73]. ExAC data can now be accessed through gnomAD. dbSNP contains small-scale variations and is a public resource. Similarly, ClinVar is a freely available database linking genetic variations to phenotypes, making ClinVar a useful resource for classification algorithms that need to be trained. HGMD aims to collate published gene lesions known to cause human inherited disease. An older version of HGMD is available free of charge to academic users, while a license can be purchased from QIAGEN for the latest release. SwissVar is a freely accessible portal to search variant entries in the UniProt Knowledgebase

(UniProtKB). ExAC aggregates WES data on over 60,000 unrelated individuals from various disease-specific studies and makes their variant database searchable free of charge.

Criticism has been voiced regarding the quality and independence of the aforementioned databases. Quality issues with variant databases arise due to discrepancies in variant classification practices of different laboratories [74–76], which manifest in public variant databases. A range of publications have questioned the pathogenic classification of variants in ClinVar [77–79]. To improve the quality of variant classifications, ClinVar has introduced a star system, indicating the reliability of each variant's classification [80]. The more accredited experts review a variant submission, the higher the star rating. The correct classification of variants is particularly important to create training datasets for variant classification algorithms. If merely the result of another classification algorithm is used by a submitter to ClinVar to determine if a variant is pathogenic, newly trained algorithms will never be able to outperform the limitations of the algorithm used for the original classification.

In addition to issues with variant classification, a lack of independent databases can pose a challenge for variant classification algorithms. To test if an algorithm performs well across a range of different datasets, the availability of datasets that are completely independent from the training dataset is crucial. For example, a classification algorithm could be trained on a ClinVar dataset of benign and pathogenic variants and subsequently tested on an HGMD dataset. If, however, a variant discovered by one clinical laboratory is catalogued in both ClinVar and HGMD, the two databases are now concordant and no longer independent. Since ClinVar is driven by variant submissions of research groups which also include variants found in publications, while HGMD is manually curated from those publications, these overlaps are inevitable. The publication of new, completely independent population variant databases that will

likely emerge from studies such as the 100,000 Genomes Project and the UK BioBank is crucial.

1.2.1.3 Existing variant prioritisation algorithms based on genotype

1.2.1.3.1 Established pathogenicity scores Based on the above-mentioned concepts (see Section 1.2.1.1) and databases (see Section 1.2.1.2), a range of algorithms were developed to predict the pathogenicity of variants in patient genomes. Hereafter, some of the most widely used algorithms for WES and WGS variant prioritisation will be explained.

An early algorithm to determine conserved elements in vertebrate genomes is PhastCons [81]. PhastCons is based on a two-state phylogenetic hidden Markov model, which is fitted to multiple alignments from several species by maximum likelihood. Based on this model, conserved elements are predicted. Output from PhastCons has become the basis for a number of VPA, including VAAST, which is discussed below.

The 'Sorting Intolerant From Tolerant' (SIFT) algorithm predicts if a single amino acid substitution (AAS) caused by a non-synonymous SNV (nsSNV) will affect protein function [21]. SIFT predicts the effect of all possible AASs based on sequence conservation (see Section 1.2.1.1). For a given protein sequence, SIFT builds a homologous reference dataset of functionally related proteins. For each position in the alignment, SIFT calculates the probabilities for all 20 possible amino acids. By scaling these probabilities by the probability of the most frequently observed amino acid at that position, the SIFT score is computed as the scaled probability of the AAS found in the patient genome. The lower the probability of the AAS, the higher the SIFT score [21].

The 'Polymorphism Phenotyping v2' (PolyPhen-2) algorithm is another method [82] for predicting the impact of SNV and resulting AASs on protein function. PolyPhen-2 uses a range of features generated from sequence annotations, multiple sequence alignments and 3-D protein structures, where available. The sequence annotations are used to determine if a variant lies in a transmembrane region, in which case the PHAT trans-membrane specific matrix score [83] is used. Similar to SIFT, multiple sequence alignments serve as an input to calculate a conservation-based score. PolyPhen-2 computes the likelihood of an amino acid occurring at a particular position relative to any position. The logarithm ratio of those likelihoods is calculated to produce a profile score. To generate a conservation score, PolyPhen-2 then calculates the difference of the profile scores for the alleles involved in the substitution. A large difference hence indicates a severe impact on the protein [82]. Lastly, PolyPhen-2 triggers the protein structure database [84] to determine if an AAS likely disrupts important structural components of the protein, for example the hydrophobic protein core. The more severe the impact on protein structure is, the more damaging a variant is assumed to be. The PolyPhen-2 score is generated with a Naïve Bayes classifier trained in a supervised learning model on the feature categories mentioned above. The algorithm is made available with two different models, one optimised for the identification of variants causing Mendelian diseases called 'HumVar', as well as another model called 'HumDiv', trained for the assessment of rare alleles at loci involved in complex phenotypes, where even mildly deleterious variants need to be detected [82].

It is important to note that conservation-based approaches suffer from two systematic limitations. First, while almost all human proteins are partially conserved across vertebrates, most also contain poorly-conserved regions. Many disease-causing alleles are located in those regions, but conservation-based approaches struggle to detect them. Second, conservation-based algorithms cannot accurately score the impact of stop

codons and frame-shifts, since these are not represented in the multiple alignments of homologous protein sequences from other species used as the basis of scoring [85].

Another example of a conservation-based GA is MutationTaster [57]. In addition to evolutionary conservation, the algorithm exploits splice-site changes, loss of protein features and potential changes in the amount of messenger RNA (mRNA) produced to predict pathogenicity. MutationTaster uses a naïve Bayes classifier and is distributed with three different prediction models, one for synonymous or intronic alterations, one for single AAS and one for complex changes in the amino acid sequence. MutationTaster2 is an expansion of MutationTaster to also rank short insertions and deletions and variants spanning exon-intron borders [86]. Furthermore, the algorithm can score regulatory features. Similar to MutationTaster, MutationTaster2 uses a naïve Bayes classifier to generate pathogenicity predictions. Since different tests are necessary to estimate the effect of alterations with different impacts on the protein sequence, MutationTaster2 incorporates individual classification models, one trained for each type of alteration.

The 'Combined Annotation-Dependent Depletion' (CADD) algorithm's pathogenicity prediction is based on the comparison of annotations of fixed or nearly fixed alleles in humans with those of simulated variants. Fixed or nearly fixed alleles in humans are variants with an allele frequency of $\geq$95%, whereas simulated variants are *de novo* variants that Kircher *et al.* [87] simulated using an empirical model. Pathogenic variants are depleted in fixed variants in humans due to natural selection, but the same does not apply for simulated variants. Using these two types of variants, the authors generated a training dataset of 29.4 million variants. Those variants were annotated with conservation-based scores, including PhastCons and PhyloP, regulatory features such as regions binding transcription factors and protein-level scores, including SIFT and PolyPhen, resulting in 63 annotations. Those annotations were used to generate features, on which a support vector machine with a linear kernel was trained to produce

a total of ten models, the average of which produces CADD's output 'C score'. The C score is genome-wide and has become a standard metric used in variant classification in many clinical genetics labs .

The rare exome variant ensemble learner (REVEL) [88] is another method developed to predict the pathogenicity of missense variants. REVEL was trained using a random forest classifier on a dataset of pathogenic and rare neutral variants that were annotated with the pathogenicity prediction scores of other existing tools: MutPred, FATHMM, VEST, Poly-Phen, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP, SiPhy, phyloP, and phastCons. In a benchmark analysis, REVEL outperformed the individual scores it was trained on.

1.2.1.3.2 WES and WGS analysis tools based on genotype The previously mentioned scores (see Section 1.2.1.3.1) significantly accelerate the process of analysing WES and WGS data for Mendelian disease patients. However, even when filtering a patient genome for variants that are predicted to be pathogenic, too many variants remain for geneticists to analyse manually. For that reason, pathogenicity scores are implemented in larger analysis pipelines to enable rapid WES and WGS data analysis.

One additional concept exploited in these pipelines are inheritance patterns. In rare genetic disease WES and WGS analysis, the inheritance pattern gives analysts meaningful clues about how to best conduct the analysis. If possible, not only the proband is sequenced, but also the parents, given that the difference between the parental genomes and the patient's contains valuable information to analyse the disease. For example, if the proband is affected but the parents are not, analysts will focus their energy on analysing *de novo* variants, which the parents do not carry, homozygous, or compound heterozygous variants. Additionally, maternally inherited heterozygous variants on the X chromosome will be investigated for male patients if the suspected inheritance pat-

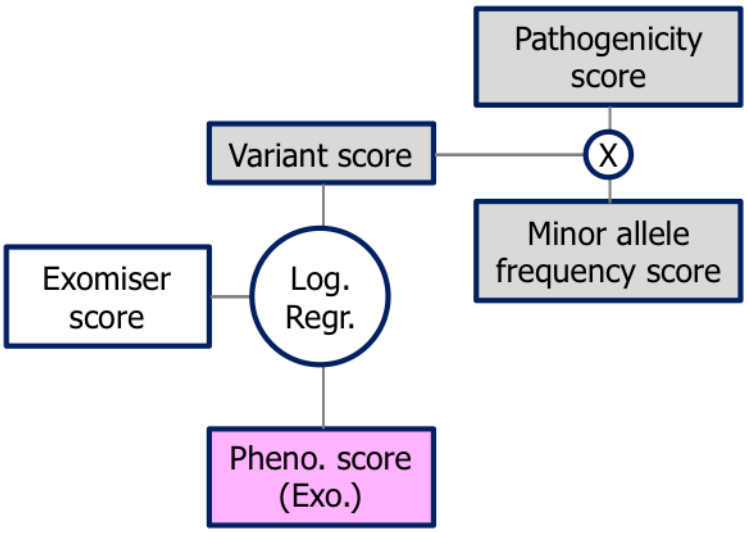


Fig. 1.1 **Overview of the Exomiser framework.** The combined Exomiser score is calculated via a logistic regression combination of a variant score and a phenotype score. The variant score is the multiplication of a variant's pathogenicity score and the variant's MAF score. The pathogenicity score of missense variants is calculated as the maximum of a variant's MutationTaster, Polyphen-2 and SIFT score, scaled from 0 to 1. For non-missense variants, the creators of Exomiser assigned fixed pathogenicity scores (see Table 1.1). The MAF score is a variant's MAF transformed from a percentage, where a lower percentage indicates a lower MAF, to a score from 0 to 1, where 0 equals 'benign' and 1 equals 'pathogenic'. A variant's MAF is the maximum MAF reported either in dbSNP or in the Exome Variant Server. If no MAF is reported, the MAF score is set to 1 [89].

tern is X-linked recessive. This filtering approach based on the suspected inheritance is efficient for reducing the number of variants that require closer investigation.

1.2.1.3.2.1 Exomiser Cases in which only the patient genome is available are called 'singletons', as opposed to cases in which patient and parental genomes can be analysed, so-called 'trios'. To facilitate singleton and trio WES or WGS analysis, several frameworks have been developed, including Exomiser [89]. Exomiser calculates a combined score based on a logistic regression combination of a 'variant score', detailed in the next paragraph, and a 'phenotype score', explained in Section 1.2.2.3. Figure 1.1 shows a high-level overview of the Exomiser framework.

The variant score is a multiplication of a MAF score and a pathogenicity score [65]. For missense variants, the pathogenicity score d is taken as the maximum of a variant's MutationTaster, PolyPhen-2 or SIFT score, scaled from 0 to 1:

$$d = max(MutationTaster\,|_{0...1}, PolyPhen-2\,|_{0...1}, 1-SIFT\,|_{0...1}) \tag{1.1}$$

If no annotation is available from either one of the three scores, Exomiser assigns a value of 0.6. For non-missense variants, fixed pathogenicity scores were assigned by the authors depending on the variant effect type. The authors differentiate between frameshift, nonsense, splice site, nonframeshift indel, stoploss, and synonymous variants (see Table 1.1).

Class	Frameshift	Nonsense	Splice site	Nonframeshift indel	Stoploss	Synonymous
Score	0.95	0.95	0.90	0.85	0.70	0.10

Table 1.1 Pathogenicity scores assigned by Exomiser for non-missense variants. For autosomal recessive inheritance with compound heterozygous variants, the variant score is computed as the average of the two highest scoring variants (adapted from [65]).

The MAF score is a variant's MAF transformed from a percentage value, where a lower percentage indicates increasing rarity, to a score ranging from '0 = benign' to '1 = pathogenic':

$$f = max(0,\ 1-0.13533\,e^{100MAF}),\ for\ MAF \leq 2\% \tag{1.2}$$

The MAF of each variant is the maximum MAF reported by a series of input databases. As of January 2021, the following databases can be used by Exomiser [90]: the 1,000 Genomes Project [49], the Exome Sequencing Project [91], ExAC [73], gnomAD [72], the TOPMed database [92], and the UK10K project [93]. If no

frequency data is available for a variant, the frequency score is set to 1. Exomiser is freely available [90].

Exomiser's variant score v is thus computed as follows:

$$v = d \times f \tag{1.3}$$

1.2.1.3.2.2 VAAST Another framework for WES and WGS trio analysis for rare genetic diseases is the 'Variant Annotation, Analysis, and Search Tool 2.0' (VAAST 2.0) [55]. VAAST 2.0 predicts the likelihood that a variant is disease-causing based on allele frequency, amino acid substitution (AAS) and phylogenetic conservation.

VAAST 2.0 is based on a conservation-controlled AAS matrix (CASM).

For each type of AAS, VAAST 2.0 calculates a severity parameter based on the conservation measure PhastCons and the relative frequency of the respective AAS in a disease and a non-damaging variant database.

Importantly, VAAST 2.0 produces a p-value-based score, making it possible to filter results based on statistical significance. This stands in contrast to Exomiser (see Section 1.2.1.3.2.1), which produces a score ranging from 0 to 1

VAAST 1.0 [94] calculates a variant severity parameter (a_i/h_i) as the ratio between the likelihood that an AAS does not contribute to disease (h_i) and the likelihood that it does (a_i). h_i is approximated with the frequency of the type of AAS in question in the 1,000 Genomes Project, whereas a_i is the frequency of the type of AAS among all disease-causing variants in HGMD. VAAST 2.0 extends the VAAST 1.0 approach by also factoring in phylogenetic conservation [55]. VAAST 2.0 first calculates a_i/h_i for each type of amino acid with a PhastCons score of 0 and 1. For any type of AAS i $(i = 1,2,...,m)$, there are n_i variants contained in the HGMD disease-causing variant

database. Each variant j ($j = 1,2,...,n_i$) has a PhastCons score P_{ij}. Since P_{ij} is a proxy for the likelihood that a variant exists in a conserved region, the likelihood that a variant is disease-causing for variants with a PhastCons score of 1 is estimated as

$$a_{i1} = \frac{(\sum_{j=1}^{n_i} P_{ij})}{C_D} \tag{1.4}$$

and for for variants with a PhastCons score of 0 as

$$a_{i0} = \frac{(\sum_{j=1}^{n_i} 1 - P_{ij})}{C_D} \tag{1.5}$$

with C_D as the total number of disease-causing variants in OMIM. To estimate h_i, the equation for variants with a PhastCons score of 1 is consequently

$$h_{i1} = \frac{(\sum_{j=1}^{n_i} P_{ij})}{C_N} \tag{1.6}$$

and

$$h_{i0} = \frac{(\sum_{j=1}^{n_i} 1 - P_{ij})}{C_N} \tag{1.7}$$

for variants with a PhastCons score of 0. C_N is the total number of non-disease-causing variants in the 1,000 Genomes Project database.

The severity parameter for AAS type i with a PhastCons score of 1 is hence a_{i1}/h_{i1} and a_{i0}/h_{i0} for a PhastCons score of 0. For variants with a different PhastCons score x ($0 < x < 1$), the ratio is calculated as follows:

$$\frac{a_{ix}}{h_{ix}} = \frac{a_{i0}}{h_{i0}} \times (1-x) + \frac{a_{i1}}{h_{i1}} \times x \tag{1.8}$$

a_{ix}/h_{ix} are the entries in the CASM, thus providing an estimate for the likelihood ratio of an AAS being pathogenic or benign.

The biggest drawback of these methods is that, for the approximately five million variants found in a human genome, the score distribution of the algorithms is close to bimodal, making it challenging to identify a single most pathogenic variant, as would be required to successfully find the disease-causing variant in a Mendelian disease. Additional information sources are required to overcome this challenge, as described in the following section.

1.2.2 Variant prioritisation algorithms based on genotype and phenotype

Reliance on genotypic information alone is challenging since human genomes contain 100 loss-of-function (LoF) variants, inactivating ≈20 genes [95]. That means that, even if the algorithms described in Section 1.2.1.3 worked perfectly and down-ranked all of the variants which do not lead to a LoF, there would still be ~100 variants with significant pathogenicity left for an analyst to screen manually. For this reason, a new class of algorithms has emerged, combining phenotypic and genotypic information to improve VPA. Hereafter, these methods will be described as genotype-and-phenotype-based VPA (GPA).

1.2.2.1 The Human Phenotype Ontology

To establish links between phenotype and genotype, GPA rely on frameworks like the "Human Phenotype Ontology (HPO)" [96–99], a standardised vocabulary aiming to capture all phenotypic abnormalities encountered in human disease[2]. The HPO is an ontology representing phenotypic terms (ontology nodes) and their inherent

[2] 13,000 HPO terms and over 156,000 hereditary disease annotations as of August 2018 [100]

hierarchy ("is_a" connections, ontology edges). Each node can be annotated with biological data like genes and diseases that have been associated with a phenotypic term (see Figure 1.2). Algorithms such as Exomiser and Phevor use this additional layer of information, connecting potentially disease-causing genes with phenotypic information, in variant prioritisation frameworks to improve the ranking mechanism (see Section 1.2.2.3).

In addition to VPA, a range of other applications have been developed based on the HPO to support rare genetic disease diagnostics. For example, computer vision algorithms that can identify phenotypic traits in photos of a patient's face, map those phenotypic traits to HPO terms and based on that infer which disease a patient might be affected by [101]. FDNA, a genetic disease diagnostic company, has even integrated one of these algorithms into a smartphone app called 'Face2Gene' [102]. Other algorithms, such as Phenomizer [66], use the collection of a patient's HPO terms to directly infer the disease a patient might be affected by.

Furthermore, platforms are emerging that try to identify patients that are similar to each other based on HPO terms. One such platform is PhenomeCentral [103], a portal for matching rare genetic disease patients with similar phenotypes and genotypes. In addition to patient matching functionality, PhenomeCentral includes an online version of Phenotips [104], an electronic medical record system for rare genetic disease patients. While PhenomeCentral matches patients on both or either of genotype and phenotype, other platforms exist that are purely focused on genetic matches, such as GeneMatcher [105]. GeneMatcher matches patients with each other that carry variants of uncertain significance in the same gene. Importantly, all of the above-mentioned matching platforms are used by clinicians, not patients, to protect patient privacy and ensure compliance with data sharing regulations.

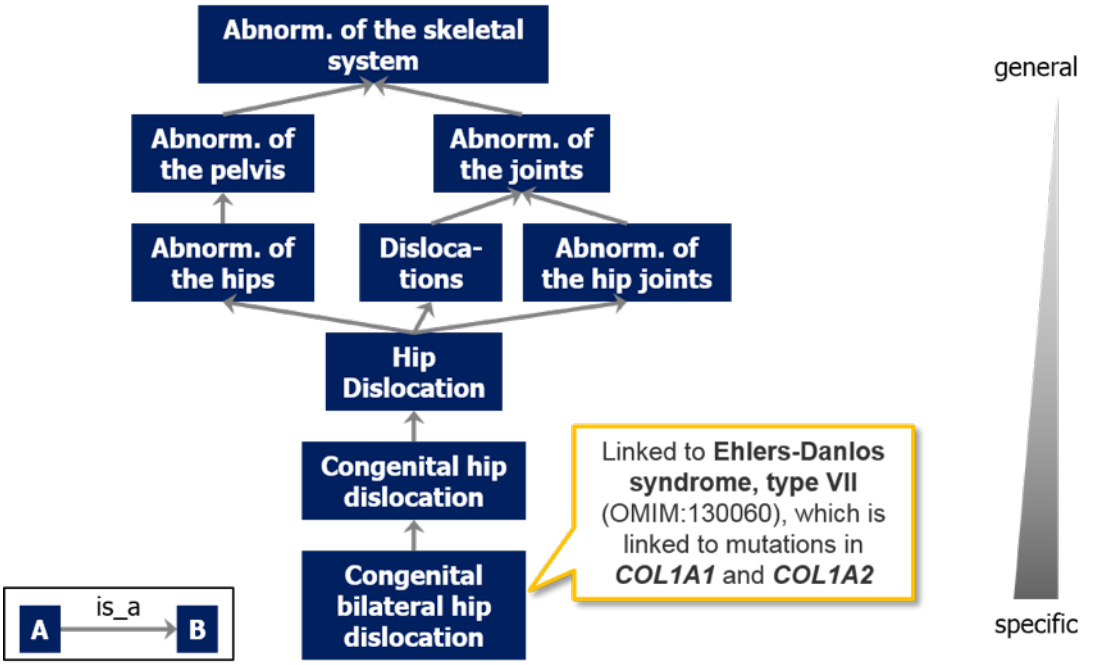


Fig. 1.2 **The Human Phenotype Ontology.** The HPO serves as a structured phenotype vocabulary. Nodes (blue boxes) represent phenotypic terms, edges (gray arrows) illustrate "is_a" connections. That way, highly specific phenotypic descriptions are linked to increasingly general descriptors. In this illustration, "congenital bilateral hip dislocation" is_a "congenital hip dislocation" is_a (...) is_a "abnormality of the skeletal system". The HPO is annotated with additional biological information. The term "congenital bilateral hip dislocation", for example, is associated with Ehlers-Danlos syndrome type 1 (OMIM:130060), which is linked to variants in *COL1A1* and *COL1A2* [96, 106].

1.2.2.2 Alternative ontologies

The HPO is only one of several established ontologies for tracking phenotypic information. The 'Unified Medical Language System' (UMLS) [107], distributed by the NIH, integrates several healthcare coding standards, which were built for different applications. While the HPO was designed as a tool to support biomedical research, the 'International Classification of Diseases and Related Health Problems, 11th Revision' (ICD-11) [108] and SNOMED [109] are vocabularies built primarily for the processing of medical billing information in healthcare systems. With the advent of large scale genomics projects that draw in patient information from various resources, the importance of ICD-11 and SNOMED for genomic analyses has increased.

Aside from ontologies built to capture phenotypic information, systems such as the 'Disease Ontology' (DO) and the 'Gene Ontology' (GO) were designed to describe relationships between diseases and genes and their affiliated biological products. The GO can for example be used to map protein-protein-interaction (PPI) networks, which provide additional insights for VPA [110].

Furthermore, ontologies can be exploited to draw connections between humans and other species. For example, the most widely-used ontology for linking human with mouse and rat phenotypes is the Mammalian Phenotype Ontology (MPO) [111].

Bioinformatics groups spend significant time on creating mappings between these resources to take advantage of all the available information for variant ranking [112].

1.2.2.3 Existing variant prioritisation algorithms based on genotype and phenotype

GPAs use the HPO to calculate how similar a patient's phenotype is to phenotypic terms linked to a candidate gene to inform ranking. For example, Phevor uses the ranked output of GA like VAAST to rank variants based on genotype and then re-ranks likely candidates based on phenotype, gene function and disease information [55, 58, 113, 94]. Exomiser is another example of a GPA. Exomiser filters out variants outside of the exonic target and variants with a too high MAF and then calculates a similarity score based on the patient's phenotype and phenotype data annotated to candidate genes. This score is combined with a genotype-based score to rank variants [89, 65]. These approaches have been shown to improve diagnostic workflows [53, 58].

GPAs combine ontologies such as the HPO with input from other tools, as described in (see Section 1.2.1.3), to better distinguish pathogenic from benign variants in patient genomes.

1.2.2.3.1 Exomiser's hiPHIVE One example, hiPHIVE, distributed as part of the Exomiser framework [89], calculates a combined score based on a variant score, described in Section 1.2.1.3.2.1, and a phenotype score, on a scale from 0 (benign) to 1 (pathogenic).

hiPHIVE's phenotype score is based on a similarity score, comparing the patient's phenotypic profile, described by HPO terms, with the phenotypic annotations of each candidate gene. Each gene receives three phenotype scores, one per species for which the phenotypic annotations were used: human via the disease-gene annotations in OMIM and Orphanet [114], mouse via the MPO and zebfrafish via zebrafish-specific phenotype databases [89]. If no phenotypic annotations are available for all three species, hiPHIVE instead uses a PPI to score how close the gene of interest is to genes with strong phenotypic similarity to the patient using a random walk algorithm [89, 115]. The resulting Exomiser phenotype score is the maximum of the species-specific phenotype scores and the PPI score.

The similarity score is calculated with a three-step process [116]. First, the ontology concepts are aligned and a significance score is computed. The significance score indicates how similar two terms are to each other. Two different algorithms, the Jaccard Index (sim_J) and the Information Content (IC), are used to estimate the significance score via a pairwise similarity comparison of each phenotypic feature.

sim_J scores similarity between an individual HPO term associated to the patient and an individual phenotypic term (HPO, MPO or Zebrafish) associated to the gene in question [117]. This results in a score between 0 (not similar) and 1 (identical). If a^p are the terms associated with the patient's HPO term p and a^g are the terms annotated to the gene's phenotypic term g, then

$$sim_J(p,g) = \frac{|a^p \cap a^g|}{|a^p \cup a^g|} \tag{1.9}$$

The *IC* of each term is calculated as follows:

$$IC(p,g) = -log_2\left(\frac{|annot_{pg}|}{|annot|}\right) \tag{1.10}$$

where $annot_{pg}$ is the number of genes annotated with the least common subsumen phenotype of the patient's phenotypic term and the gene's annotated phenotypic term. *annot* is the total number of terms in the HPO. In essence, that ratio describes the likelihood of a gene or patient profile being annotated with that particular phenotypic term. The resulting significance score s for each pairing of terms is calculated as the geometric mean of sim_J and *IC*:

$$s_{pg} = \sqrt{IC(p,g)\, sim_J(p,g)} \tag{1.11}$$

Second, overall phenotypic similarity scores are calculated between the patient's HPO profile and each gene's phenotypic profile. For this, the pairwise comparisons between each concept are considered, only using the best scoring matches of patient HPO term and gene phenotypic concept. The raw overall similarity of the patient's HPO profile and the gene's phenotypic annotations can be computed as either the maximum of the individual score matches (*maxScore*) or as their mean average (*avgScore*).

If $i = 1...m$ are the phenotypic terms for the patient's HPO profile p and $j = 1...n$ are the phenotypic terms for the gene g, then

$$maxScore(p,g) = max(s(i,j)),\ i = 1...m,\ j = 1...n \tag{1.12}$$

and

$$avgScore(p,g) = \frac{\sum_{i=1}^{m} max(s(i,j))_{j=1...n} + \sum_{j=1}^{n} max(s(i,j))_{i=1...m}}{m+n} \tag{1.13}$$

Third, since these phenotypic scores are not scaled between 0 and 1, making it difficult to compare results with each other, the phenotypic similarity scores are normalised. For each HPO term in the patient's profile, the optimal theoretical phenotypic term match is calculated. To scale the raw scores of the pairwise comparisons, the raw score is divided by the optimal theoretical score:

$$maxPercentageScore(p,g) = 100 \times \frac{maxScore(p,g)}{maxScore(p,\ optimal\ match\ for\ p)} \tag{1.14}$$

and

$$avgPercentageScore(p,g) = 100 \times \frac{avgScore(p,g)}{avgScore(p,optimalmatchforp)} \tag{1.15}$$

The phenotype score ϕ assigned by hiPHIVE for the similarity of a patient's HPO profile and a gene's phenotypic annotations is then calculated as:

$$\phi = avg(maxPercentageScore(p,g), avgPercentageScore(p,g)) \tag{1.16}$$

A logistic regression classifier was trained based on 10,000 disease-causing variants from HGMD and 10,000 benign variants from the 1000 Genomes Project to calculate

the combined Exomiser score E based on the variant score v (see Equation 1.3) and phenotype score ϕ [118]:

$$E = \frac{1}{1 + e^{-(-13.96+11.61(\phi+v))}} \tag{1.17}$$

Cipriani *et al.* [119] evaluated Exomiser's hiPHIVE on a cohort of 134 WES datasets from patients with confirmed molecular diagnoses for rare retinal diseases. Exomiser ranked 74% of the disease-causing variants as the top candidate and 94% in the top 5. Not using the HPO profiles of patients and only relying on genotype-based variant prioritisation reduced the performance, ranking 3% of candidate variants first and 27% in the top 5.

1.2.2.3.2 Phevor Another example of a GPA framework, Phevor [113], uses the output from algorithms like VAAST 2.0 and reranks it based on a phenotype score. Information about the patient is supplied in the HPO format or, alternatively, terms from the DO, MPO, GO or OMIM can be supplied. Based on the supplied terms, Phevor generates a candidate gene list using each ontology's annotations. This gene list can then be used to relate different ontologies to each other. Figure 1.3 shows how two generic ontologies A and B are connected with each other based on a hypothetical candidate gene list consisting of Gene X, Gene Y, and Gene Z.

The genes contained in the list are then used as starting points to propagate across each ontology to assign a phenotypic relevance score to each term in the ontology, as illustrated in Figure 1.4. Each ontology node i linked directly to a gene from the list receives a node score S_i of 1. From those starting nodes, the propagation moves to the children of each node. Each time an edge is crossed, the current value of the previous node is divided by two. If a node has two children, the starting node's value is divided by two twice, once for every child. Following the same procedure, the original seed

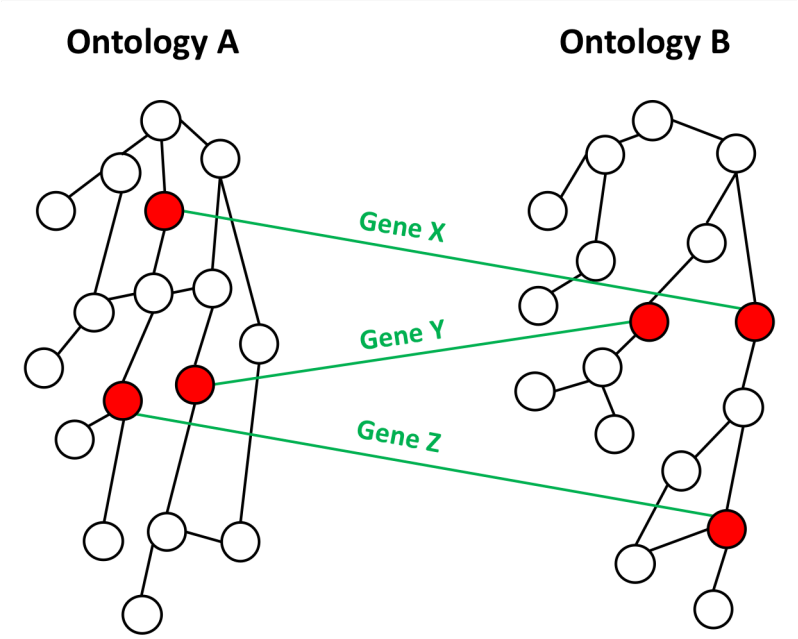


Fig. 1.3 **Combining ontologies with Phevor.** Phevor connects different ontologies with each other based on their gene annotations. Examples of ontologies used in this way include the DO, MPO, GO or OMIM. Shown here are two example ontologies A and B which are connected via three genes, Gene X, Gene Y, and Gene Z, represented by green lines. This cross-linking of ontologies allows Phevor to use annotations from different ontologies for the gene prioritisation process, for example phenotypic information from the HPO, and information about gene function from the GO. Red nodes represent the starting points for the propagation process described in Figure 1.4. This figure was adapted from Singleton *et al.* [113].

scores are also propagated to parent terms, until the root of the ontology is reached. Once propagation is completed, the node values are normalised to a score between 0 and 1 by dividing each value by the sum of all node scores in the ontology.

Next, each gene g annotated to the ontology is assigned an ontology-specific gene score $G_{g,ontology}$ equal to the maximum score of any node in the ontology which the gene is annotated to. Each gene receives a score for every ontology it is annotated to, resulting in multiple scores per gene. Those scores are subsequently summed up and renormalised to produce a final gene score G_g between 0 and 1:

$$G_g = \frac{G_{g,HPO} + G_{g,GO} + G_{g,MPO} + ...}{\sum_{h=1}^{n} G_h} \tag{1.18}$$

where G_h is the gene score of gene h and n is the total number of gene scores.

Importantly, genes which have no annotations in any ontology can receive significant scores in the propagation process, since they can be located at intersections connecting multiple highly-annotated nodes. The propagation process enables Phevor to also score novel genes that lack ontology annotations.

Subsequently, genes are ranked by their gene sum scores and percentile ranks are calculated. Phevor then calculates a disease association score D_g for every gene:

$$D_g = (1 - V_g) \times N_g, \tag{1.19}$$

where N_g is the percentile rank of the renormalised gene score G_g for gene g and V_g is the gene's percentile rank as received from the VPA used as input, for example SIFT, except for VAAST, for which the p-value can be used directly.

Next, Phevor also calculates a score H_g for each gene, indicating the likelihood that a gene is not associated with disease:

$$H_g = V_g \times (1 - N_g). \tag{1.20}$$

The Phevor score S_g is ultimately calculated as

$$S_g = log_{10} D_g / H_g. \tag{1.21}$$

With tools like Exomiser and the combination of VAAST and Phevor, which primarily focus on coding variants, the diagnostic yield in WGS studies is approximately 25-40% [120] for rare disease patients that did not receive a diagnosis from single gene or multi-gene panel testing. Comparisons of the algorithms have produced mixed results, with Exomiser achieving better results than VAAST combined with Phevor in some papers and vice versa in other publications [58, 113]. **The goal of Chapter 3 is therefore to compare the performance of the two combinations of algorithms with each other based on a number of patient cases from the HICF2 study.**

1.2.3 Variant prioritisation algorithms based on genotype, phenotype, and expression data

As mentioned previously (see Section 1.2.2), significant performance improvements in prioritising disease-causing variants in patient genomes were achieved by adding additional data sources that are independent from the original inputs used in the first wave of ranking algorithms. Yet, it remains challenging to quickly and accurately identify variants that are highly pathogenic but reside in genes which haven not previously been annotated with phenotypic information, despite approaches like Exomiser's usage of additional ontologies and PPI networks and Phevor's propagation of multiple ontologies. Therefore, analysing cases with novel genes is hard and further improvements are needed. Afia and her relatives are one such case, with a likely disease-causing

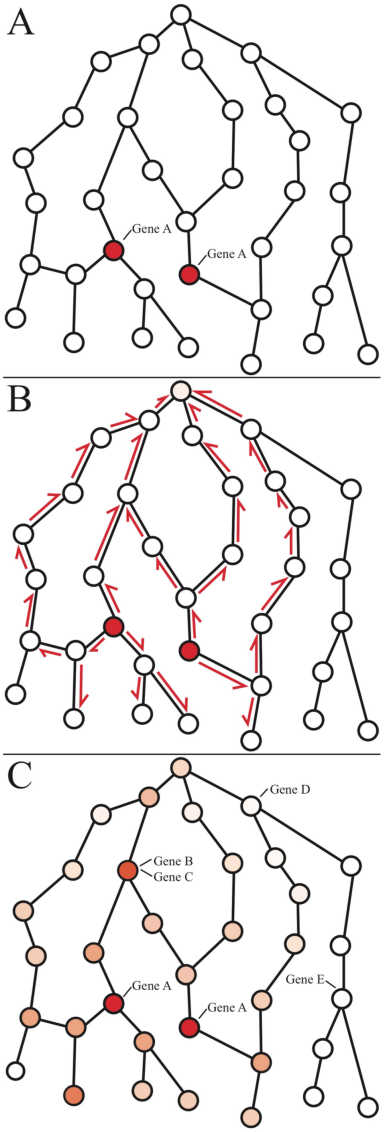


Fig. 1.4 **Ontological propagation with Phevor.** The tree-like structure illustrates an ontology used for the phenotype score calculation, for example the HPO. Each node represents a term from the ontology. **Panel A** shows an ontology with two user-provided terms (nodes) shown in red. Here, Gene A has previously been annotated to both of these terms. At the start of the propagation, the two red-highlighted nodes receive a score of 1. Subsequently, the propagation to children and parent nodes starts, as illustrated in **panel B**. Every time an edge is crossed to reach a neighbouring node, the score of the previous node is divided by two. In **panel C**, the final result of the propagation is shown. The node colors represent the magnitude of the propagation scores, with dark red nodes receiving the highest scores and white nodes receiving the lowest scores. In this example, propagation has identified two new gene candidates, Gene B and Gene C, which are not associated with the original seed node, but are annotated to nodes that received high propagation scores. Gene D and Gene E represent examples of genes annotated to nodes that did not receive high scores. This figure was reproduced from Singleton *et al.* [113].

variant in *HDLBP*, a gene without HPO term annotations. To support variant ranking for genes without phenotypic annotations, expression data can be a useful additional source of evidence.

1.2.3.1 Gene expression databases

The proliferation of high-throughput RNA sequencing technologies has enabled the construction of large tissue-specific gene expression databases, providing data to delineate the effect of genetic variation on gene expression in human tissues [121].

One of the most-widely used tissue-specific gene expression databases was generated by the Broad Institute's Genotype-Tissue Expression (GTEx) project [121]. In August 2018, GTEx contained RNA-seq data on 53 tissues from 11,688 samples [122] collected post-mortem. Donor eligibility is broad and not limited to specific disease or conditions. GTEx is freely available for academic users.

Another widely used database is the EMBL-EBI's Human Protein Atlas: comparable to GTEx, but smaller in size, the Human Protein Atlas, based on samples from 122 human donors, holds RNA-seq data on 37 tissue types and protein expression data on 44 tissue types [123].

The Genomics Institute of the Novartis Research Foundation (GNF) Gene Atlas is an earlier version of a gene expression database [124]. The GNF Gene Atlas contains expression data on 17,761 human genes and 79 human tissues, generated via microarray chips from a small number of donors.

Other gene expression databases include the Gene Expression Atlas [125], the RNA-Seq Expression Atlas [126], the ArrayExpress database [127] and the Gene Expression Omnibus [128].

1.2.3.2 Established links between gene expression and disease-associated genes

Various studies suggest that elevated expression of genes in specific healthy tissues implies an increased relevance of those genes for the function of those tissues. This indicates that pathogenic germline variants are more likely to affect tissues where the variant-carrying genes are more highly expressed. Oellrich *et al.* [129] demonstrated that 72-76% of phenotypes in knock-out mice are associated with disruption of genes expressed in the affected tissue. The authors subsequently used their gene expression-phenotype map to add an additional information layer to their disease gene ranking tool 'Phenotype comparisons for Disease Genes and Models' (PhenoDigm) [116]. PhenoDigm was originally designed to rank potential disease gene candidates based on their annotations with various model organism phenotypes. The combination of PhenoDigm with gene expression data showed improved performance compared to the original version of the algorithm. The authors do, however, note that a significant number of phenotype-tissue associations are spatially separated, for example the phenotype 'total body fat', which is correlated with genes expressed in various brain tissues. In a different study, Barshir *et al.* [130] demonstrated that, out of a set of 233 disease-associated genes with pathogenic germline variants, a majority had elevated expression levels in the affected tissues, as well as increased numbers of tissue-specific protein-protein interactions.

Furthermore, it has been shown that *de novo* disease candidate variants in patients with congenital heart disease can be successfully distinguished from controls based on gene expression in heart tissue [131]. In addition to the above-mentioned examples, there is a large literature corpus on examples of genes that are associated with specific diseases and exclusively expressed in the affected tissues. Examples include *MYH6*, *MYH7* and *TNNI3*, which are linked to a number of cardiac diseases and have highly elevated expression in the heart, compared to other tissues [132, 133]. Similarly, the Parkinson's disease- and dementia-associated genes *SNCA* and *SNCB* are highly

expressed in several brain regions and nervous tissue [132]. At the same time, related effects have been observed for conditions affecting multiple tissues, such as cystic fibrosis (CF). CF is a multi-system disease affecting several tissues, including the pancreas, lungs and sinus, all of which the primary CF gene *CFTR* is highly expressed in [132].

Mosley *et al.* [134] showed correlations between genotype, gene expression in the form of protein levels and phenotype. Instead of analysing associations between the measured expression levels of variant-harbouring genes in tissues linked to a specific phenotype, the group simulated a 'virtual proteome' by linking genotype and expected protein levels. With a study of over 40,000 individuals, they were able to demonstrate correlations between 55 simulated proteins and 89 distinct disease phenotypes.

However, a large number of genes are highly expressed in a number of tissues, not just disease-associated tissues. Approaches exclusively focusing on relative expression of single genes across different tissues may therefore be limited in their ability to identify genes associated with disease based on their tissue-specific expression [130, 135, 136]. Additionally, the expression profiles of known disease-associated genes can also be counter-intuitive, with lower expression in tissues relevant for the phenotype. One such example is *SIK1*, which is associated with severe developmental epilepsy, but is lowest expressed in brain and highest in skin, without a known skin phenotype [137]. Certain studies suggest that, instead of looking at individual genes, the focus should be on protein complexes, which collectively are elevated in affected tissues, while this is not necessarily the case for individual genes [138]. There is additional evidence suggesting that tissue-specific co-expression networks can further help decipher the genetic signature of diseases [135].

1.2.3.3 Disease-gene prioritisation algorithms based on expression data

To test the power of gene expression data for ranking candidate genes arising from WES and WGS data, several groups have implemented prioritisation algorithms.

1.2.3.3.1 CANDID Hutz *et al.* [139] developed CANDID, a gene prioritisation algorithm that ranks genes according to their importance for complex human traits. For the ranking, the algorithm uses up to eight different criteria: cross-species conservation, PPI, information from publications, protein domain descriptions, linkage analysis results, association analysis results, custom data and gene expression profiles [139]. For each criterion, CANDID assigns a score. Criterion-specific scores are normalised and weighted according to user-specific weighting settings, producing a final ranking of all candidate genes.

NCBI's HomoloGene database is used to score genes according to cross-species conservation. HomoloGene annotates each gene with a label indicating the organism harbouring the most distant homolog. Human-specific genes receive a score of 0, whereas genes labeled 'Eukaryota' receive the highest score of 1. Scores for other used species are distributed evenly.

PPI is factored in by CANDID through the NCBI Gene database, which contains information on PPI. A gene's PPI score is simply the sum of the publication and protein domain scores of all of the gene's interacting partners. That way, genes that receive a 0 for all other scores can still achieve a high PPI score.

To score genes according to information from publications, users supply CANDID with a list of keywords that would typically be used in a literature search for the complex human trait of interest. CANDID then searches PubMed with those keywords and ranks genes according to their association with PubMed search results. CANDID was developed before wide-spread adoption of the HPO.

To utilise protein domain information, users supply a list of keywords related to the trait of interest, which are used to search the National Center for Biotechnology Information (NCBI) Conserved Domain Database (CDD). The CDD search produces a list of protein domains and genes containing those domains. Every gene that is at least contained in one search result output receives a score of 1, all other genes get a score of 0.

Furthermore, CANDID utilises linkage data by assigning a linkage score to each gene based on the LOD score of the gene's location.

SNV association data can be factored in by supplying CANDID with a list containing SNV and their associated p-values. Each gene is assigned an association score based on the SNV with the lowest p-value in that gene.

CANDID also accepts custom scores for genes with higher user interest, which result in a custom data score for genes.

To factor gene expression into the analysis, CANDID queries the GNF Gene Atlas. CANDID compares each gene's expression across the 79 human tissues covered in the GNF Gene Atlas. Genes receive a score of 1 for the tissue they are most highly expressed in and relative scores, according to their relative expression compared to the tissue with the highest expression, for all other tissues. That way, genes receive a high score for the tissue they are highest expressed in and low scores for all other tissues. Importantly, house keeping genes, which are generally equally highly or lowly expressed across tissues, receive roughly the same score across tissues. Users must supply tissue code(s) corresponding to the tissue(s) of interest. A gene's overall expression score then corresponds to the normalised sum of all scores in the tissues indicated by the user.

The final CANDID score is calculated by multiplying each of the individual scores with the user's score-specific weighting and summing the result up. Genes are ranked in order of their final CANDID scores.

1.2.3.3.2 The Gene Tissue Expression Ranker (GeneTIER) Another algorithm utilising gene expression data, GeneTIER, was developed by Antanaviciute *et al.* [140]. While CANDID draws on knowledge-based inference as well as gene expression data, GeneTIER is purely built on experimental gene expression data. GeneTIER draws on expression data from the Gene Expression Atlas, RNA-Seq Expression Atlas, ArrayExpress, and Gene Expression Omnibus.

GeneTIER calculates an expression score for genes based on three factors: expression levels in affected tissues, variance of expression across tissues, and differences in expression levels between affected and unaffected tissues. The algorithm individually calculates scores for all genes in the databases from human RNA sequencing, human microarrays, mouse RNA sequencing and mouse microarrays respectively. The dataset- and species-specific scores are then combined to produce the final GeneTIER score. Users can adjust the relative contribution of human vs mouse data to the final result. Users supply GeneTIER with a list of affected tissues and a list of candidate disease genes. GeneTIER then uses its algorithms to rank the genes according to their likelihood of causing the disease that is affecting the supplied tissues.

To test the algorithm, the authors generated a benchmark dataset using the HPO. Via the HPO, the authors identified a list of 1,000 gene-disease associations for testing. HPO terms annotated to the selected diseases were mapped to tissues they likely affect. In the benchmark assessment, GeneTIER achieved an area under the curve (AUC) of 0.78-0.83, depending on the specific testing methods, thus not only providing validation for GeneTIER's ranking algorithm, but also indirectly demonstrating that

the HPO implicitly contains information about expression levels of genes in specific tissues.

1.2.3.3.3 Endeavour A third example is Endeavour by Tranchevent *et al.* [141]. To prioritise candidate genes, users first supply the algorithm with a process of interest. This can either be a list of genes that are known to be associated with the process of interest, or an OMIM disease. Next, the user supplies a list of candidate genes of yet unknown significance for the process of interest. Subsequently, Endeavour ranks the candidate genes according to their relevance for the process of interest.

Based on the process of interest, Endeavour builds a model of the likely characteristics of genes that are relevant to the process. Endeavour uses 75 different data sources and one sub-model is created and trained for each data source. The data sources represent a range of input factors, including PPI, chemical data, gene and protein function, sequence-based features, phenotypic information, and expression data. For example, the GO is one such data source. For the genes associated to the process of interest, Endeavour determines which features from the GO are overrepresented, thus giving an indication for which features are important for that process. Endeavour then ranks the candidate genes based on how closely they comply with the trained sub-model, i.e. the overrepresented features in the case of the GO. Once a ranking is determined for each individual data source, the scores for each gene are combined into one overall score, resulting in the final gene ranking.

The authors validated the algorithm based on a number of gold standard gene datasets, consisting of genes which are already known to be associated with the process of interest. For genes derived from the HPO, Endeavour achieves an AUC of 0.8834, both validating the algorithm's performance and, similar to GeneTIER, suggesting a significant link between gene expression levels and the HPO.

While there is room for improvement for algorithms such as GeneTIER and Endeavour, the demonstrated results indicate that gene expression data can be used for disease gene prioritisation.

1.2.3.3.4 The GeneNetwork Assisted Diagnostic Optimization (GADO) algorithm Deelen *et al.* [142] published the GADO algorithm. GADO ranks candidate genes based on a patient's HPO terms.

GADO's predictions are based on observed co-regulation of genes annotated to a specific HPO term with other genes. By exploting co-regulation, GADO can also rank genes that are not annotated with HPO terms.

The authors analysed co-regulation patterns of genes based on RNA-seq data from 31,499 samples in the European Nucleotide Archive [143], resulting in a prioritisation score for each gene for each HPO term. Since every gene is annotated with a prioritisation score for every HPO term, GADO works for genes with known HPO term annotations, as well as genes without an established link to a phenotype.

Co-regulation patterns are established with a principal component analysis (PCA) of all genes in their sample. Using this approach, the authors identified 1,588 principal components, each of which describes co-regulation between genes. Each gene in the dataset has a coefficient assigned for each principal component, describing the relevance of the principal component for that gene.

Next, the authors calculated the relevance of each principal component for each HPO term. In the HPO, each term is annotated with a list of genes known to cause the specific phenotype. For example, dilated cardiomyopathy (HPO term HP:0001644) is annotated with 142 different genes[3], including *ACTC1*, *COX1*, and *ABCC9*. The authors then calculated how relevant each principal component is for the set of genes

[3] as of October 2020

annotated to dilated cardiomyopathy, resulting in a relevance coefficient for each HPO term and each principal component.

Once a relevance coefficient was computed for each gene-prinicipal component relationship and each HPO term-principal component relationship, the authors were able to calculate the relevance of each gene for each HPO term. Similarly, when supplied with a list of HPO terms, GADO can calculate the combined relevance of each gene for the HPO profile. Thus, when supplied with a patient's HPO terms, GADO produces a list of genes ranked according to their relevance for the patient's phenotypic profile.

The authors benchmarked GADO against Exomiser on a WES cohort of 83 patients with confirmed genetic diagnoses. While GADO achieves a lower median rank for the causative variant than Exomiser (12.5 compared to 21), Exomiser ranks more variants in the top three (28 compared to 14). Furthermore, the authors applied GADO to a cohort of 61 patients that did not previously receive a diagnosis via WES, yielding causative gene candidates for ten cases.

1.2.3.4 Links between the Human Phenotype Ontology and tissue-specific disease gene expression

GADO does not differentiate between tissues in its analysis of expression data. While the analyses conducted for the development of GeneTIER and Endeavour implicitly suggest that the HPO contains meaningful information on tissue-specific gene expression levels, Feiglin *et al.* [132] conducted a study explicitly investigating this relationship.

The authors analysed the correlation between the HPO terms annotated to disease genes from OMIM and the expression level of those genes in tissues assumed to be affected based on the disease's HPO terms. For this purpose, they created a mapping

between 25 high-level HPO terms and GTEx tissues. For example, the HPO term 'Abnormality of the adrenal glands' (HP:0000834) is linked to the tissue 'adrenal gland'. Hence, genes annotated with that HPO term (and all of their children) are assumed to have increased expression in the adrenal glands [132]. To approximate 'elevated expression levels', Feiglin *et al.* [132] introduced two relative metrics: the expression of a gene in question in a phenotype-implicated tissue relative to all other tissues (referred to as the 'cross-tissue' expression level), and the expression of a gene in question in a phenotype-implicated tissue relative to all other genes in that same tissue (hereafter referred to as the 'cross-gene' expression level). Using the mentioned mapping as well as cross-tissue and cross-gene scores, the authors demonstrated that genes linked to OMIM rare genetic diseases generally are more highly expressed in tissues implicated by HPO terms linked to the disease than they are in tissues not implicated by the phenotype.

Overall, these results hold up for each individual expression metric (cross-tissue and cross-gene) as well as for their combination. Results vary, however, depending on the tissue. The cross-gene score for Nerve tissue achieves the highest AUC ($\approx$0.70), compared to the lowest AUC for vagina tissue's cross-gene score ($\approx$0.45).

1.2.3.5 A novel algorithm using genotypic, phenotypic, and tissue-specific gene expression data for variant prioritisation

Prioritisation of variants in genes with no known phenotypic annotations remains a challenge.

GPAs like Exomiser and Phevor predict the disease-causing variant to be pathogenic and thus assign a high rank relative to other variants. The phenotype-based scoring stage (see Sections 1.2.2.3.1 and 1.2.2.3.2) however will downgrade the candidate variant, due to a lack of phenotypic annotations. Exomiser relies on phenotypic matches

with model organisms as well as PPI data for novel gene discovery, while Phevor builds on its ontology propagation approach.

Deelen *et al.* [142] demonstrated with their GADO algorithm that gene expression data can be a valuable resource to prioritise genes based on gene co-regulation (see Section 1.2.3.3.4). While Deelen *et al.* used gene expression data regardless of the tissue of origin, Feiglin *et al.* [132] showed that variants confirmed to be disease-causing are on average more highly expressed in tissues that are likely affected based on the disease's HPO terms (see Section 1.2.3.4). Gene expression data in tissues indicated by HPO terms could thus be an informative input for variant prioritisation, particularly for genes lacking phenotypic annotations.

In addition to GADO, other approaches linking genotypic, phenotypic, and expression data exist, but with limited scope. Tools such as QueryOR are using GTEx expression data to annotate genes to have additional information for the result interpretation step, but they lack a link to HPO terms and do not use the data to calculate a ranking score to facilitate analysis [144]. MendelScan by Koboldt *et al.* [145] integrates segregation data, MAF, predicted protein effect and gene expression data, but expression data has to be individually supplied for the disease of interest, which limits its scope to diseases primarily affecting one tissue.

The goal of Chapter 4 is therefore to develop an algorithm that successfully combines genotypic, phenotypic and tissue-specific expression data for rare genetic disease WGS variant prioritisation. The goal of Chapter 5 is to validate this novel algorithm on real patient cases.

1.3 Functional validation of variants arising from whole genome sequencing data

VPA are useful to surface potentially disease-causing candidate variants. To determine if a candidate variant should be reported back to the patient, many clinical geneticists adhere to the ACMG guidelines published in 2015 [67], as previously mentioned (see Section 1.2.1.1). The classification of candidate variants in genes that are novel for the observed phenotype, such as Afia with FLS, can require additional evidence from functional studies. In this section, I describe how data from *in vitro* or *in vivo* functional studies affects the classification of candidate variants (see Section 1.3.1) and provide details on FLS (see Section 1.3.2) and the disease's link to a novel candidate gene, *HDLBP* (see Section 1.3.3).

1.3.1 Impact of *in vitro* and *in vivo* functional studies on variant classification

The ACMG guidelines allocate variants to one of five categories: 'pathogenic', 'likely pathogenic', 'uncertain significance', 'likely benign', and 'benign' [67]. The categories are assigned using a set of rules that determine how evidence of pathogenicity and evidence of a benign impact should be evaluated.

Evidence for pathogenicity is classified into four criteria: 'very strong', 'strong', 'moderate', and 'supporting'. Evidence for a benign impact is classified into three criteria: 'stand-alone', 'strong', and 'supporting'. Each criterion is further divided into several categories. For example, a variant leading to a frameshift in a gene where LoF is a known mechanism of disease is considered 'very strong' evidence for pathogenicity.

In vitro or *in vivo* functional studies supportive of a damaging effect on the gene or gene product, referred to as 'PS3' level evidence in the ACMG guidelines, are classified as 'strong' evidence of pathogenicity. One PS3 criterion on its own is not sufficient to classify a variant as pathogenic or likely pathogenic. Instead, at least an additional two 'supporting' or one to two 'moderate' pathogenicity criteria are required to classify a variant as likely pathogenic.

Initiatives such as Solve-RD, funded by the European Commission, are building up processes to systematically functionally validate candidate variants in newly discovered disease genes [146]. Functional studies examine the link between the disease phenotype (see Section 1.3.2) and the effect of the candidate variant on the suspected disease gene (see Section 1.3.3).

1.3.2 Fine-Lubinsky syndrome

FLS is a very rare genetic disease [147]. While FLS cases have previously been reported, no molecular diagnoses have been made. The phenotype includes plagiocephaly, megalocornea, cleft palate, digital abnormalities, dysmorphic facial features, moderate developmental delay and body asymmetry [148, 147]. Structural brain abnormalities, deafness and shallow orbits have also been reported (see Figure 1.5). Aside from two affected siblings, most reported cases have been sporadic [147]. Here, I study a consanguineous family with five affected relatives, one of which is Afia.

1.3.3 *HDLBP* and its potential links to FLS

The protein expressed by *HDLBP*, commonly referred to as 'vigilin' and the lead-candidate for our FLS case, is the largest RNA-binding protein (RBP) in the KH domain-containing family and one of the largest known RBPs [149]. Vigilin has

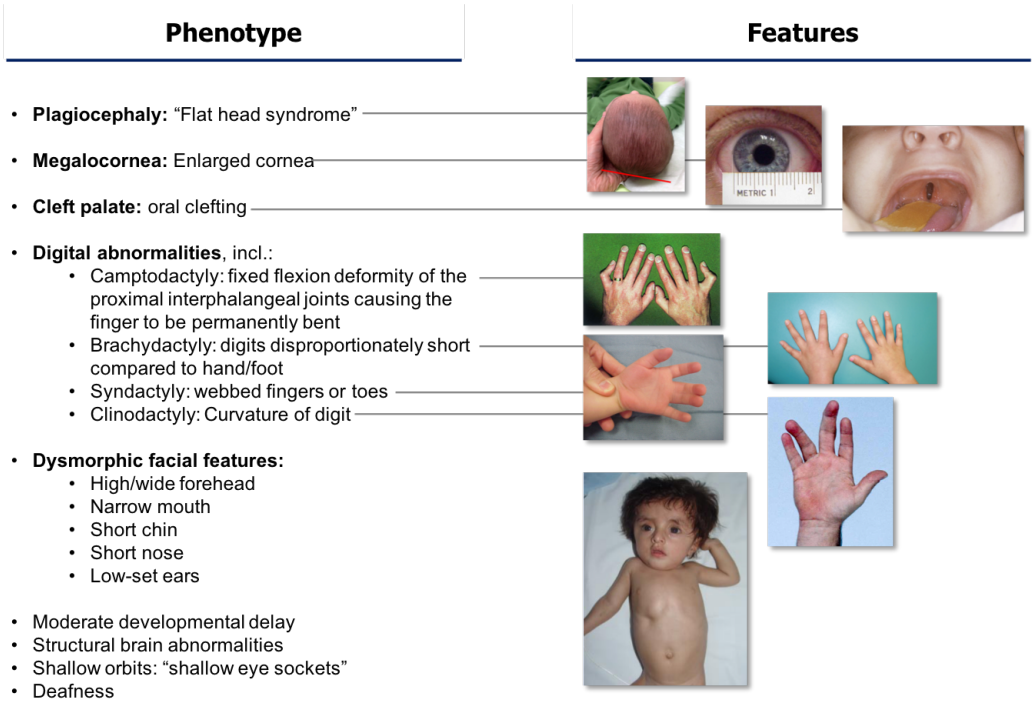


Fig. 1.5 **The Fine-Lubinsky syndrome phenotype.** The FLS phenotype includes plagiocephaly, megalocornea, a cleft palate, and a range of digital abnormalities, including camptodactyly, brachydactyly, syndactyly, and clinodactyly. Furthermore, dysmorphic facial features including a high or wide forehead, a narrow mouth, a short chin, a short nose, and low-set ears are observed. Affected patients additionally display moderate developmental delay, structural brain abnormalities, shallow orbits, and deafness.

been shown to bind over 700 mRNAs [149], including Apolipoprotein-B, Serpin-A1, Fetuin-A, and Clusterin in mouse primary hepatocytes.

Vigilin has a total of 14 KH domains. The KH13 and KH14 domains represent *HDLBP*'s main mRNA-binding interface [149]. As will be discussed in Chapter 6, there is evidence that the candidate splice site variant found in our patient[4] causes in-frame skipping of exon 14, which makes up approximately 55% (41/75 amino acids) of the RNA-binding KH6 domain of the encoded protein vigilin. RBPs play an important role in gene expression and consequently in human disease [150–153]. The variant is hypothesised to affect vigilin's ability to bind RNA, which could be related to the phenotype.

HDLBP has not previously been linked to FLS. However, several reports have described patients affected by '2q37-deletion syndrome' (OMIM:600430) [154] with deletions of the 2q37 locus, which *HDLBP* is located in.

While some phenotypic features associated with 2q37-deletion syndrome overlap with the phenotype of our patients, including brachydactyly, facial dysmorphism, low-set ears and intellectual deficiency, other features were absent in our patients, including seizures, obesity and short stature. Furthermore, patients in our family displayed phenotypic features distinct from reported symptoms of 2q37-deletion syndrome, including plagiocephaly and camptodactyly. Leroy *et al.* [155] describe a cohort of 14 patients with 2q37 deletions, in 13 of which *HDLBP* is one of the deleted genes. In the cohort of 14, morphological dysmorphisms, including brachydactyly type E and facial features were among the most easily recognisable phenotypic features. Only one patient, whose deletion did not include *HDLBP*, did not present with facial dysmorphism.

Furthermore, Felder *et al.* [156] showed that *HDLBP* was considerably down-regulated in lymphoblastoid cell lines of a 2q37-deletion syndrome patient with autism

[4] c.1731+1G>A, NM_203346.4, segregates in two branches of same family

and brachymetaphalangy [156]. The authors note that vigilin is structurally similar to the protein encoded by (*FMR1*), LoF variants in which cause Fragile-X syndrome, a mental retardation condition. Similar to vigilin, *FMR1* contains KH domains, which serve in RNA binding and transport from the nucleus to the cytoplasm.

Vigilin is found in two distinct but similar protein complexes in the cytoplasm and nucleus, called the nuclear and cytoplasmic vigilin-containing complexes (VCCn and VCCc) [149]. Both complexes contain tRNA and tRNA-binding is mediated by vigilin. Furthermore, the two complexes bind exclusively to tRNA, not to other RNA species. Vigilin itself is hypothesised to be involved in tRNA nuclear export [149] and their subsequent delivery to the translation machinery. Cheng *et al.* [149] hypothesise that vigilin could be instrumental for the translation of mRNAs encoding membrane-associated and secreted proteins. Additionally, *HDLBP* is highly expressed in bone and cartilage [156], tissues relevant for the FLS phenotype. Furthermore, as a key regulator in proliferating cells, *HDLBP* plays an important role during embryonic development, suggesting that even moderate down-regulation of expression could have a severe impact.

In addition to that, knock-down mice studies have linked vigilin to cholesterol metabolism and *HDLBP* is linked to obesity in 2q37-deletion syndrome patients [157, 155].

Vigilin also regulates gene expression of the insulin-like growth factor-2 gene (*IGF2*). Vigilin interacts with CCCTC binding factor (*CTCF*) to maintain the imprinting of *IGF2*, under mediation by long noncoding RNA (lncRNA) [158]. In particular, the authors show that KH1-7 domains, i.e. including the KH6 domain, which is spliced out in our patient, functionally interact with zinc-finger domains of *CTCF*. Furthermore, they demonstrate that lncRNA is involved in interaction between vigilin and *CTCF*.

The combined existing literature suggests a potential association of *HDLBP* with FLS due to an impact on the protein's RNA-binding activity. **Therefore, the goal of Chapter 6 is to functionally investigate the potential link between the candidate variant in *HDLBP* in Afia and her relatives with FLS by examining vigilin's RNA-binding activity. *HDLBP* has not previously been investigated as a potentially disease-causing variant in a large human pedigree.**

1.4 Thesis objectives

While the increasing number of rare genetic diseases for which a molecular diagnosis exists makes it possible to shorten the diagnostic odyssey of more and more patients, it is becoming increasingly difficult to decipher the underlying causes of the remaining, yet to be fully characterised conditions. Broadly, the aim of this thesis is therefore to examine and improve upon the status quo in the analysis and functional validation of rare genetic disease variants arising from WGS data.

The main objectives are as follows:

1. To analyse and compare the performance of the GPA Exomiser and Phevor on rare genetic disease cases from the HICF2 study in Chapter 3
2. To develop an algorithm that successfully combines genotypic, phenotypic, and tissue-specific expression data for the identification of candidate variants in novel disease-causing genes in Chapter 4
3. To validate the performance of the newly developed algorithm on the patient cases analysed for objective 1 in Chapter 5
4. To assess the impact of a splice-site variant on the RNA-binding activity of vigilin, the protein encoded by *HDLBP*, in order to functionally characterise the gene as a potential candidate for FLS in Chapter 6

Chapter 2

Materials and methods

In this chapter, I describe materials and methods used throughout the thesis. The chapter is split into a description of the HICF2 project (see Section 2.1), cases from which are used for analyses in Chapters 3, 5, and 6, methods and materials relevant for VPA-based analyses in Chapters 3, 4, and 5 (see Section 2.2), and methods and materials for a functional validation study in Chapter 6 (see Section 2.3). Chapter-specific adaptations and use cases for the methods and materials are described in the respective chapters (see Sections 3.2, 4.2, 5.2, and 6.2).

2.1 The Health Innovation Challenge Fund 2 project

Rare genetic disease patient cases described in Chapters 3, 4, and 5 of this thesis were part of a whole genome sequencing project funded by the Health Innovation Challenge Fund 2 (HICF2) situated at the University of Oxford. In this section, I describe the HICF2 project setup (see Section 2.1.1), the materials and methods used for sequencing (see Section 2.1.2), the HICF2 variant interpretation pipeline (see Section 2.1.3), and the collection of HPO terms for HICF2 cases (see Section 2.1.4). Details on individual patient cases can be found in Chapters 3 and 5 (see Sections 3.2

and 5.2). All work described in Section 2.1, with the exception of the collection of HPO terms described in Section 2.1.4, was conducted prior to my analyses by members of the HICF2 consortium.

2.1.1 The HICF2 project setup

The HICF2 project was governed by a Genomic Medicine Multi-Disciplinary Team (GM-MDT) that verified the eligibility of patients for WGS, approved patients for participation in the HICF2 project, and reviewed case findings. The GM-MDT consisted of members representing different specialties and roles, including clinical scientists, genetic counsellors, non-clinical researchers, physician-scientists, and physicians.

To enrol a patient in the HICF2 study, referring clinicians had to submit an application form summarising clinical information on the patient, including a detailed description of the patient's disease diagnosis and phenotype, as well as a family pedigree.

Once approved for the HICF2 study, patient samples were submitted for whole genome sequencing (see Section 2.1.2 for details) and analysis through the HICF2 consortium (see Section 2.1.3 for details).

All cases presented in this thesis were analysed by senior post-doctoral geneticists in the research group of Professor Jenny Taylor. Candidate variants and supporting evidence identified by the analysts were assessed with the referring clinician and subsequently reviewed by the GM-MDT prior to reporting results back to the patient or their caretakers.

2.1.2 Whole genome sequencing

For whole genome sequencing, 3 μg patient DNA were extracted from whole blood and used to prepare DNA libraries using an Illumina TruSeq DNA PCR-free library preparation kit. The Illumina HiSeq 2500 and HiSeq 4000 systems were used with a 100 bp paired-end read protocol to perform whole genome sequencing to an average depth of 30x.

2.1.3 The HICF2 bioinformatics research pipeline

WGS data from all cases described in this thesis were analysed by senior post-doctoral geneticists using a bioinformatics research pipeline developed for the HICF2 study as described in this section (see Figure 2.1).

WGS FASTQ files were downloaded from Illumina BaseSpace. Reads were merged and mapped to the human reference genome (GRCh37/hs37d5) using the Burrows-Wheeler Alignment tool (BWA, version 0.7.10-r789) [14] and Stampy [15]. PCR duplicates were removed with Picard Tools [159].

B-allele frequencies (BAF) were computed from the aligned BAM files via ngCGH [160] and Nexus Copy Number Software [161] was used for the investigation of CNV and regions of loss of heterozygosity. Shared genomic regions were identified via CODOC [162].

Platypus (version 0.8.1) [13] was used for variant calling (settings see Table 2.1). Singletons were run as single samples on Platypus, while sibling pairs and trios were run as multi-sample files. Two separate tools were used in parallel to identify candidate variants analysed further in Chapters 3 and 5: an annotation and filtering tool called VARAN developed in-house by Dr Niko Popitsch, and the commercial product IVA by Ingenuity [46]. VARAN used the Variant Effect Predictor (VEP, version 69) [19] to

annotate variants. VCFs analysed with IVA were annotated in IVA. After annotation, VCFs were filtered. Variants with a MAF < 1% in reference databases including dbSNP, ExAC, and the 1,000 Genomes Project that were consistent with the suspected inheritance pattern for the particular case were included in the analysis. Remaining variants were interpreted based on the variant's pathogenicity as predicted by several scores, including SIFT [21], CADD, and Polyphen-2 [82], biological pathways, as well as relevant literature and the patient's phenotype. The quality of all candidate variant calls was evaluated with the Integrative Genome Viewer (IGV) version 2.3.32 [163]. Short lists of candidate variants were discussed with the referring physician to either confirm one or more candidate variants as disease-causing based on existing evidence for inclusion in a research report, or to discuss next steps for further assessment of candidate variants, including functional validation. All candidate variants determined to be disease-causing were further validated by Sanger sequencing.

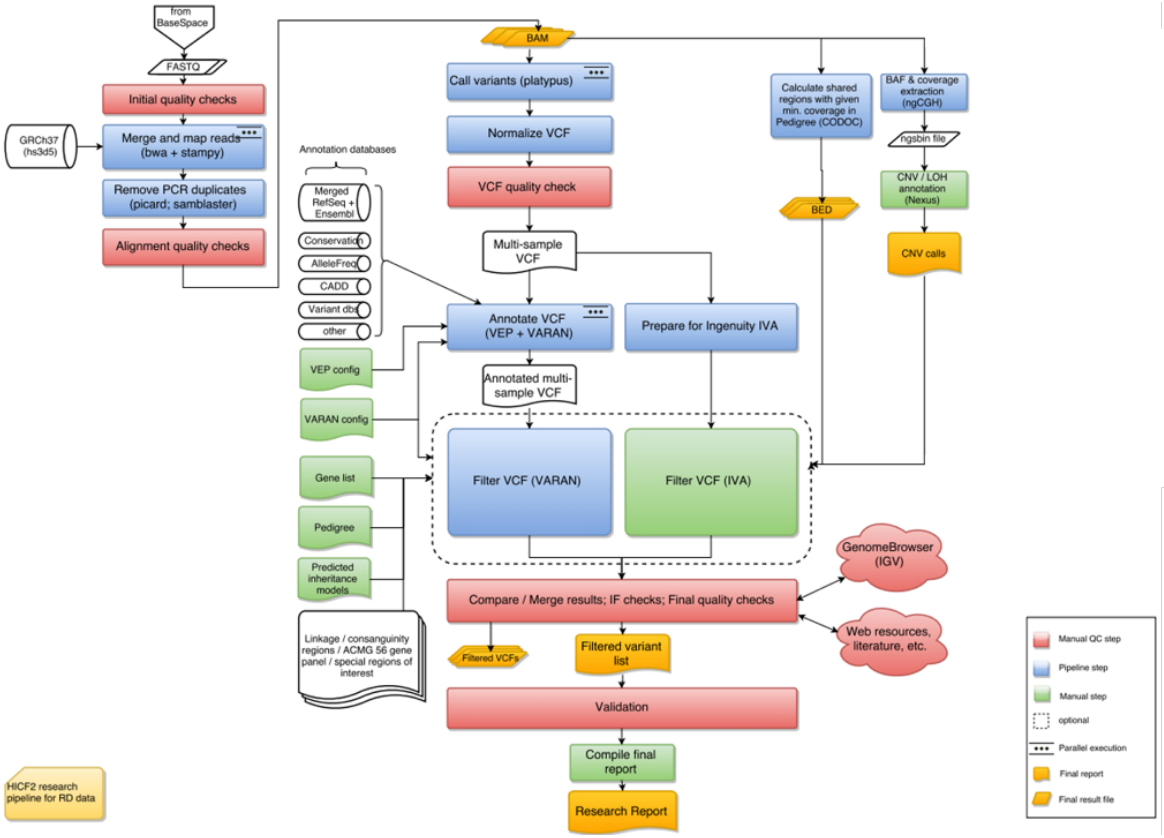


Fig. 2.1 **HICF2 bioinformatics research pipeline.** Benchmark variants for all cases described in this thesis were identified using the bioinformatics pipeline described in this figure. FASTQ files were downloaded from Illumina BaseSpace to merge and map reads with bwa and stampy. PCR duplicates were removed with Picard tools and variants were called using Platypus. BAF were computed from aligned BAM files via ngCGH, and CNV and regions of loss of heterozygosity were investigated using Nexus Copy Number Software. Variants were called using Platypus and VCFs were analysed with the in-house tool VARAN, as well as Ingenuity's IVA. Variants were filtered based on a MAF < 1% in reference databases and each case's suspected inheritance pattern. Variants were subsequently interpreted based on pathogenicity scores such as SIFT, CADD, and Polyphen-2, as well as biological pathway information, relevant literature, and the patient's phenotype. Call qualities were evaluated using IGV. Candidate variants identified by the HICF2 analysts were discussed with the referring physician for reporting to the patient, or to initiate further functional validation. Variant calls were validated by Sanger sequencing (Figure provided by HICF2 consortium member and senior post-doctoral researcher Dr Niko Popitsch).

Option	Setting
assemblyRegionSize	1500
trimReadFlank	0
assembleBadReads	1
minVarDist	9
trimSoftClipped	1
minReads	2
qualBinSize	1
maxHaplotypes	50
filterVarsByCoverage	1
maxSize	1500
originalMaxHaplotypes	50
skipDifficultWindows	0
parseNCBI	0
skipRegionsFile	None
noCycles	0
trimAdapter	1
minPosterior	5
assembleAll	1
trimOverlapping	1
filterDuplicates	1
abThreshold	0.001
minFlank	0
bufferSize	1000000
fileCaching	0
useEMLikelihoods	0
coverageSamplingLevel	30
calculateFlankScore	0
nCPU	8
filterReadsWithUnmappedMates	1
qdThreshold	10
maxVariants	8
scThreshold	0.95
filterReadsWithDistantMates	1
maxReads	5000000
badReadsWindow	11
genIndels	1
largeWindows	0
minMapQual	20
maxVarDist	15
maxGOF	30
rlen	150
minGoodQualBases	20
refCallBlockSize	1000
countOnlyExactIndelMatches	0
longHaps	0
HLATyping	0
filterReadPairsWithSmallInserts	1
minBaseQual	20
getVariantsFromBAMs	1
genSNPs	1
assemble	0
assemblerKmerSize	15
minVarFreq	0.05
verbosity	2
compressReads	0
rmsmqThreshold	40
filteredReadsFrac	0.7
outputRefCalls	0
badReadsThreshold	15
hapScoreThreshold	4
regions	None
sbThreshold	0.001
assembleBrokenPairs	0
mergeClusteredVariants	1
maxGenotypes	1275
nInd	3

Table 2.1 **Platypus settings.** This table shows all settings used to run Platypus for variant calling on WGS data in this thesis.

2.1.4 HPO term collection

HPO terms for every case discussed in this thesis were collected using the software Phenotips (version 1.2.3) [104]. For each case, I generated an initial list of HPO terms based on the clinical information submitted by the referring clinician through the HICF2 application form (see Section 2.1.1). The HPO profiles were refined by the referring clinician in in-person case review meetings.

2.2 Variant prioritisation algorithms

In Chapters 3, 4, and 5, VPA are used to evaluate candidate variants previously identified as part of the HICF2 project. In this section, settings for VPA used across chapters are described. For details on the chapter-specific utilisation of each method, see Sections 3.2, 4.2, and 5.2).

Exomiser version 7.2.1 and version 11.0.0 were used for different analyses. Depending on the version, Exomiser draws allele frequencies from different databases. For Exomiser version 11.0.0, the following allele frequency databases were used: the 1,000 Genomes Project [49], the Exome Sequencing Project [91], ExAC [73], gnomAD [72], the TOPMed database [92], and the UK10K project [93]. For Exomiser version 7.2.1, the 1,000 Genomes Project, the Exome Sequencing Project, and ExAC were used. Variants were filtered to have a call quality of $\geq$5.0 and a maximum MAF of $\leq$1%. Exomiser's phenotype score was calculated with Exomiser's hiPHIVE algorithm setting. Pedigree (PED) files were used to pass pedigree information to Exomiser. Table 2.2 shows the structure of a PED file. Exomiser's inheritance filter was used to filter variants based on the suspected inheritance pattern of each case. Exomiser 7.2.1 differentiates between 'autosomal_dominant', which filters for heterozygous variants and thus produces a list of variants following autosomal dominant or *de novo*

inheritance, ‘autosomal_recessive’, which includes autosomal recessive and compound heterozygous inheritance, and ‘X_recessive’ for X-linked recessive inheritance. The inheritance pattern filter of Exomiser 11.0.0 is more granular, also differentiating between ‘autosomal_recessive_hom_alt’ inheritance for autosomal recessive homozygous variants, ‘autosomal_recessive_comp_het’ for compound heterozygous variants, ‘x_dominant’, ‘x_recessive_hom_alt’, and ‘x_recessive_comp_het’ for X-linked dominant, recessive homozygous, and compound heterozygous inheritance respectively. Settings were passed to Exomiser using yml files. Table 2.3 and Table 2.4 show the yml file settings for all analyses using Exomiser version 11.0.0 and 7.2.1 respectively.

Family ID	Individual ID	Paternal ID	Maternal ID	Sex	Phenotype
F1	proband_1	proband_1_father	proband_1_mother	1	2
F1	proband_1_father	0	0	1	1
F1	proband_1_mother	0	0	2	1

Table 2.2 **PED file structure.** This table shows the structure of an example PED file (F1.ped) used as input for Exomiser’s inheritance pattern filter. PED files consist of the following six columns: family ID, individual ID, paternal ID, maternal ID, sex, and phenotype. Members of the same family are assigned the same family ID, in this case F1. Each individual is assigned an individual ID referring to the sample columns used in the input multisample VCF. In this example, three members of the same family are analysed: the proband (proband_1), the proband’s father (proband_1_father), and the proband’s mother (proband_1_mother). Each individual is represented by a line entry in the PED file. The family relationship is indicated by providing the ID of each individual’s father and mother in the paternal and maternal ID columns (0=unknown). The individual’s sex is supplied in the sex column (1=male, 2=female, 0=unknown) and the affected status is provided in the phenotype column (1=unaffected, 2=affected) (adapted from [89]).

Option	Setting
vcf	path to input vcf file
ped	path to ped file
modeOfInheritance	depending on the suspected inheritance pattern of each case: AUTOSOMAL_DOMINANT, AUTOSOMAL_RECESSIVE, or X_RECESSIVE
analysisMode	PASS_ONLY
geneScoreMode	RAW_SCORE
hpoIds	comma-separated list of the proband's HPO terms
frequencySources	THOUSAND_GENOMES, ESP_AFRICAN_AMERICAN, ESP_EUROPEAN_AMERICAN, ESP_ALL, EXAC_AFRICAN_INC_AFRICAN_AMERICAN, EXAC_AMERICAN, EXAC_SOUTH_ASIAN, EXAC_EAST_ASIAN, EXAC_FINNISH, EXAC_NON_FINNISH_EUROPEAN, EXAC_OTHER
pathogenicitySources	POLYPHEN, MUTATION_TASTER, SIFT
qualityFilter	minQuality: 5.0
variantEffectFilter	remove: UPSTREAM_GENE_VARIANT, INTERGENIC_VARIANT, REGULATORY_REGION_VARIANT, CODING_TRANSCRIPT_INTRON_VARIANT, NON_CODING_TRANSCRIPT_INTRON_VARIANT, DOWNSTREAM_GENE_VARIANT
frequencyFilter	maxFrequency: 1.0
pathogenicityFilter	keepNonPathogenic: true
inheritanceFilter	{}
hiPhivePrioritiser	{}
outputPassVariantsOnly	false
numGenes	0
outputPrefix	prefix for output file
outputFormats	TSV-GENE, TSV-VARIANT, VCF, HTML

Table 2.3 **Exomiser version 7.2.1 yml settings.** This table shows all settings used to run Exomiser version 7.2.1 via a yml input file.

Option	Setting
genomeAssembly	hg19
vcf	path to input vcf file
ped	path to ped file
proband	name of proband's ID used in the vcf
hpoIds	comma-separated list of the proband's HPO terms
inheritanceModes	depending on the suspected inheritance pattern of each case: AUTOSOMAL_DOMINANT: 1.0, AUTOSOMAL_RECESSIVE_HOM_ALT: 1.0, AUTOSOMAL_RECESSIVE_COMP_HET: 1.0, X_DOMINANT: 1.0, X_RECESSIVE_HOM_ALT: 1.0, X_RECESSIVE_COMP_HET: 1.0
analysisMode	PASS_ONLY
frequencySources	THOUSAND_GENOMES, TOPMED, UK10K, ESP_AFRICAN_AMERICAN, ESP_EUROPEAN_AMERICAN, ESP_ALL, EXAC_AFRICAN_INC_AFRICAN_AMERICAN, EXAC_AMERICAN, EXAC_SOUTH_ASIAN, EXAC_EAST_ASIAN, EXAC_FINNISH, EXAC_NON_FINNISH_EUROPEAN, EXAC_OTHER, GNOMAD_E_AFR, GNOMAD_E_AMR, GNOMAD_E_EAS, GNOMAD_E_FIN, GNOMAD_E_NFE, GNOMAD_E_OTH, GNOMAD_E_SAS, GNOMAD_G_AFR, GNOMAD_G_AMR, GNOMAD_G_EAS, GNOMAD_G_FIN, GNOMAD_G_NFE, GNOMAD_G_OTH, GNOMAD_G_SAS
pathogenicitySources	POLYPHEN, MUTATION_TASTER, SIFT
qualityFilter	minQuality: 5.0
variantEffectFilter	FIVE_PRIME_UTR_EXON_VARIANT, FIVE_PRIME_UTR_INTRON_VARIANT, THREE_PRIME_UTR_EXON_VARIANT, THREE_PRIME_UTR_INTRON_VARIANT, NON_CODING_TRANSCRIPT_EXON_VARIANT, UPSTREAM_GENE_VARIANT, INTERGENIC_VARIANT, REGULATORY_REGION_VARIANT, CODING_TRANSCRIPT_INTRON_VARIANT, NON_CODING_TRANSCRIPT_INTRON_VARIANT, DOWNSTREAM_GENE_VARIANT
frequencyFilter	maxFrequency: 1.0
pathogenicityFilter	keepNonPathogenic: true
inheritanceFilter	{}
hiPhivePrioritiser	{}
outputContributingVariantsOnly	false
numGenes	0
outputPrefix	prefix for output file
outputFormats	HTML, JSON, TSV_GENE, TSV_VARIANT, VCF

Table 2.4 **Exomiser version 11.0.0 yml settings.** This table shows all settings used to run Exomiser version 11.0.0 via a yml input file.

VAAST 2.0 and Phevor were executed using Omicia Opal version 4.24.0 [164]. For a detailed description of VAAST 2.0 and Phevor settings, see Chapter 3.

2.3 Functional validation

In this section, materials and methods used to functionally characterise *HDLBP* as a candidate gene for FLS are described (see Chapter 6). An analysis of one of the rare disease cases described in Chapter 3 identified a splice-site variant in *HDLBP* as a potential cause of FLS. The splice-site variant causes in-frame skipping of exon 14, which is part of the RNA-binding KH6 domain of the encoded protein vigilin. Reduced RNA-binding activity due to exon skipping was assessed as a potential cause of FLS.

Stable cell lines were established to compare the effect of the variant to the wildtype, subsequently referred to as '*HDLBP* mutant' and '*HDLBP* wildtype' (see Section 2.3.1) in various assays described hereafter. Quality controls were conducted to assess if the stable cell lines were established successfully (see Section 2.3.2). Western Blots of vigilin were performed to examine protein instability as a cause of the phenotype (see Section 2.3.3) and time-dependant protein decay was analysed (see Section 2.3.4). Potential differences in intra-cellular protein localisation between wildtype and mutant as a cause of the phenotype were assessed via fluorescence microscopy (see Section 2.3.5). The tertiary structure of vigilin's KH6 domain was simulated to assess the potential impact of the splice-site variant on RNA binding (see Section 2.3.6). Finally, the RNA-binding activity of wildtype and mutant vigilin was assessed using an Oligo(DT) capture method (see Section 2.3.7).

2.3.1 Establishment of stable cell lines for the *HDLBP* wildtype and mutant

Stable cell lines for the *HDLBP* wildtype and mutant had to be established to assess the impact of the *HDLBP* candidate variant on the RNA-binding activity of vigilin. cDNA fragments for *HDLBP* wildtype and *HDLBP* mutant (see Figure 6.5 for details, the full sequences are included in the Appendix) were ordered from ThermoFisher GeneArt [165] and were delivered lyophilised in a pMK-RQ vector backbone. The following steps were conducted in parallel for *HDLBP* wildtype and mutant. Lyophilised plasmids were resuspended in distilled water at 100 ng/l. For plasmid amplification, 1 μl of plasmid was mixed with 10 μl of KCM buffer (500 mM KCl, 150 mM $CaCl_2$, 250 mM $MgCl_2$) in a 50 μl reaction and incubated with 50 μl chemically competent DH5α *E. coli* cells (Invitrogen, Cat. No. 12297-016) for 20 min on ice, followed by 10 min at room temperature. Bacteria were then cultured in 900 μl of super optimal broth with catabolite repression medium (Sigma-Aldrich, Cat. No. S1797) at 37°C for one hour and plated in corresponding antibiotic-enriched LB-agar plates (ThermoFisher, Cat. No. 22700041) for overnight incubation at 37°C. Single colonies were grown overnight in LB medium (ThermoFisher, Cat. No. 10855001). Plasmids were then purified using the QIAprep Spin Miniprep Kit (Qiagen, Cat. No. 27106). Subsequently, 1 μg of plasmid was digested with XhoI (New England Biolabs, Cat. No. R0146M) and KpnI (New England Biolabs, Cat. No. R0142M) to excise the *HDLBP* gene insert. Digested products were run on a 1% agarose gel and the *HDLBP* gene insert was isolated using the QIAquick Gel Extraction Kit (Qiagen, Cat. No. 28706) following the manufacturer's instructions. In parallel, 1 μg of the expression plasmid, pcDNA5-FRT-TO-precision-linker-eGFP (see Figure 6.6 and Figure 6.7 for detailed plasmid maps), was digested with XhoI and KpnI and gel-purified as above. 25 ng *HDLBP* gene insert were cloned into 25 ng of the digested expression plasmid using Quick-Stick Ligase (Bioline, Cat. No. BIO-27028) and following manufacturer's

instructions. *HDLBP* gene insert-containing expression vectors were amplified by transforming 2 μl into chemically competent DH5α cells as described above. The *HDLBP*-containing plasmids were extracted using the Qiagen Plasmid Midi Kit (Qiagen, Cat. No. 12143) following the provided protocol. All constructs were then verified by Sanger sequencing (GATC Biotech services) and restriction digestion with the enzymes XhoI and KpnI. The following primers (Eurofins Genomics [166]) were used for sequencing:

- Primers for *HDLBP* wildtype:
  - **Primer 1:** 5'-CCACGCTGTTTTGACCTCCA-3'
  - **Primer 2:** 5'-GTATAATAGACTGGTTGGCG-3'
  - **Primer 3:** 5'-AGCGTACCAAGGATCTAATC-3'
  - **Primer 4:** 5'-ATCACCATCATTGGAAAGGA-3'
  - **Primer 5:** 5'-GCTACAGGCCGAGCAGGAGG-3'
- Primers for *HDLBP* mutant:
  - **Primer 1:** 5'-CCACGCTGTTTTGACCTCCA-3'
  - **Primer 2:** 5'-GTATAATAGACTGGTTGGCG-3'
  - **Primer 3:** 5'-AGCGTACCAAGGATCTAATC-3'
  - **Primer 4:** 5'-CACCGCCACTTCGTCATCCG-3'
  - **Primer 5:** 5'-CCGGTTGGAGCATGACGTGA-3'

The cDNA-containing plasmids were transfected into HeLa and HEK293 cells using X-tremeGENE 9 Transfection Reagent (Sigma-Aldrich, Cat. No. 06365779001) with the provided protocol. Transfected cells were cultured in T150 flasks (Sigma-Aldrich, Cat. No. CLS430825) in DMEM medium (Gibco, Cat. No. 11995-065) containing Hygromycin B (TOKU-E, Cat. No. 31282-04-9), 10% Fetile Bovine Serum, and 1% PenStrep for one week.

2.3.2 Polymerase chain reaction

Transfected, confluent HEK293 cells were harvested using trypsin (Sigma- Aldrich, Cat. No. T3924) and the *HDLBP* mutant and wildtype cDNA-containing plasmids were extracted using the Qiagen Plasmid Midi Kit (Qiagen, Cat. No. 12143) using the provided protocol. All PCR work was done by Dr Pamela Kaisaki, a senior post-doctoral geneticist in the Taylor Group. PCR of mutant and wildtype *HDLBP* was performed with the FastStart Taq DNA polymerase kit (Roche, Cat. No. 12 032 902 001) using custom primers ordered from Eurofins Genomics (Eurofins Genomics [166]) that amplify from exon twelve to exon 16. The following primers were used:

- **Primer 1, *HDLBP*-12 forward:** 5'-AATTTGATCCGCATCGAGGG-3'
- **Primer 2, *HDLBP*-16 reverse:** 5'-TGCTGGAAGGTCGATTTTGG-3'

PCR was performed using a 96X Universal peqSTAR thermal cycler (VWR, Cat. No. 732-2887), starting with a denaturation step at 95°C for two minutes, followed by 35 amplification cycles of 30 seconds at 95°C, 30 seconds at 55°C, and 30 seconds at 72°C, followed by six minutes at 72°C.

2.3.3 Western blot

Transfected, confluent HeLa and HEK293 cells were harvested using trypsin (Sigma-Aldrich, Cat. No. T3924), expression was induced with 1 µg/mol doxycycline and the cells were grown in petri dishes overnight. Next, the cells were harvested and resuspended in RIPA lysis buffer (50 mM Tris ph 7.5, 150mM NaCl, 1% (vol/vol) Triton-X100, 0.1% SDS, 0.5% Na (wt/vol) deoxychol, 1xAEBSF, MilliQ) and spun down. 5 mmol DTT and loading buffer (NuPage Invitrogen, Cat. No. 1771478) were added to the supernatant and the solution was boiled [five minutes, 95°C]. Next, the samples were run on a gel [50 minutes, 180V] with Precision Plus Protein Standards

dual color (250kD) (Bio-Rad, Cat. No. 161-0394) and ColorPlus Prestained Protein Ladder, Broad Range (230kD) (New England BioLabs, Cat. No. P7711S). Thereafter, the traces were transferred onto a nitrocellulose membrane and developed in two steps, with Rat monoclonal [3H9] to GFP (Chromotek) as the primary and IRDye 800CW Goat Anti-Rat as the secondary antibody. Images were acquired on an Odyssey Li-Cor.

2.3.4 Protein decay

To assess if protein decay over time differs between the vigilin wild type and mutant, the respective transfected HEK293 cell lines were grown to 50% confluence in T75 flasks (Sigma-Aldrich, Cat. No. CLS3290). Six repeats were created for the vigilin wildtpye, five for the mutant. The cells were harvested using 2 ml trypsin and 8 ml DMEM media per flask. For each repeat, 200 μl of each solution were added to a well in a black 96 well plate with a transparent bottom (Bio-Rad, Cat. No. HSP9666). Expression was induced with 0.4 µg/mol doxycycline per well overnight. The next morning, all media was removed, the cells were washed once with PBS, 200 μl media were added, and the GFP measurement was started in a plate reader using 475 nm and 509 nm as the excitation and emission wavelengths for GFP. Measurements were taken every 15 minutes over 48h for a total of 192 time points. The time-dependant decay of wild type and mutant vigilin is inferred by measuring the intensity of GFP linked to the vigilin wild type and mutant in the vigilin-GFP fusion proteins over time.

2.3.5 Fluorescence microscopy for intra-cellular protein localisation

Transfected HeLa cells were grown on High Precision Coverslips (Marienfeld, Cat. No. 0107052) and expression was induced with 1 µg/mol doxycycline overnight for coverslips used for fluorescence analysis, while expression was not induced for

controls. The cells were fixed in PBS +4% formaldehyde and permeabilised in PBS + 0.1% TX100. After several PBST washing steps, the cells were blocked in PBST+2% BSA and incubated with the secondary antibody in PBST + 2% BSA in the dark. The coverslips were washed in PBST + 2% BSA and incubated with DAPI in PBS1X. After a few washing steps, they were treated with Vectashield medium (Vector Laboratories, Cat. No. H-1000), dried and the borders were sealed with nail polish. Next, the images were acquired using an API DeltaVision Elite widefield fluorescence microscope with an Olympus PlanApo 100x/1.40 IX70 oil objective (1.60 auxiliary magnification) and deconvolved with the Resolve3D module of softWoRx-Acquire version 4.1.2, release 1 (GE Healthcare; image properties: XY dimensions: 512 x 512, ZWT dimensions (expected): 11 x 2 x 1, pixel size: 0.09669 0.09669 0.150, binning: 1x1, flat-field calibration: off).

2.3.6 Tertiary structure simulation of vigilin

The tertiary structure of vigilin's KH6 domain was simulated by Dr Matteo Ferla, a senior post-doctoral bioinformatician in the Taylor Group, to assess the impact of a splice-site variant causing skipping of exon 14 on the RNA-binding KH6 domain. To visualise vigilin's KH6 domain, the pregenerated threaded model protein structure of human vigilin's KH6 domain was downloaded from SWISS-Model [167]. The initial RNA position was inferred based on the third KH domain of KH-type splicing regulatory protein (PDB:4B8T) [168] and further refined by finding the lowest energy conformation with Rosetta Relax [169]. The figure was generated with the software package PyMOL (version 2.0, student licence) [170]. The default settings of the software were modified with 'set ray_trace_mode, 3' and 'bg_color white'.

2.3.7 Oligo(dT) capture

The polyadenylated RNA-binding activity of the vigilin-GFP fusion protein was measured using the protocol described below. The following solutions are referenced throughout the protocol:

- **Lysis buffer:** 20 mM Tris-HCl (pH 7.5) (Sigma-Aldrich, CAS Number 1185-53-1), 500 mM LiCl (Sigma-Aldrich, CAS Number 7447-41-8), 0.5% LiDS (wt/vol, stock 10%) (Sigma-Aldrich, CAS Number 2044-56-6), 1 mM EDTA (Sigma-Aldrich, CAS Number 60-00-4), 0.1% IGEPAL (NP40) (Sigma-Aldrich, CAS Number 9002-93-1), and 5 mM DTT (Sigma-Aldrich, CAS Number 3483-12-3).
- **Buffer 1:** 20 mM Tris-HCl (pH 7.5), 500 mM LiCl, 0.1% LiDS (wt/vol), 1 mM EDTA, 0.1% IGEPAL (NP40), and 5 mM DTT.
- **Buffer 2:** 20 mM Tris-HCl (pH 7.5), 500 mM LiCl, 1 mM EDTA, 0.1% IGEPAL (NP40), and 5 mM DTT.
- **Buffer 3:** 20 mM Tris-HCl (pH 7.5), 200 mM LiCl, 1 mM EDTA, and 5 mM DTT.
- **Elution buffer:** 20 mM Tris-HCl (pH 7.5) and 1 mM EDTA.

The following cell lines were used:

- **Cell lines to be assessed:** HEK293 cells with the eGFP-tagged *HDLBP* mutant and wildtype, respectively
- **Positive control:** a HEK293 cell line transfected with a plasmid containing eGFP-tagged Heterogeneous Nuclear Ribonucleoprotein Q (*hnRNPQ*), a known RNA-binding protein

- **Negative control:** a HEK293 cell line transfected with a GFP-containing plasmid, which does not bind RNA

The cell lines, established as described in Section 2.3.1, were cultured in two 15 cm dishes per cell line to reach 50% confluence. eGFP expression was induced by adding 1 $\mu g/mol$ doxycycline and the cells were incubated overnight.

The media was removed from the 15 cm dishes, cells were washed with 10 ml of PBS and the PBS was subsequently discarded.

The dishes were set on ice, the lids were removed, and UV light at 254 nm and 0.15 Jcm^{-2} was applied to cross-link the polyadenylated mRNA with the eGFP-tagged fusion protein for all conditions. 0.9 ml lysis buffer, kept at 37°C, were applied per dish. The cellular material was collected with a rubber scraper in a 1.5 ml Eppendorf tube per condition. The viscous mixture was homogenised by passing it through a 5 ml syringe with a 27G needle three times.

300 μl lysate for each condition, hereafter referred to as the 'input', were stored on ice to be used for normalisation during the fluorescence measurement at the end of the protocol.

Per condition, 300 μl of the homogenised lysate solution were added to a tube and 300 μl oligo$(dT)_{25}$ magnetic beads (New England Biolabs, Cat. No. S1419S), washed previously in lysis buffer, were added to the lysate. The mixture was incubated for 1h in a cold room whilst gently rotating the tubes.

The tubes were put on ice and a magnet was used to separate the beads, now linked to the polyadenylated mRNA and therefore the eGFP-tagged protein, and the lysate. The lysate was kept on ice to be used afterwards.

1.8 ml of lysis buffer were added to the tubes containing the beads and the mixture was incubated on ice for five minutes, inverting the tube every minute. The beads were collected with the magnet and the supernatant was discarded.

1.8 ml of buffer 1 was added to the beads and the solution was incubated on ice for five minutes, inverting the tubes every minute. The beads were collected and the supernatant was discarded.

1.8 ml of buffer 2 was added, the tubes were inverted 10 times, the beads were collected using the magnet, and the supernatant was discarded.

Next, 1.8 ml of buffer 3 were added, the tubes were inverted 10 times, the beads were collected using the magnet, and the supernatant was discarded.

200 μl of the elution buffer were added to the beads, and the mixture was incubated for three minutes at 55°C. The beads were separated from the supernatant using a magnet and the supernatant (later referred to as eluate #1) was stored on ice. RNA concentrations of the supernatant were measured using a nanodrop (Thermo Scientific - NanoDrop 1000).

The beads were recycled using 400 μl of 0.1M NaOH and subsequently incubated at 55°C for five minutes. Next, the beads were removed from the NaOH using the magnet.

Thereafter, the beads were washed three times in lysis buffer. After removing the lysis buffer, the beads were resuspended in the lysate supernatant that was previously kept on ice. The mixture was rotated for 1h.

All steps described above were repeated one more time and the resulting eluate #2 was pooled with eluate #1, reaching a total volume of 800 μl. For storage longer than one day, samples are stored in a -80°C freezer.

Fluorescence measurement Fluorescence was measured in a microplate reader with a black 96 well plate with a transparent bottom (Bio-Rad, Cat. No. HSP9666), using 475 nm and 509 nm as the excitation and emission wavelengths for eGFP.

Measurements were conducted for the three proteins in question: *HDLBP* wildtype and mutant, as well as the GFP control. The measured eGFP intensity serves as a

proxy for the respective protein's RNA binding activity. For each protein p, the RNA binding activity relative to GFP $b_{RNA}(p)$ was determined as follows:

$$b_{RNA}(p) = \frac{100 \sum_{i=1}^{n} \frac{I_{(p)eluate,i}}{I_{(p)input,i}}}{n I_{GFP,eluate}} \tag{2.1}$$

.

n is the number of replicates, $I_{(p)eluate,i}$ and $I_{(p)input,i}$ represent the measured GFP intensity of the eluate and input respectively for repeat i, and $I_{GFP,eluate}$ is the measured GFP intensity of the GFP control's eluate. In total, two biological and three technical replicates each were created. The higher $b_{RNA}(p)$ is, the better the RNA-binding activity of the respective protein.

Chapter 3

Prioritisation of variants from whole genome sequencing data using genotypic and phenotypic information – a comparison

3.1 Introduction

VPA are an increasingly widely used approach for clinical WGS analysis of rare disease patients. VPA based on allele frequency, phylogenetic conservation, protein impact and variant type, such as SIFT, CADD, MutationTaster, PolyPhen and VAAST (see Section 1.2.1.3.1) are widely used to facilitate variant analyses. However, given the approximately 100 biologically relevant LoF alleles in an average human genome [95], identifying the one disease-causing variant for a monogenic rare disease is challenging. For that reason, methods extending the mentioned approaches were introduced by including phenotypic information in the ranking process. Two of the most widely used VPA using allele frequency, conservation and phenotypic information are the

Exomiser and VAAST+Phevor frameworks (see Chapter 1, Section 1.2.2.3 for an in-depth algorithm description). Both algorithms were used by teams analysing rare genetic disease data from the 100,000 Genomes Project, the largest WGS RD cohort that exists to date [171, 99]. Exomiser and Phevor each rely on two sets of scores, one based on allele frequency and conservation and one based on phenotypic information, which are combined to produce a final ranking score. The individual scores of each of the two algorithms, however, are calculated differently. Exomiser's hiPHIVE algorithm combines a minor-allele frequency score and a pathogenicity score into the Exomiser variant score, which in turn is combined with the Exomiser phenotype score, based on human, mouse, zebra fish and PPI data, through a logistic regression, producing the Exomiser combined score for each variant. Phevor, on the other hand, reranks input from tools such as VAAST, by using phenotypic information with a cross propagation approach that links data from multiple ontologies through candidate genes.

Deciding which algorithm to use is challenging for molecular genetics diagnostic laboratories and academic research groups. Molecular diagnoses only exist for approximately half of the more than 7,000 rare diseases [172]. Thus, analysts require tools to identify disease-causing variants in both known and novel genes. To determine which algorithms are best suited for this challenge, algorithm comparisons using both in silico and real patient WGS data have been conducted.

Pengelly *et al.* [173] conducted a comparison of four algorithms based on 21 RD cases: eXtasy, the Online Variant Analysis (OVA) tool [174], Exomiser with CADD, Exomiser and PhenIX. Benchmark variants, all of which lie in known genes for the disease phenotype, were identified by clinical genetics experts prior to their benchmark analysis. In their analysis, the PhenIX algorithm performed best, with Exomiser's hiPHIVE as a close second. PhenIX is part of the Exomiser framework and based on an algorithm similar to hiPHIVE, but PhenIX uses a different approach to calculate the phenotype score. In contrast to hiPHIVE, PhenIX only uses human

phenotypic data and not mouse or zebra fish phenotypes or PPI data, since it is optimised for the identification of disease-causing variants in genes known for the human phenotype. hiPHIVE, also drawing on mouse and zebrafish phenotypic data, as well as PPI networks, is designed for known and novel disease gene discovery. In this benchmark analysis, PhenIX outperformed hiPHIVE, perhaps unsurprisingly, since the gene candidates in all 21 cases are known, not novel. The authors state that they also included Phevor in the comparison, but no results for Phevor are presented.

Requena *et al.* [175] analysed the concordance of the results produced by their in-house-developed Pathogenic Variant (PAVAR) score, VAAST, VAAST combined with Phevor, Exomiser version 2, CADD, and FATHMM. The authors do not specify which of the PhenIX, PHIVE and hiPHIVE algorithms in the Exomiser framework were used, making a comparison of their results with existing literature challenging. For the concordance analysis, they used four trios and one case consisting of the proband and only one parent with familial Meniere's disease (FMD). The usefulness of concordance as a metric to evaluate algorithm performance, however, is limited. Concordance is not directly related to the accuracy of each algorithm, but instead merely highlights how similar the results of each algorithm are to each other, regardless of whether or not they are correct. Furthermore, the five cases presented by the authors all carry variants in known FMD genes, thus not providing insight into the comparative performance of the used algorithms for novel gene discovery.

In 2014, Javed *et al.* [176] published Phen-Gen, a variant prioritisation algorithm similar to Exomiser and VAAST+Phevor, which uses sequencing data and phenotypic data as inputs. Included in the publication is a benchmark analysis comparing Phen-Gen with eXtasy, VAAST and Phevor on simulated patient data, in which Phen-Gen outperforms the other algorithms. The Exomiser framework, however, is not included in the analysis.

Furthermore, Smedley *et al.* [58], the creators of Exomiser, and Singleton *et al.* [113], the creators of VAAST+Phevor, conducted their own comparisons. In the Smedley *et al.* paper, Exomiser's hiPHIVE algorithm outperforms PhenIX, Phevor, Phen-Gen and eXtasy, while VAAST+Phevor achieves better results than Exomiser's hiPHIVE in Singleton *et al.*'s study. Phen-Gen and PhenIX had not yet been published at the time of Singleton *et al.*'s study and eXtasy was not included in the analysis.

To provide an independent point of view of the performance of the two main analysis frameworks originally used by the 100,000 Genomes Project, Exomiser's hipHIVE and VAAST+Phevor, for known and novel disease gene discovery alike, I conducted an in-depth comparison of the algorithms. The performance of the algorithms is showcased using eleven rare genetic disease patient cases from the HICF2 cohort. Furthermore, the analysis serves to assess the advantages of GPAs compared to GAs. **The goal of this chapter is thus to compare Exomiser's hiPHIVE and VAAST+Phevor with each other to produce a novel independent source validating their performance, whilst also examining the usefulness of GPAs compared to GAs.**

3.2 Materials and Methods

3.2.1 Patient cases

Eleven patient cases from the HICF2 study were included in the analysis. The cases were selected to include different pedigree structures (singletons and trios), different inheritance patterns (autosomal/*de novo* dominant, autosomal recessive and X-linked recessive) and likely disease-causing variants in known and, at the time of analysis, novel gene candidates for the specific diseases. Phenotypic data was described using the HPO and collected as discussed in Chapter 2. Candidate variants, identified

manually by senior post-doctoral geneticists with our clinical analysis pipeline (see Section 2.1.3), were used as benchmarks to test the algorithms' effectiveness in prioritising likely disease-causing variants from WGS data. Table 3.1 shows an overview of all patient cases. At the time of the analysis, data from the 100,000 Genomes Project was not yet available, which could have otherwise been used for the benchmark comparison.

	Diagnosis	Gene	Variant	# samples	Inheritance	HPO terms	Known/novel
1	Distal arthrogryposis	*TNNI2*	c.466C>T [NM_001145829.1], p.Arg156*	Singleton	AD or *de novo*	Micrognathia, downslanted palpebral fissures, high palate, malar flattening, limited shoulder movement, hip dislocation, ulnar deviation of finger, distal arthrogryposis, tapered finger, flexion contracture of finger, abnormality of the hand, talipes	known
2	Bilateral hippocampal sclerosis	*CACNA1E*	c.5702G>A [NM_001205293.1], p.Arg1901His	Singleton	AD or *de novo*	Seizures, dysgenesis of the hippocampus	known
3	Severe epileptic encephalopathy	*WWOX*	c.705dupG [NM_016373.2], p.His236AlafsTer34	Singleton	AR	Coarse facial features, deep palmar crease, atrial septal defect, hypertonia, profound global developmental delay, epileptic encephalopathy, hypsarrhythmia, infantile spasms, abnormal hand morphology	known
4	Dilated cardiomyopathy	*ACTC1*	c.664G>A [NM_005159.4], p.Ala222Thr	Trio	*de novo*	Weight for age (decreased body weight (<-2SD)), endocardial fibroelastosis, dilated cardiomyopathy, cardiomegaly, respiratory tract infection	known
5	Majeed syndrome	*PSTPIP1*	c.748G>A [NM_003978.4], p.Glu250Lys	Trio	*de novo*	Recurrent skin infections, bone marrow hypocellularity, episodic fever, splenomegaly, recurrent infections	known
6	Undefined immunodysregulatory disorder	*SAMD9L*	c.3353A>G [NM_152703.2], p.Tyr1118Cys	Trio	*de novo*	Nystagmus, psoriasis, respiratory tract infection, arthropathy, colitis, abnormality of the intestine, clumsiness, increased CSF protein, abnormality of the cerebral white matter, gait ataxia, cerebellar atrophy, thrombocytopenia, decreased antibody level in blood	known
7	Fine-Lubinsky syndrome	*POR*	c.1493G>C [NM_000941.2], p.Arg498Pro	Trio	AR	Cleft palate, Narrow mouth, micrognathia, short chin, shallow orbits, agenesis of permanent teeth, plagiocephaly, megalocornea, preauricular skin tag, cupped ear, arthrogryposis multiplex congenita, hypospadias, renal agenesis, moderate global developmental delay, talipes	known*
8	Congenital erythrocytosis	*SLC30A10*	c.823T>A [NM_018713.2], p.Trp275Arg	Trio	AR	Abnormality of the cardiovascular system, cerebral hemorrhage, hypotension, hemangioma, varicose veins, stroke, abnormality of blood and blood-forming tissues, increased hemoglobin, peripheral thrombosis, increased hematocrit, increased red blood cell mass, abnormality of the nervous system, headache, abnormality of the integument, plethora, neoplasm, constitutional symptom	known
9	Atypical Klippel-Trenaunay syndrome	*RBPJ*	c.535T>G [NM_005349], p.Leu179Val	Trio	*de novo*	Localised skin lesion, hemangioma, juvenile onset, large hemangioendothelioma involving right buttock, right thigh since the age of 6, splenomegaly, abnormal thrombosis	novel
10	Fatal acute encephalitis	*DOCK11*	c.1679C>T [NM_144658.3], p.Ser560Leu	Trio	X	Encephalitis, abnormality of the spleen, abnormality of bone marrow cell morphology, lymphadenopathy	novel
11	Fine-Lubinsky syndrome	*HDLBP*	c.1731+1G>A [NM_203346.4], p.Val540_Leu577del	Two cousins	AR	Uplifted earlobe, short nose, short chin, shallow orbits, severe global developmental delay, plagiocephaly, narrow mouth, low-set, posteriorly rotated ears, hypertelorism, contracture of the proximal interphalangeal joint of the 5th finger, brain atrophy, bilateral camptodactyly, abnormal cornea morphology	novel

Table 3.1 **Characteristics of the HICF2 patient cases used in this thesis.** This table summarises the diagnoses, candidate genes, specific variants, number of samples used per analysis, and suspected inheritance patterns (AD = Autosomal dominant, AR = Autosomal recessive, X = X-linked recessive) for each HICF2 patient case analysed in this thesis. Furthermore, each patient's phenotype, described in HPO terms, is listed. Finally, it is indicated whether or not a candidate gene was known or novel for the patient's phenotype. (*) *POR* is a known gene for Antley-Bixler syndrome, a disease with a closely overlapping phenotype with FLS.

3.2.2 Whole genome sequencing and variant identification

WGS was conducted for probands and, where available, parents. WGS data was analysed and benchmark variants for each case were identified by senior post-doctoral geneticists using the HICF2 rare disease research bioinformatics pipeline (see Chapter 2 for a detailed description).

3.2.3 Algorithm settings for algorithm comparison

Prior to variant ranking, VCF files were filtered to only include variants in coding regions as well as 25 base pairs reaching into intronic regions from the start and end of exons. For coding region filtering, a corresponding BED file was downloaded from UCSC [177]. The bedtools algorithm was used for filtering [178]. For case 11 with FLS and a candidate variant in *HDLBP*, two affected cousins were sequenced (see Figure 6.1 in Chapter 6 for a detailed pedigree diagram). Since FLS is suspected to be autosomal recessive and the disease-causing variant must segregate in both branches of the pedigree, the VCF was further filtered to only include variants that were homozygous in both sequenced cousins. Subsequently, the filtered VCFs were processed by Exomiser and VAAST/VAAST+Phevor for the algorithm comparison.

Two different versions of Exomiser were used for two different types of analysis. Exomiser version 11.0.0 was used for a comparison with VAAST+Phevor. Exomiser versions 11.0.0 and 7.2.1 were used to analyse differences in performance of the framework for versions released approximately two-and-a-half years apart.

All variants were annotated using Exomiser's built-in version of Jannovar and the filtering tool included in the pipeline [89]. For a detailed description of the Exomiser settings used for the case analyses, see Chapter 2.

To test VAAST+Phevor, the Opal platform [54] (pipeline version b37:6.0.6, see Table 3.2 for details on database versions) by Fabric Genomics (formerly Omicia) was used, which included release 3.0.4.2 of VAAST and version 2.1 of Phevor. First, variants were filtered based on a minimum variant call quality of $\geq$5, and a MAF of $\leq$1% in the 1,000 Genomes project, the Exome Variant Server (EVS) and ExAC, as well as based on the suspected inheritance pattern. Importantly, Opal's inheritance pattern filter is more granular than Exomiser's. Opal differentiates between autosomal dominant, *de novo*, autosomal recessive and X-linked. Therefore, rankings produced using Exomiser's 'AD' or 'AR' filter generally contain more variants than Opal's output, where the 'AD' setting is split into autosomal dominant and *de novo*. After filtering in Opal, VAAST, which is included in the Opal platform, was used to rank variants. Finally, ranked output variants from VAAST were re-ranked with Phevor.

Name	Full Name	Release
1000 Genomes Project	1000 Genomes Project	Phase 3 Version 5 (2013-05-02)
CADD	Combined Annotation Dependent Depletion	v1.0
ClinVar	ClinVar	2016-09-12
COSMIC	Catalogue Of Somatic Mutations In Cancer	v78 (2016-09)
dbSNP	Single Nucleotide Polymorphism Database	v147
dbVar	dbVar	May 2016 Release
DGV	Database of Genomic Variants	2016-05-15
Ensembl	Ensembl	v83
EVS6500	Exome Variant Server	v.0.0.30 (2014-11-03)
ExAC	Exome Aggregation Consortium	v0.3
GeneSplicer	GeneSplicer Splice Site Prediction	2003-02-19
GERP++	Genome Evolutionary Rate Profiling	2010 release
GRCh	Genome Reference Consortium Human Genome Build	v37

Table 3.2 **Databases and database versions used for the analyses conducted with Fabric Genomics' Opal platform, build b37:6.0.6**. This table lists all external databases and version numbers utilised in Fabric Genomics' Opal platform that were used for the analysis of all HICF2 patient cases in this thesis.

3.2.4 Evaluation of ranking results

Manhattan plots were created to compare the individual algorithms' ranking performance based on the eleven RD benchmark cases. The quality of high ranking candidate variants was evaluated with the Integrative Genome Viewer (IGV) version 2.3.32 [163].

VAAST+Phevor assign a score and a unique rank to each variant in the VCF file. Two variants cannot have the same rank. The versions of Exomiser used for my analyses, however, do not assign ranks to variants and can assign the same score to multiple variants in the same VCF. For example, two different variants found in one patient could both be assigned an Exomiser variant score of 1.0. In order to be able to compare the two algorithm frameworks based on the rank assigned to the benchmark variants of interest, I assigned ranks to each variant in Exomiser's output VCF, starting with the first variant with the highest score in the output file receiving the first rank and ending with the last variant with the lowest score receiving the last rank. Importantly, no two variants received the same rank, even if they received the same Exomiser variant, phenotype, or combined score.

3.2.5 Determination of number of variants for further investigation

To analyse if the GPAs reduced the number of potentially disease-causing variants that require further analysis by a bioinformatician, I determined significance cut-offs for each algorithm. For VAAST, a p-value $p \leq 0.05$ was chosen. The Phevor score is a combination of the VAAST p-value ≤ 0.05 and a prior on the correlation between a gene and the patient's phenotype. A prior < 0.5 would mean the gene is negatively correlated with the proband's phenotype. Hence, to calculate the Phevor cut-off, a prior $q \geq 0.5$ was chosen. The Phevor cut-off F is calculated as follows [113]:

$$F = \frac{log(\frac{(1-p)q}{p(1-q)})}{log(10)} = 2.3 \tag{3.1}$$

For each of the three Exomiser scores, I computed the average score of all benchmark variants as a significance cut-off. Based on the determined significance cut-offs,

I calculated the percentage of ranked variants that would be regarded as significant and would thus require further analysis.

3.3 Results

The performance of all algorithms was analysed both on a summary (see Section 3.3.1) and individual case result level (see Section 3.3.2).

3.3.1 Summary results

To gain an in-depth understanding of the overall performance of the individual algorithms, I analysed the ability of the VPA to rank benchmark variants high, for known and novel disease gene candidates (see Section 3.3.1.1), and their ability to reduce the number of likely non-disease-causing variants that receive a high rank (see Section 3.3.1.2).

3.3.1.1 Ranking performance

To assess the ranking performance of the algorithms, I conducted two different types of analyses. First, I compared the, at the time of analysis, most up-to-date versions of Exomiser using the hiPHIVE algorithm and VAAST+Phevor with each other (Exomiser v11.0.0, VAAST release 3.0.4.2 and version 2.1 of Phevor, see Section 3.3.1.1.1). The goal of this analysis is to compare the performance of the two main analysis frameworks initially used by the world's largest RD study, the 100KG project. Thereafter, I compared two versions of Exomiser with each other that were released approximately two-and-a-half years apart (v7.2.1, released on January 5th, 2016, and v11.0.0, released on September 21st, 2018). This analysis was conducted to examine perfor-

mance differences based on database versions and minor algorithm modifications (see Section 3.3.1.1.2).

3.3.1.1.1 Ranking performance of the latest version of Exomiser compared with VAAST+Phevor

The overall performance of all five VPA was compared based on eleven HICF2 patient cases. The heights of the bars in Figures 3.1, 3.2, 3.3, and 3.4 summarise the number of benchmark variants that were ranked first, in the top 5, 10, or 20.

GAs vs GPAs: Overall, the two GPAs - Exomiser's combined score and VAAST + Phevor - outperform their respective GAs in the 'ranked first', 'top 5' and 'top 10' analyses. Exomiser's combined score ranks five variants first, in comparison to four variants for Exomiser's variant score and one variant for Exomiser's phenotype score. Similarly, VAAST+Phevor ranks eight variants first, improving upon VAAST's one first ranked variant. As the frame of analysis is expanded from first ranked to top 5 and top 10, that trend is continued, with the GPAs outperforming their respective GAs. Finally, Exomiser's combined score ranks a total of nine variants in the top 20, compared to Exomiser's phenotype score with seven and Exomiser's variant score with four variants. In contrast to that, VAAST catches up with VAAST+Phevor and both algorithms rank a total of nine variants in the top 20.

GPA vs GPA: VAAST+Phevor ranks more benchmark variants first than Exomiser's combined score (eight compared to five). As the analysis frame is expanded to the top 5, 10 and 20, Exomiser's combined score eventually catches up and both algorithms rank nine of the eleven benchmark variants in the top 20.

GA vs GA: Exomiser's variant score ranks four benchmark variants first, compared to one for VAAST. However, VAAST outperforms Exomiser's variant score in the remaining analyses, eventually catching up to the GPAs with a total of nine benchmark variants ranked in the top 20.

Performance for novel genes: Both analysis frameworks manage to capture all three novel gene candidate variants in the top 20 of their rankings. VAAST+Phevor even ranks all three novel gene candidate variants first, with Exomiser's combined score only ranking the candidate variant in *HDLBP* first due to its high predicted pathogenicity, but closing up in the top 20. Importantly, VAAST on its own also ranks one novel gene candidate variant - again, *HDLBP* - first and performs on par with the two GPAs for novel gene discovery, already ranking all three novel gene variant candidates in the top 5. Section 3.3.2.2.3 contains a detailed analysis of the candidate variant in *HDLBP* and its likely contribution to the FLS phenotype.

Variants not captured in top 20: Each of the two ranking frameworks only captures a total of nine out of eleven benchmark variants in the top 20. Both frameworks fail to rank the candidate variant in *CACNA1E* high, with Exomiser assigning a rank of 364 compared to VAAST+Phevor with 109 (see Figure 3.11). While the variant's predicted pathogenicity is high, *CACNA1E* does not carry any annotations in the HPO, which explains part of the low phenotype rank. Section 3.3.2.1.2 contains a more detailed description. Furthermore, VAAST+Phevor do not rank the candidate variant in *WWOX* due to an error in their annotation pipeline. The Opal platform, on which VAAST+Phevor are run, wrongly assigns the *WWOX* candidate variant to transcript ENST00000566780, where the variant falls into an intron and the algorithms thus do not rank it. The variant should, however, be mapped to transcript ENST00000355860, where it is correctly classified as a coding-region variant and thus receives a score. Lastly, the candidate variant in *POR* is only ranked 27th by Exomiser's combined score, just missing the top 20 cut-off. The reason for this is the large number of pathogenic candidate variants in the *POR* genome due to a loss of heterozygosity in that region caused by consanguinity (see Section 3.3.2.1.7 for details).

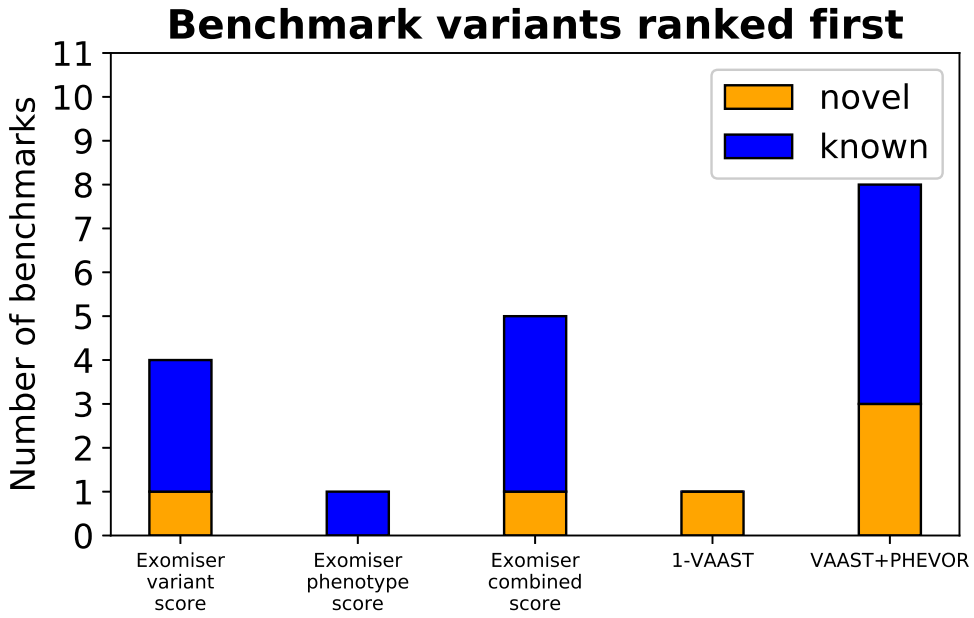


Fig. 3.1 Number of cases for which benchmark variants were ranked first for all Exomiser scores, VAAST, and VAAST+Phevor, differentiated by whether a gene was known (blue) or novel (orange) for the phenotype.

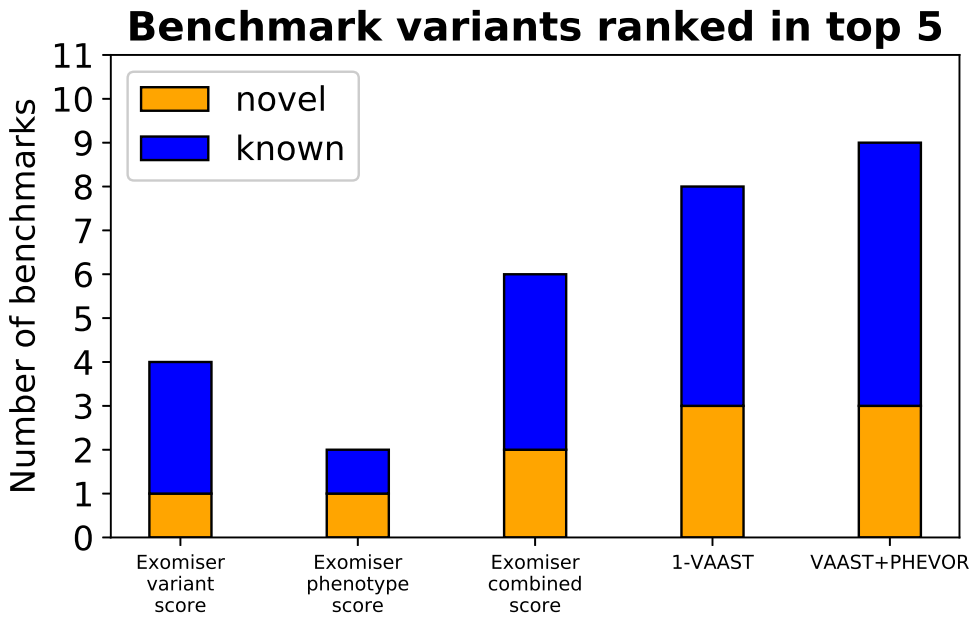


Fig. 3.2 Number of cases for which benchmark variants were ranked in top 5 for all Exomiser scores, VAAST, and VAAST+Phevor, differentiated by whether a gene was known (blue) or novel (orange) for the phenotype.

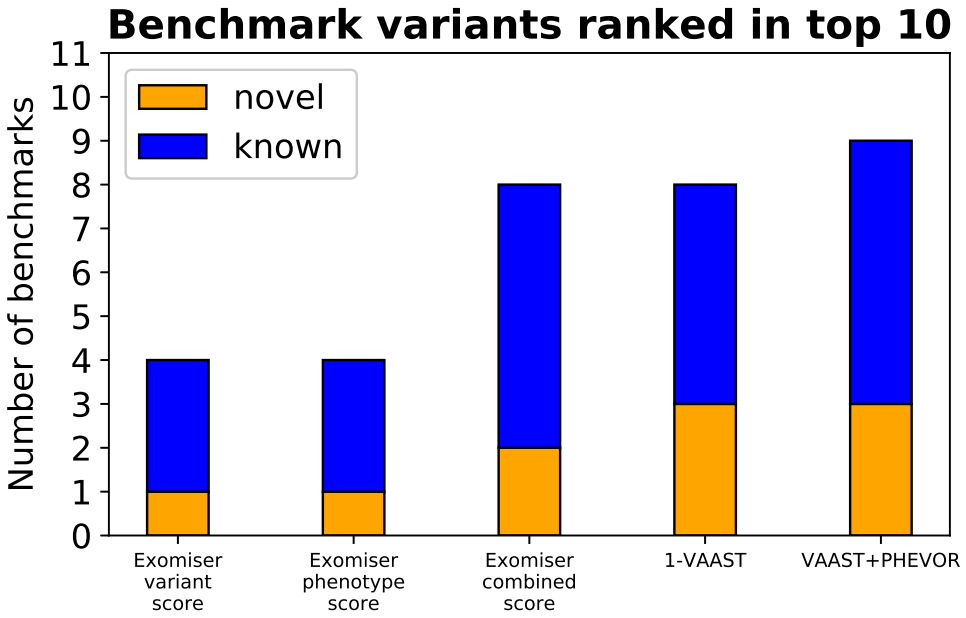


Fig. 3.3 Number of cases for which benchmark variants were ranked in top 10 for all Exomiser scores, VAAST, and VAAST+Phevor, differentiated by whether a gene was known (blue) or novel (orange) for the phenotype.

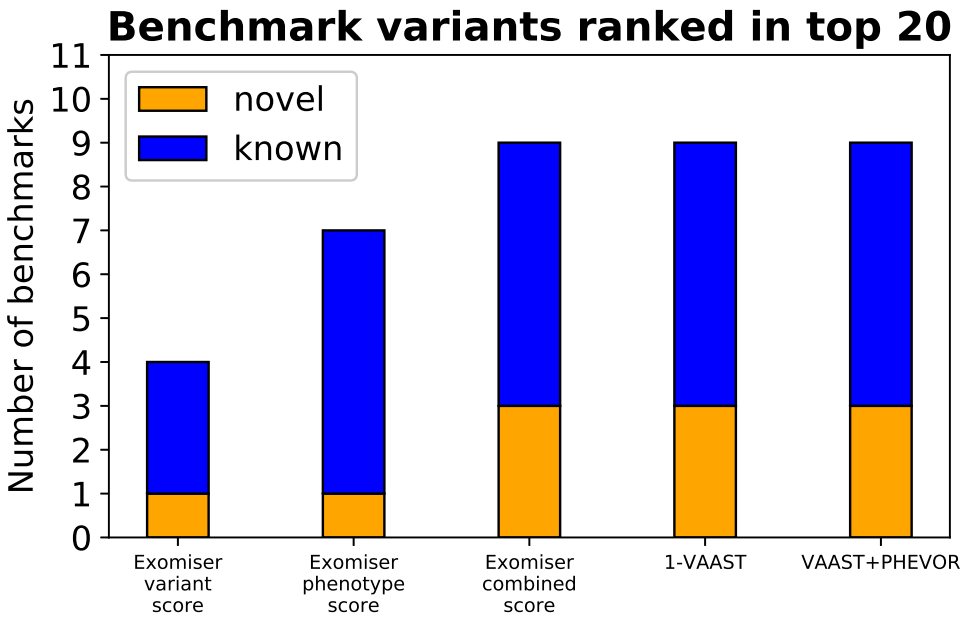


Fig. 3.4 Number of cases for which benchmark variants were ranked in top 20 for all Exomiser scores, VAAST, and VAAST+Phevor, differentiated by whether a gene was known (blue) or novel (orange) for the phenotype.

3.3.1.1.2 Ranking performance of different versions of Exomiser The field of rare genetic disease diagnostics is rapidly advancing, leading to genetics databases and VPA being updated frequently. Therefore, it is important to continuously re-analyse patient data to uncover likely disease-causing variants that were missed in prior analyses. To test the impact of running analyses at different time points on the aforementioned HICF2 cohorts, I analysed the eleven patient cases with two different versions of the Exomiser framework that were published over two-and-a-half years apart from each other (v7.2.1, released on January 5th, 2016, and v11.0.0, released on September 21st, 2018). VAAST+Phevor were only accessible through Fabric Genomics' Opal platform, which did not allow for a comparative assessment of different algorithm and database versions.

Figures 3.5, 3.6, 3.7 and 3.8 show the ranking performance of the two different algorithm versions for the number of benchmark variants ranked first, in the top five, ten and 20 respectively.

GPA new vs GPA old: The new version of Exomiser's combined score overall ranks more variants first than the old version (five compared to four), with that gap narrowing when the analysis window is widened to the top 20. Furthermore, the new version ranks all three novel gene candidates in the top 20, as opposed to the old version, which misses the candidate variant in *HDLBP*. Details on the *HDLBP* analysis can be found in Section 3.3.2.2.3. Both versions, however, rank one novel gene benchmark variant first. While the new version of Exomiser ranks more novel candidate genes in the top 20, the old version ranks one more known candidate variant in the top 20 than the new version. The gene that does not make it into the top 20 anymore in the new version of Exomiser is *POR*, which receives a lower rank in the new version than in the old version in all three scores. The respective patient case stems from a consanguineous marriage, which results in a loss of heterozygosity. Thus, the patient's genome harbours numerous homozygous variants that are predicted to be

pathogenic, which Exomiser's variant score cannot effectively discern from each other. Simultaneously, since the release of the old version of Exomiser, more genes have been associated with the patient's phenotype, resulting in the candidate gene receiving a comparatively lower phenotype score ranking. Those two effects combined result in the benchmark variant receiving a lower ranking.

GPAs (old and new) vs GAs (old and new): Irrespective of the release date, Exomiser's combined score always outperforms its respective Exomiser variant score for all four analyses.

GA new vs GA old: The new version of Exomiser's variant score outperforms the old version for the first-ranked and top 5-ranked variants, but succumbs to the old version for the top 10 and 20 analyses.

Phenotype score new vs phenotype score old: Counterintuitively, the phenotype score comparison follows a different pattern. With the exception of the top 20 analysis, the new version performs less well than the old version. One possible explanation is that, as studied cohorts grow and more genes are associated with phenotypic terms, the signal of purely phenotype-based scores for the ranking of known genes weakens. The fact that the average number of annotations per term has risen from ≈5.3 in 2010 (9,500 terms and 50,000 annotations [97]) to 12 in 2019 (13,000 terms and 156,000 annotations [179]) supports that hypothesis. As the HPO grows, additional features to weight the importance of individual term annotations for genes and diseases, such as the expected frequency of a phenotypic term for a specific disease, will likely become more relevant.

Irrespective of the varying performance of the phenotype score between versions, the overall GPA performs best in its latest version. The cases in this chapter thus were analysed using Exomiser v11.0.0. The patient cases in chapter five however were analysed at a different time point with Exomiser v7.2.1 (see Chapter 5).

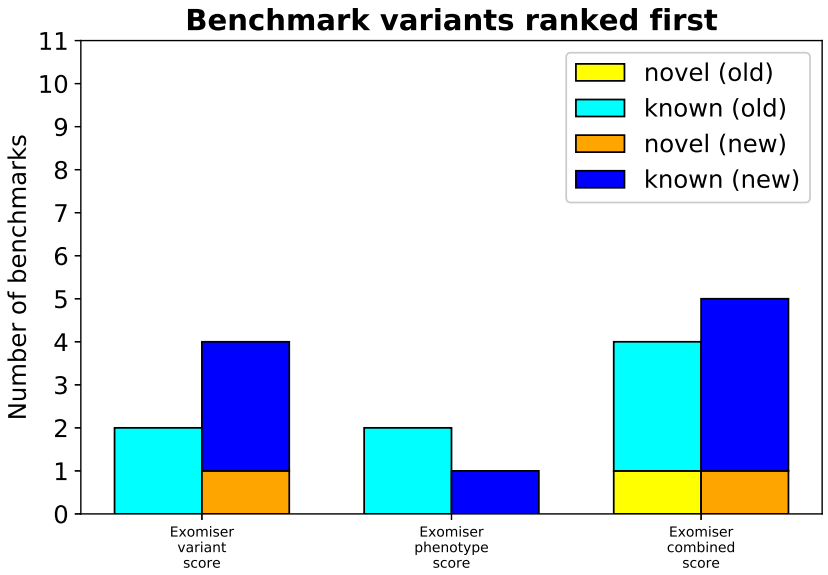


Fig. 3.5 Number of cases for which benchmark variants were ranked first for two different versions of Exomiser: v7.2.1, signified as 'old', and v11.0.0, signified as 'new'. Categories are further differentiated by whether a gene was known (turquoise and blue) or novel (yellow and orange) for the phenotype.

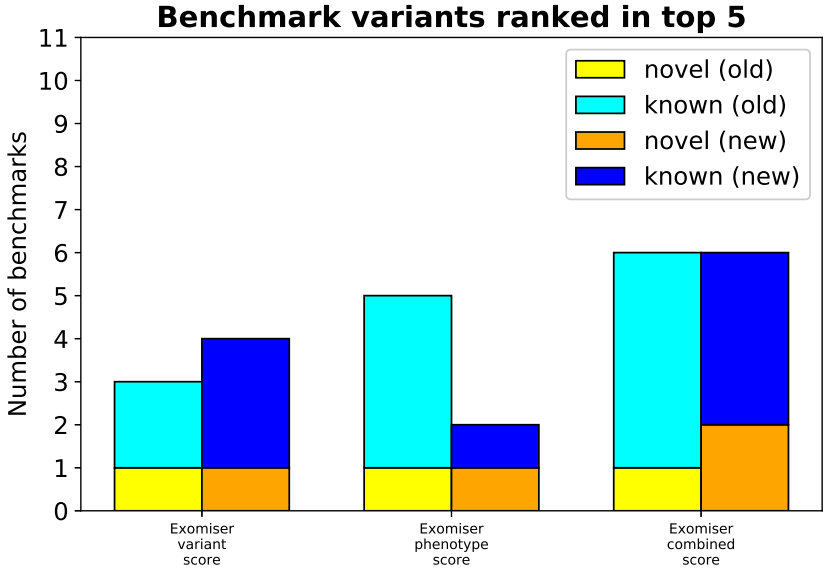


Fig. 3.6 Number of cases for which benchmark variants were ranked in the top 5 for two different versions of Exomiser: v7.2.1, signified as 'old', and v11.0.0, signified as 'new'. Categories are further differentiated by whether a gene was known (turquoise and blue) or novel (yellow and orange) for the phenotype.

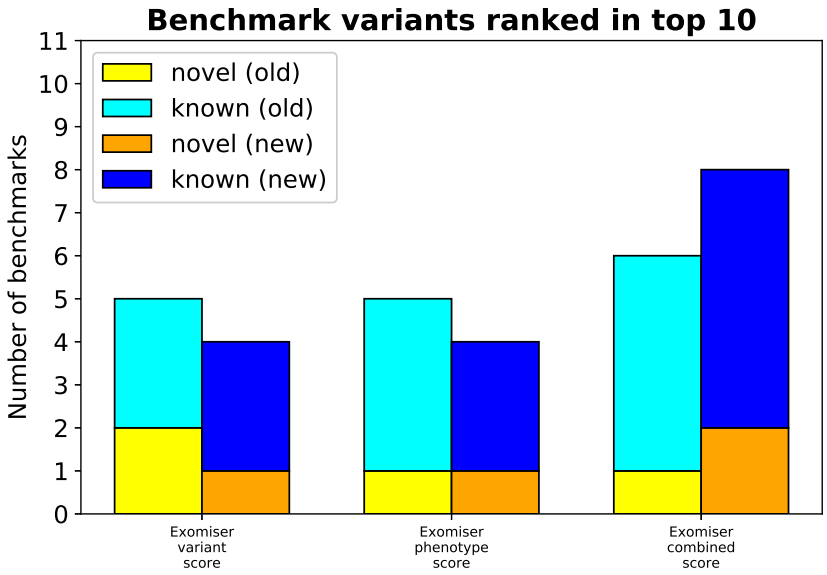


Fig. 3.7 Number of cases for which benchmark variants were ranked in the top 10 for two different versions of Exomiser: v7.2.1, signified as 'old', and v11.0.0, signified as 'new'. Categories are further differentiated by whether a gene was known (turquoise and blue) or novel (yellow and orange) for the phenotype.

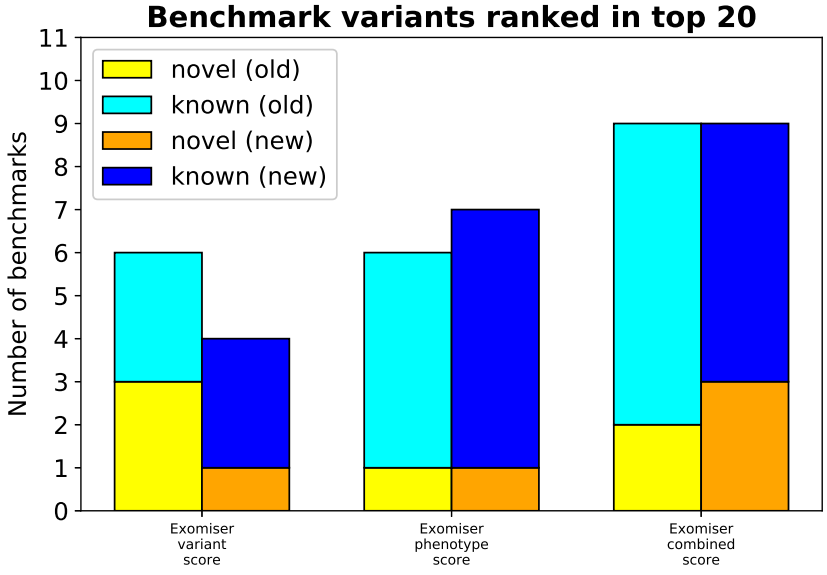


Fig. 3.8 Number of cases for which benchmark variants were ranked in the top 20 for two different versions of Exomiser: v7.2.1, signified as 'old', and v11.0.0, signified as 'new'. Categories are further differentiated by whether a gene was known (turquoise and blue) or novel (yellow and orange) for the phenotype.

3.3.1.2 Variant distribution

In addition to the ranking of the benchmark variants, I investigated the percentage of all ranked variants that receive a significant score by each algorithm and thus require further investigation by analysts.

It is important to note that the two frameworks used for benchmarking, Exomiser and VAAST+Phevor, rank different numbers of variants due to differences in their filtering approaches. When combining the variants of all eleven cases used for the benchmark analysis, Exomiser ranks 23,031 variants, in contrast to 1,254 for VAAST+Phevor (see Figure 3.9). The difference stems from different inheritance model settings and allele frequency databases used. For details on the individual filtering pipelines, see Section 3.2.3.

Thus, my analysis focuses on the relative reduction of the percentage of 'significant' candidate variants between the GA and the GPA in each pipeline, rather than the absolute score.

VAAST produces a p-value, providing a natural statistical cut-off of $p < 0.05$ that analysts can use for filtering lists of candidate variants. Based on VAAST's significance cut-off, the corresponding threshold for Phevor is 2.3 (see Section 3.2.5 for details). Figure 3.9, panel 4 shows a histogram for all variants ranked by VAAST for all eleven cases combined. The number of variants is plotted over $1 - VAAST$. For a cut-off of $VAAST < 0.05$ (or $1 - VAAST > 0.95$), 25.4% of the ranked variants are considered to be significant candidates. That percentage is reduced considerably by applying Phevor to the VAAST output, resulting in 11.2% of variants with a significant ranking result (> 2.3).

In contrast to VAAST+Phevor, Exomiser's combined score is produced with a logistic regression. This generates a likelihood that a variant is disease-causing with output values ranging from 0 to 1, but no statistical significance cut-off exists.

I therefore used the average score achieved by all of my benchmark variants in Exomiser's variant, phenotype and combined scores respectively to estimate the number of variants that receive a significant score and thus require further investigation. The average Exomiser variant score achieved by the benchmark variants is 1.0 (see Figure 3.9, panel 1), since all benchmark variants are predicted to be highly pathogenic, resulting in 32.3% of variants to be considered significant, as well as a score of 0.62 and 2.6% for Exomiser's phenotype score (see Figure 3.9, panel 2). The average significance cut-off for Exomiser's combined score is 0.88. A score of > 0.88 is achieved by 2.2% of all ranked variants.

While the percentage of variants with a significant score for Exomiser's phenotype score is lower than Exomiser's combined score, the benchmark performance data discussed in Section 3.3.1.1 shows that the overall best performance for variant prioritisation is achieved by Exomiser's combined score.

In summary, both GPAs reduce the percentage of ranked variants with a significant score that require further investigation compared to their GA counterparts, whilst also achieving better ranking performance metrics.

In addition to an analysis of the relative reduction of variants, I examined the output further. Different algorithms produce different distribution shapes. The Exomiser-specific algorithms largely cluster, with Exomiser's variant score almost exhibiting binary behaviour, resulting in a large number of variants with a very high or very low score, and few variants in between (see Figure 3.9, panel 1). Exomiser's phenotype score leads to two distinct clusters, one at 0.0 and one at 0.5 (see Figure 3.9, panel 2). Variants that are not annotated in any of the databases used to produce Exomiser's phenotype score, including the HPO, MPO, zebrafish-specific phenotype databases or the PPI network, receive a score of 0.0. The peak at 0.5 is a result of the PPI algorithm. The PPI scores range between 0.5 and 1.0. Thus, if a variant receives a score <0.5 from the species-specific phenotype algorithms and is annotated in the

PPI network, the variant will automatically score 0.5 or higher. In the histogram, ≈14,300 variants are plotted in the 0.5 bin, making up ≈62% of all variants plotted, which indicates that a significant number of variants are not annotated in any of the phenotype-specific databases. Exomiser's combined score behaves similarly to the variant and phenotype scores, where the 0.5 cluster observed for Exomiser's phenotype score migrated to a cluster at ≈0.75 (see Figure 3.9, panel 3). This is easily explained using Equation 1.3 (see Section 1.2.2.3.1). For a variant with an Exomiser variant score of 1.0 and an Exomiser phenotype score of 0.5, the Exomiser combined score is ≈0.75, corresponding to the large peak at the significance cut-off in Figure 3.9, panel 3.

VAAST produced a smoother distribution than Exomiser's variant score with ≈26.0% of all ranked variants receiving a score of ≈1.0 (see Figure 3.9, panel 4). VAAST's result distribution alone would make it difficult to discern disease-causing variants from variants that are merely predicted to be pathogenic. Phevor, however, re-ranks the variants, producing a close-to-Gaussian distribution that results in a smaller number of variants receiving significant scores (see Figure 3.9, panel 5).

While Exomiser, in absolute terms, assigns a significant score to a smaller percentage of all ranked variants, VAAST+Phevor's score distribution is more conducive to differentiating disease-causing variants from variants that are merely predicted to be pathogenic. One caveat, of course, is the determination of what makes a 'significant' score. While VAAST+Phevor's statistical framework can easily be analysed using the produced p-values, I had to determine a significance cut-off for Exomiser by averaging the scores of the benchmark variants.

The presented result summary demonstrates the advantages of GPAs over GAs for WGS analysis of RD patients, whilst also highlighting still existing shortcomings. The following in-depth analysis of the eleven patient cases will further elucidate the performance of the assessed algorithms for different analysis settings.

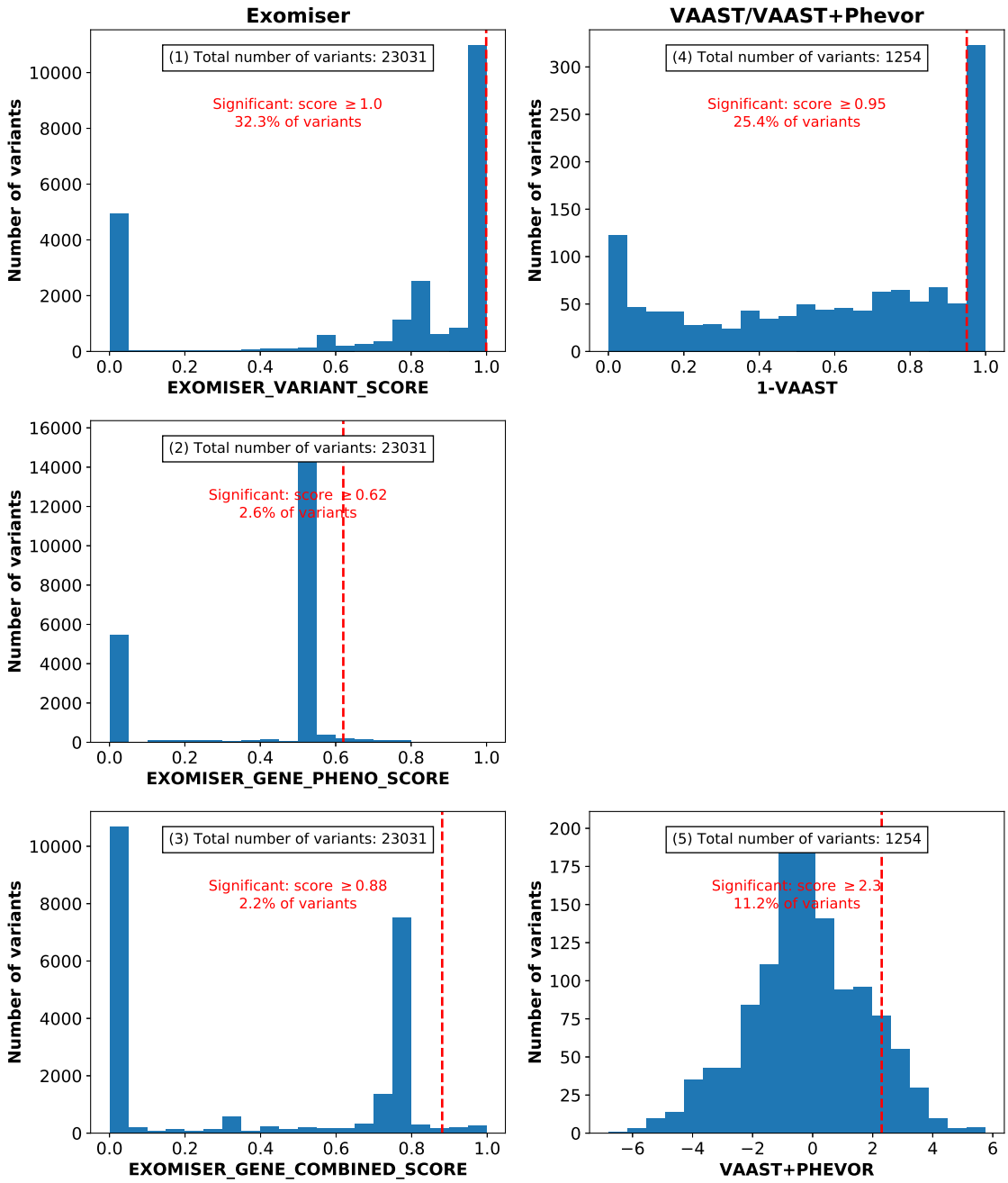


Fig. 3.9 Histogram of prioritisation scores, combining all variants for all benchmark cases for each algorithm. The red dotted light shows the significance cut-off for each algorithm: Exomiser variant score $\geq$ 1.00, Exomiser phenotype score $\geq$ 0.62, Exomiser combined score $\geq$ 0.88, 1-VAAST(p-value) $\geq$ 0.95 and Phevor score $\geq$ 2.3.

3.3.2 Case results

In addition to the summary-level analyses conducted to compare the variant ranking performance of Exomiser's hiPHIVE and VAAST+Phevor, I compare results for individual RD patient genome analyses. The following sections show the benchmarking results for cases with variants in both known (see Section 3.3.2.1) and novel genes (see Section 3.3.2.2) for the respective phenotype.

3.3.2.1 Known genes

3.3.2.1.1 Case 1: Distal arthrogryposis linked to stop-gain variant in *TNNI2*

3.3.2.1.1.1 Phenotype Case 1 is a patient with distal arthrogryposis. The distal arthrogryposis phenotype includes clenched fists, overlapping fingers, camptodactyly, ulnar deviation and positional foot deformities. The patient presented with all of those features.

3.3.2.1.1.2 Candidate gene The benchmark variant is a heterozygous stop-gain variant in the Troponin I2, Fast Skeletal Type (*TNNI2*) gene[1]. Variants in *TNNI2* affect the fast skeletal muscle through the alteration of Ca^{2+} concentrations and have previously been identified in patients with distal arthrogryposis with autosomal dominant inheritance [180]. The variant is predicted to be pathogenic by MutationTaster and is a known pathogenic variant for distal arthrogryposis in ClinVar (RCV000013249.18). Since the patient's benchmark variant is heterozygous, the suspected inheritance pattern is autosomal dominant or *de novo*. Parental genomes were not available and only the patient was sequenced. Thus, the inheritance pattern could not be further delineated. This case was selected to test the algorithms on a singleton with dominant or *de novo* inheritance of a pathogenic variant in a known gene. It is generally more

[1] c.466C>T [NM_001145829.1], p.Arg156*

challenging to identify disease-causing variants with a dominant or *de novo* inheritance pattern in patient genomes if no parental genomes are available to filter out non-disease causing variants.

3.3.2.1.1.3 Analysis VAAST achieved the worst performance of the VPA for *TNNI2*, ranking the benchmark variant at rank 14. The remaining algorithms ranked *TNNI2* first. Exomiser's variant score did so at the expense of clustering at 1.0, in contrast to Exomiser's combined score and VAAST+Phevor. In addition to successfully assigning the highest rank to the benchmark variant, Exomiser's combined score and VAAST+Phevor significantly reduced the number of variants that require further attention by the analyst by reducing the number of variants achieving a high score. Both GPAs resulted in a better ranking resolution than the GAs (see Figure 3.10).

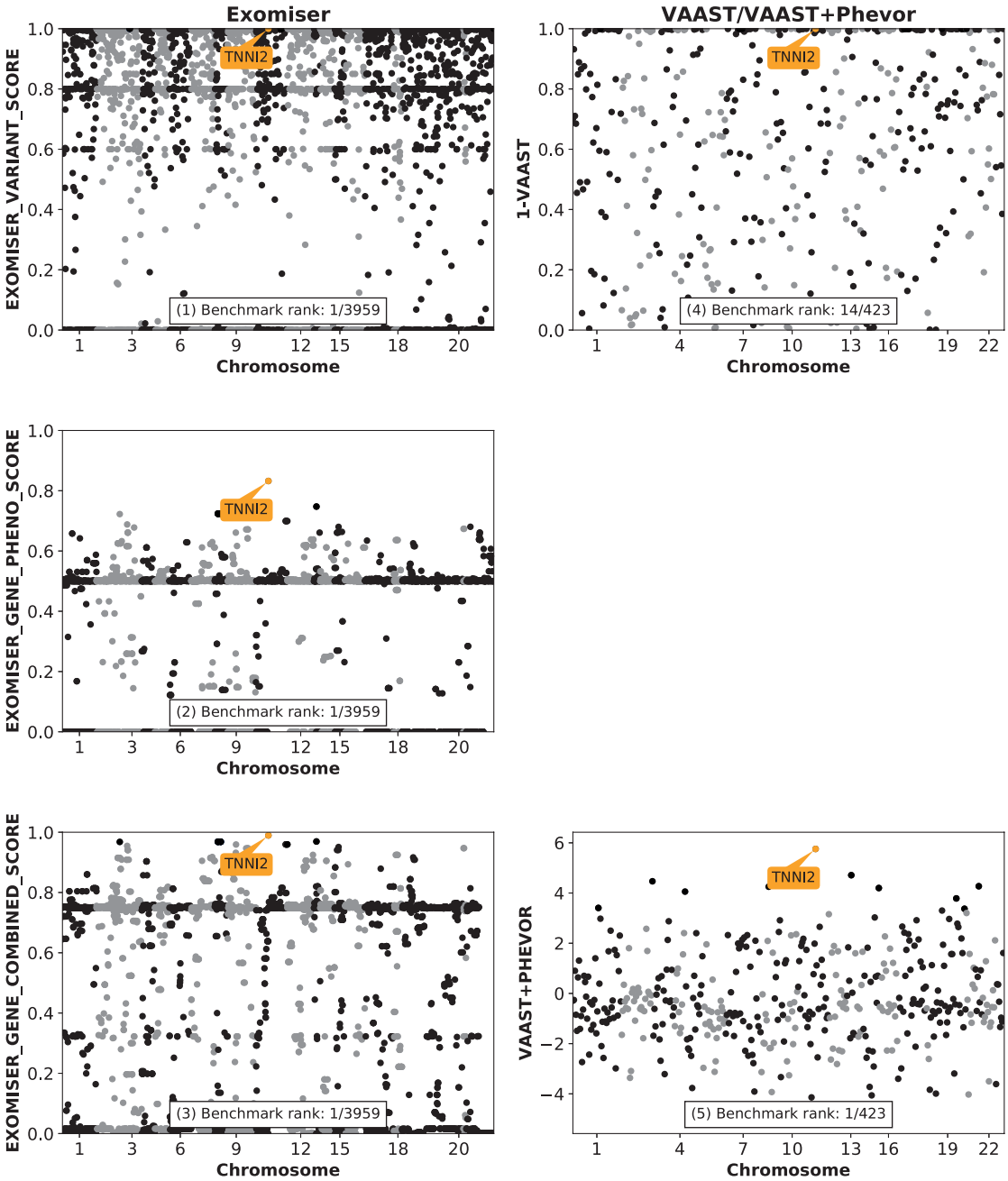


Fig. 3.10 Manhattan plots for case 1 (singleton, AD or *de novo*): distal arthrogryposis with stop-gain variant in *TNNI2*. Genomic coordinates on x-axes, 'importance' of variants on y-axes (variant score and combined score for Exomiser, $1 - p\_value$ for VAAST and the Phevor score for VAAST+Phevor). The benchmark variant is highlighted orange. (1) Exomiser's variant score ranks the *TNNI2* variant first. (2) Exomiser's phenotype score ranks the *TNNI2* variant first too, supported by existing HPO annotations to the gene for the patient's distal arthrogryposis phenotype. (3) Combining the results from Exomiser's component scores, the combined score ranks the *TNNI2* variant first with few close, but lower-ranked candidates. (4) VAAST ranks numerous variants high, including the *TNNI2* variant on rank 14. (5) VAAST+Phevor clearly identifies the *TNNI2* variant as the most likely candidate.

3.3.2.1.2 Case 2: Bilateral hippocampal sclerosis with missense variant in *CACNA1E*

3.3.2.1.2.1 Phenotype The proband in case 2 is affected by bilateral hippocampal sclerosis, a severe congenital neurological condition presenting with febrile convulsions, status epilepticus, cognitive difficulties and psychiatric comorbidities [181]. The patient's phenotype includes seizures and hippocampal dysgenesis.

3.3.2.1.2.2 Candidate gene Only a sample for the proband was available and sequenced. Prior analysis identified a heterozygous missense variant in the calcium voltage-gated channel subunit alpha1 E (*CACNA1E*) gene[2]. The *CACNA1E* gene encodes the voltage-dependent R-type calcium channel subunit alpha-1E protein. Voltage-dependent calcium channels mediate the entry of calcium ions into excitable cells [182]. Heterozygous *de novo* variants in *CACNA1E* are known to cause epileptic encephalopathy [183]. This variant is predicted to be pathogenic by MutationTaster (score: 1.0), SIFT (score: 0.0009), and Polyphen2 (score: 0.994) and was rare in gnomAD (0.012% in non-Finnish Europeans). Since the patient's benchmark variant is heterozygous and no parental samples were available for sequencing, the suspected inheritance pattern is autosomal dominant or *de novo*.

3.3.2.1.2.3 Analysis The benchmark variant is deleterious and thus ranked highly by Exomiser's variant score and VAAST. However, a large number of variants passed the filtering stages, in part because filtering based on inheritance was limited due to the unavailability of parental samples. Thus, the benchmark variant's score is not significant compared to the many other highly ranking variants. Despite the phenotype-based algorithms, Exomiser's phenotype score and Phevor, non-disease-causing variants are not sufficiently down-graded to create a clear picture. At the time of the analysis, *CACNA1E* was a known disease-causing gene for epileptic

[2] c.5702G>A [NM_001205293.1], p.Arg1901His

encephalopathy, but the gene had not yet been annotated with the disease and the respective phenotypic terms in the HPO. Hence, the algorithms were not able to rank the benchmark variant in *CACNA1E* in the top 20 or higher (see Figure 3.11).

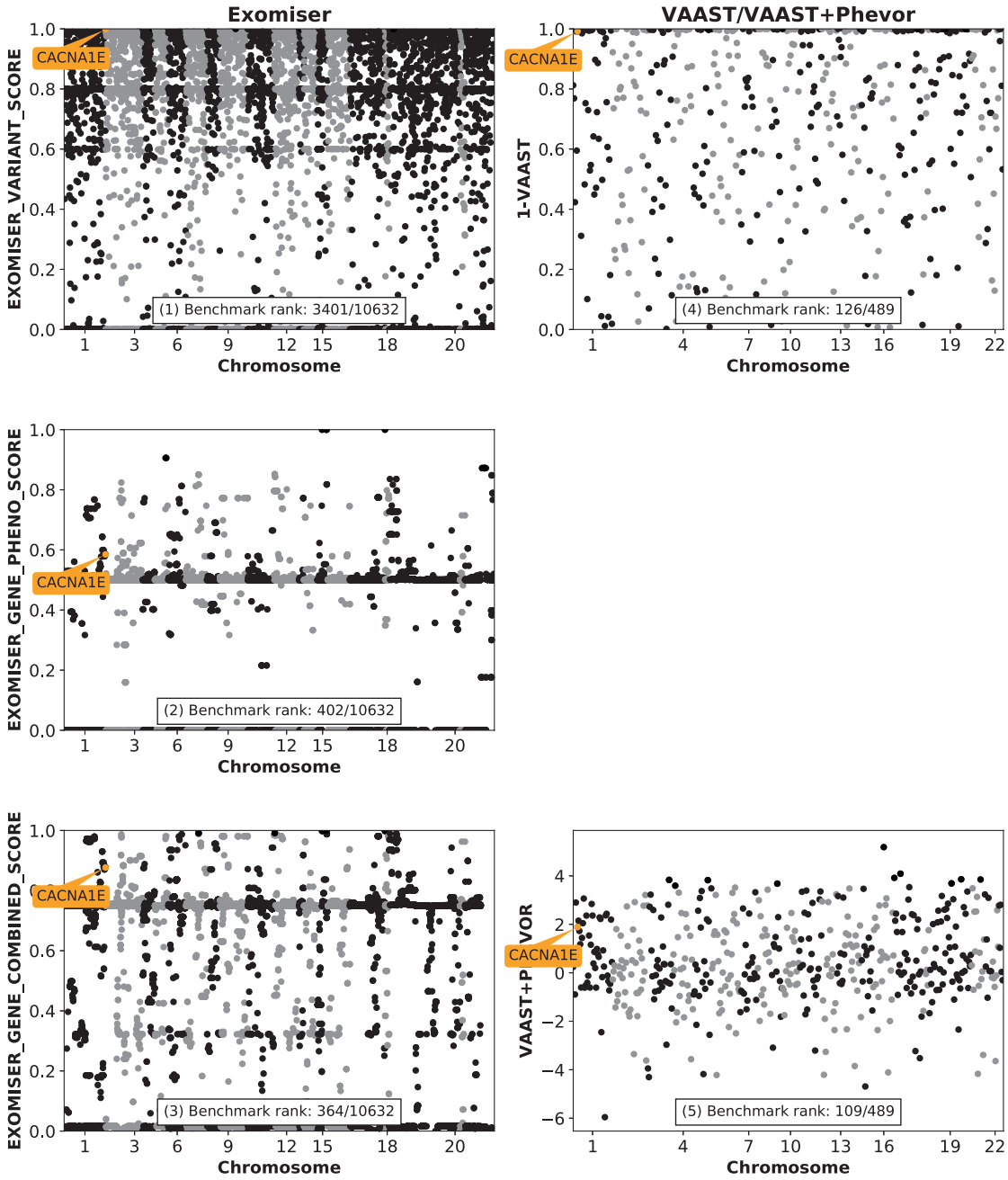


Fig. 3.11 Manhattan plots for case 2 (singleton, AD or *de novo*): bilateral hippocampal sclerosis with missense variant in *CACNA1E*. Genomic coordinates on x-axes, 'importance' of variants on y-axes (variant score and combined score for Exomiser, $1 - p\_value$ for VAAST and the Phevor score for VAAST+Phevor). The benchmark variant is highlighted orange. (1) Exomiser's variant score ranks *CACNA1E* in position 3401. (2) Exomiser's phenotype score ranks *CACNA1E* in 402nd place. (3) Exomiser's combined score ranks *CACNA1E* 364th, far from the top of the list. (4) VAAST ranks numerous variants high, including *CACNA1E* on rank 126. (5) VAAST+Phevor ranks *CACNA1E* 109th.

3.3.2.1.3 Case 3: Severe epileptic encephalopathy with frameshift variant in *WWOX*

3.3.2.1.3.1 Phenotype This case is a patient with severe early-onset epilepsy with epileptic encephalopathy. In a comprehensive study, Piard *et al.* [184] describe a condition affecting patients with pathogenic germline variants in the WW Domain-Containing Oxidoreductase (*WWOX*) gene, called *WWOX*-related epileptic encephalopathy (WOREE) syndrome. The WOREE phenotype involves dysmorphic features, including a round face with full cheeks, scoliosis or kyphosis, feeding and respiratory problems, severe developmental delay, hypertonia, early-onset epilepsy with drug-resistant daily seizures, as well as hypsarrhythmia in several cases, infantile spasms and visual impairment. The phenotype of the patient described here is largely consistent with WOREE (see Table 3.1).

3.3.2.1.3.2 Candidate gene The parents of the patient are unaffected and their genomes were not sequenced. The preceding HICF2 analysis identified a heterozygous frameshift insertion in *WWOX*[3] that was called homozygous on account of an exonic deletion. *WWOX* is a signalling protein involved in protein-protein interactions, the interruption of which has been named as a reason for the susceptibility of affected patients to seizures [185]. As described by Piard *et al.* [184] and others [186, 187], variants in *WWOX* have previously been associated with a range of early-onset epileptic encephalopathy types with autosomal recessive variants and is thus considered a known gene for the phenotype. The benchmark variant is rare and was not listed in gnoMAD, ExAC, or TOPDMED at the time of the analysis.

3.3.2.1.3.3 Analysis Similar to case one, Exomiser's variant score ranked the benchmark variant first, accompanied by a large number of variants predicted to

[3] c.705dupG [NM_016373.2], p.His236AlafsTer34

be pathogenic that cluster around a score of 1.0 (see Figure 3.12). In Exomiser's phenotype score, *WWOX* is rivalled by a handful of other genes with related phenotypic annotations. The variants in those genes, however, are not predicted to be pathogenic by Exomiser's variant score. Therefore, the *WWOX* variant achieves the highest Exomiser combined score. The phenotype score was thus helpful to discern the disease-causing variant from a large number of other predicted-to-be-pathogenic variants. In the latest version of the Opal platform, used in this analysis to run VAAST and VAAST+Phevor, the ENST00000566780 transcript is used for *WWOX*, in which the benchmark variant lies in an intronic region. This is a platform-internal error which should be avoided, since most bioinformatics pipelines generally use the gene transcript corresponding to the most severe consequence for the protein. Hence, the variant is not scored by VAAST and VAAST+Phevor.

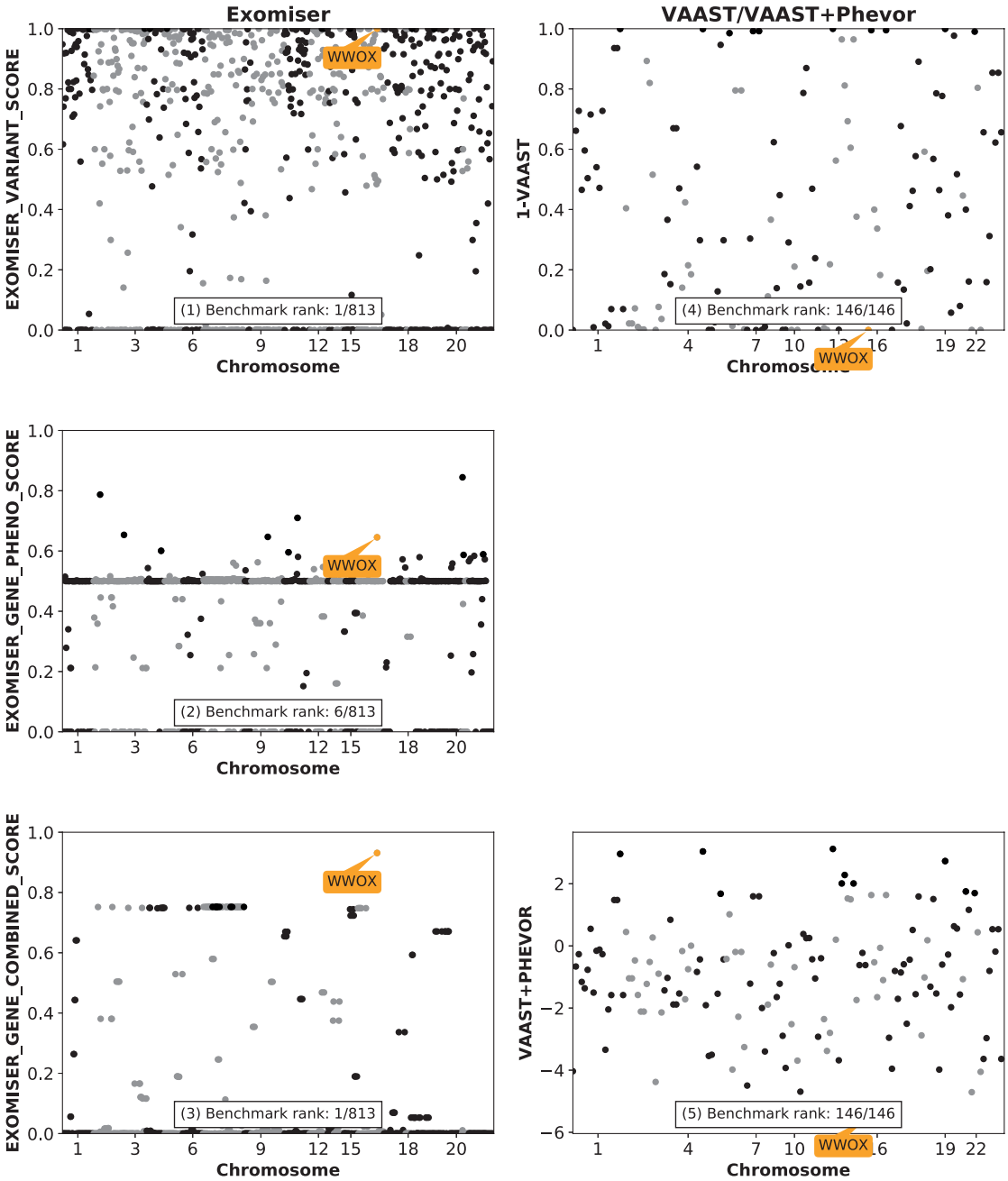


Fig. 3.12 Manhattan plots for case 3 (singleton, AR): severe epileptic encephalopathy for candidate frameshift insertion in *WWOX*. Genomic coordinates on x-axes, 'importance' of variants on y-axes (variant score and combined score for Exomiser, $1 - p\_value$ for VAAST and the Phevor score for VAAST+Phevor). The benchmark variant is highlighted orange. (1) Exomiser's variant score ranks benchmark variant first, but as part of a large cluster of variants. (2) Exomiser's phenotype score ranks the benchmark variant on rank six, only behind five other genes with relevant HPO annotations. (3) Exomiser's combined score ranks the benchmark variant first, as a result of a high predicted pathogenicity and phenotypic relevance. (4+5) VAAST and VAAST+Phevor do not rank the benchmark variant in *WWOX* due to a different transcript, in which the variant is intronic.

3.3.2.1.4 Case 4: Dilated cardiomyopathy (DCM) in childhood with missense variant in *ACTC1*

3.3.2.1.4.1 Phenotype This case involves DCM in childhood, a severe condition causing a dilated left ventricle and systolic dysfunction [188]. The proband presented with decreased weight for his age (<-2SD), endocardial fibroelastosis, dilated cardiomyopathy, and cardiomegaly.

3.3.2.1.4.2 Candidate gene The proband's parents were unaffected and their genomes were sequenced. The identified benchmark variant is a *de novo* missense variant in the Actin, Alpha, Cardiac Muscle 1 (*ACTC1*) gene[4]. The encoded protein has been shown to be essential for normal structure and function of cardiac myocytes and variants in *ACTC1* are known to cause DCM [189, 188]. The benchmark variant is predicted to be pathogenic by Polyphen2 (score: 0.968) and MutationTaster (score: 1.0) and was not listed in gnomAD or TOPMed at the time of analysis.

3.3.2.1.4.3 Analysis Exomiser's variant score for the benchmark variant is 1.0, ranking it at the top of the distribution in position 38, jointly with ≈40 other variants. *ACTC1* is a known gene for DCM and thus carries HPO annotations related to the patient's phenotype, which enable Exomiser's phenotype score to assign a score of ≈0.7 to the benchmark variant.

Only one gene carrying false positive variants, as determined in IGV, supersedes the *ACTC1* variant's Exomiser combined score. In comparison to Exomiser's variant score, the use of existing phenotypic annotations for *ACTC1* enables Exomiser to downgrade other variants with high pathogenicity scores that are not relevant for the DCM phenotype (see Figure 3.13).

[4] c.664G>A [NM_005159.4], p.Ala222Thr

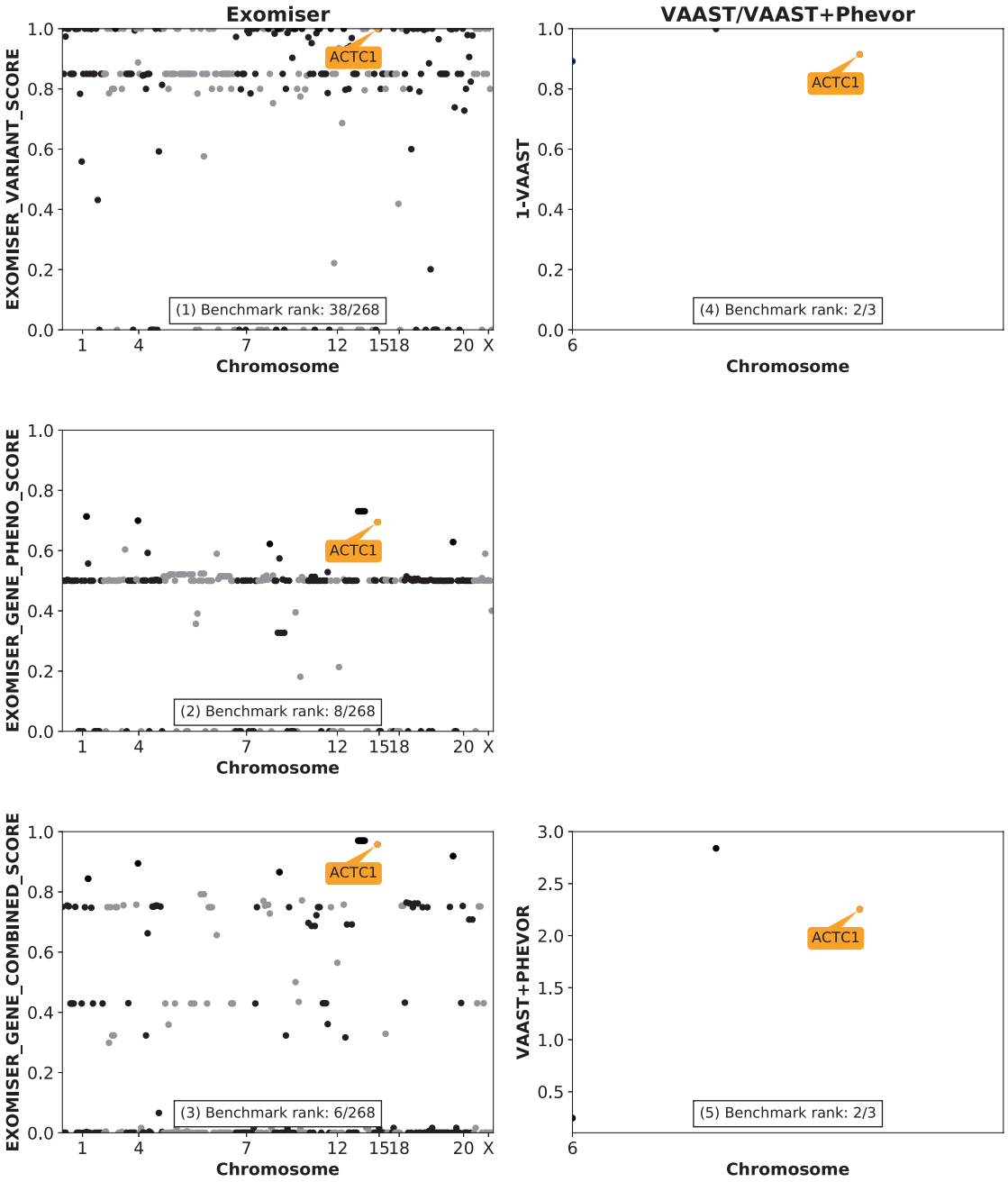


Fig. 3.13 Manhattan plots for case 4 (trio, *de novo*): DCM in childhood with candidate heterozygous missense variant in *ACTC1*. Genomic coordinates on x-axes, 'importance' of variants on y-axes (variant score and combined score for Exomiser, $1 - p\_value$ for VAAST and the Phevor score for VAAST+Phevor). The benchmark variant is highlighted orange. (1) Exomiser's variant score ranks *ACTC1* highly, in 38th place among a cluster of likely pathogenic variants. (2) Exomiser's phenotype score ranks *ACTC1* in 8th place due to existing HPO annotations for the known gene. (3) Exomiser's combined score ranks *ACTC1* 6th overall, close to the top of the list. (4) VAAST ranks the benchmark variant second, out of just three candidate variants due to effective filtering. (5) Similar to VAAST, VAAST+Phevor ranks *ACTC1* in second place.

3.3.2.1.5 Case 5: Majeed syndrome with missense variant in *PSTPIP1*

3.3.2.1.5.1 Phenotype This case involves Majeed syndrome, an auto-inflammatory condition with neutrophilic dermatosis, chronic recurrent multifocal osteomyelitis, and congenital dyserythropoietic anemia [190]. The patient presented with a largely overlapping phenotype with Majeed syndrome, including recurrent skin infections, bone marrow hypocellularity, episodic fever and splenomegaly.

3.3.2.1.5.2 Candidate gene The proband's parents were unaffected and their genomes were sequenced. The identified benchmark variant is a *de novo* missense variant in the proline-serine-threonine phosphatase interacting protein 1 (*PSTPIP1*) gene[5]. Pathogenic variants in *PSTPIP1* have previously been reported as causative for pyogenic sterile arthritis, pyoderma gangrenosum, and acne (PAPA) syndrome [191], a condition that's closely related to Majeed syndrome [192]. Candidate variants in *PSTPIP1* produce a hyper-phosphorilated *PSTPIP1* protein, altering its role in interleukin-1 (IL-1β) production [191]. The benchmark variant is predicted to be pathogenic by Polyphen2 (score: 1.0), MutationTaster (score: 1.0), and SIFT (score: 0.002) and was not reported in gnoMAD or TOPMed at the time of analysis and has since been submitted to ClinVar (VCV000097810.6) multiple times as pathogenic for PAPA.

3.3.2.1.5.3 Analysis The benchmark variant is highly deleterious and thus ranked highly by Exomiser's variant score, jointly with $\approx$350 other variants. In comparison to Exomiser's variant score, the use of existing phenotypic annotations for *PSTPIP1* enables Exomiser's combined score to downgrade other variants with high deleteriousness scores that are not relevant for the Majeed syndrome phenotype significantly, bringing the benchmark variant to rank 13 for Exomiser's stand-alone

[5] c.748G>A [NM_003978.4], p.Glu250Lys

phenotype score and rank 9 for the combined score. VAAST already ranks the benchmark variant in second place, while the re-ranking with Phevor brings the benchmark variant to rank number 1 by leveraging phenotypic information (see Figure 3.14).

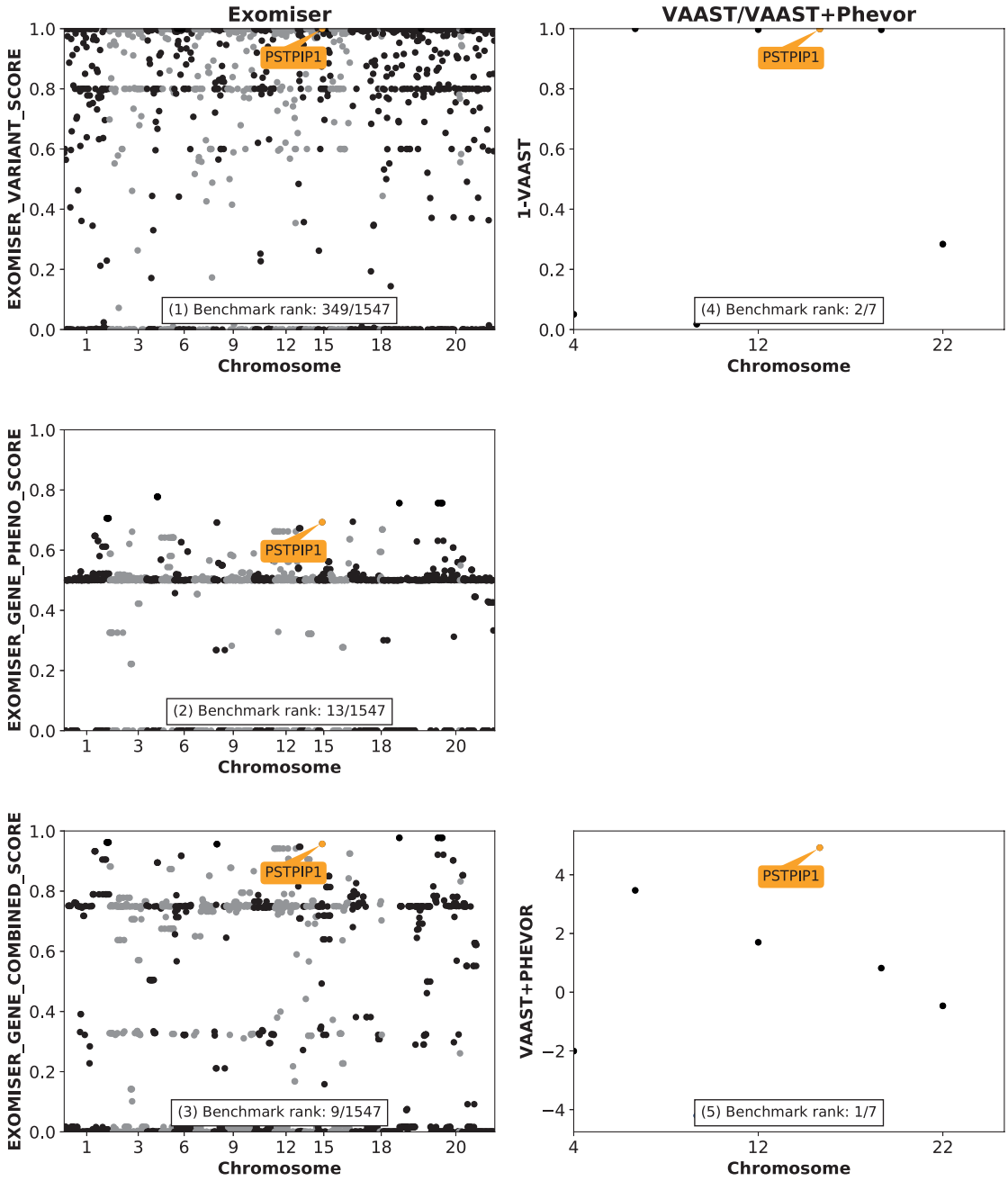


Fig. 3.14 Manhattan plots for case 5 (trio, *de novo*): Majeed syndrome for candidate missense variant in *PSTPIP1*. Genomic coordinates on x-axes, 'importance' of variants on y-axes (variant score and combined score for Exomiser, $1 - p\_value$ for VAAST and the Phevor score for VAAST+Phevor). The benchmark variant is highlighted orange. (1) Exomiser's variant score ranks *PSTPIP1* in 349th place among a cluster of likely pathogenic variants. (2) Exomiser's phenotype score ranks *PSTPIP1* in 13th place due to existing HPO annotations for the known gene. (3) Exomiser's combined score ranks *PSTPIP1* 9th overall, close to the top of the list. (4) VAAST ranks the benchmark variant second, out of just seven candidate variants due to effective filtering. (5) VAAST+Phevor successfully ranks *PSTPIP1* in first place.

3.3.2.1.6 Case 6: Undefined immunodysregulatory disorder with missense variant in *SAMD9L* or *DCXR*

3.3.2.1.6.1 Phenotype The patient in this case suffers from an undefined immunodysregulatory disorder with hypogammaglobulinaemia, enteropathy, seronegative arthropathy, psoriatic skin disease, cerebellar atrophy, and white and grey matter changes.

3.3.2.1.6.2 Candidate gene Samples from the unaffected parents were included in the analysis. Prior analysis identified two lead candidate genes. The first candidate gene is the Sterile Alpha Motif Domain Containing 9 Like (*SAMD9L*) gene, which carries two non-synonymous *de novo* missense variants[6], both of which lie on a common paternal haplotype, as confirmed via long-read single-molecule sequencing with an Oxford Nanopore Technologies (ONT) MinION [193]. Although further functional validation is required, there is evidence to suggest that the p.Tyr1118Cys allele is more likely to be disease-causing. p.Tyr1118Cys is a novel variant, in contrast to p.Arg359Gln, which is present in gnomAD at an allele frequency of 2/245,750. p.Tyr1118Cys further scores slightly higher in pathogenicity prediction algorithms (PolyPhen2 value: 0.701, SIFT value: 0.001) than p.Arg359Gln (PolyPhen2 value: 0.620, SIFT value: 0.028). Additionally, now-published disease-causing variants in *SAMD9L* are associated with Ataxia-Pancytopenia Syndrome, a condition presenting with cerebellar ataxia, predisposition to marrow failure, myeloid leukemia, and variable hematologic cytopenias, and thus an overlapping phenotype with our patient. Those disease-causing variants are almost exclusively located at the protein C terminal [194–198].

[6](1) c.3353A>G [NM_152703.2], p.Tyr1118Cys (2) c.1076G>A [NM_$NM_1$52703], $p.Arg359Gln$

The second candidate was the Dicarbonyl And L-Xylulose Reductase (*DCXR*) gene, carrying a mosaic *de novo* missense variant[7]. When this analysis was conducted in 2016, both genes had not been described in connection to this phenotype. Due to the patient's severe phenotype, which dictates an impactful variant, and the fact that the *DCXR* variant was mosaic, which is generally associated with less pronounced phenotypes, analysts considered *SAMD9L* to be the lead candidate. *SAMD9L* was subsequently connected to Ataxia-Pancytopenia Syndrome, a condition with a high phenotypic overlap with our patient, thus confirming the gene as a candidate [194], whereas to date, there is no evidence supporting *DCXR* as a causative gene. The case is included in the analysis as an additional example to test the algorithm performance for, at the time, novel gene candidates with *de novo* inheritance in a trio.

3.3.2.1.6.3 Analysis The analysis results are shown in Figure 3.15. *SAMD9L* outranks *DCXR* for all of Exomiser's scores, as well as VAAST+Phevor. VAAST ranks the benchmark variants in both genes equally high. Phenotypic evidence captured by Exomiser and Phevor respectively suggests a stronger link of *SAMD9L* to the patient's phenotype, ranking *SAMD9L* significantly higher than *DCXR* (Exomiser: combined score of >0.95 for *SAMD9L* vs. ≈0.2 for *DCXR*, VAAST+Phevor: >5.0 for *SAMD9L* vs. ≈2.0 for *DCXR*, missing the significance cut-off at 2.3).

[7] c.643G>A [NM_016286.3], p.Val215Met

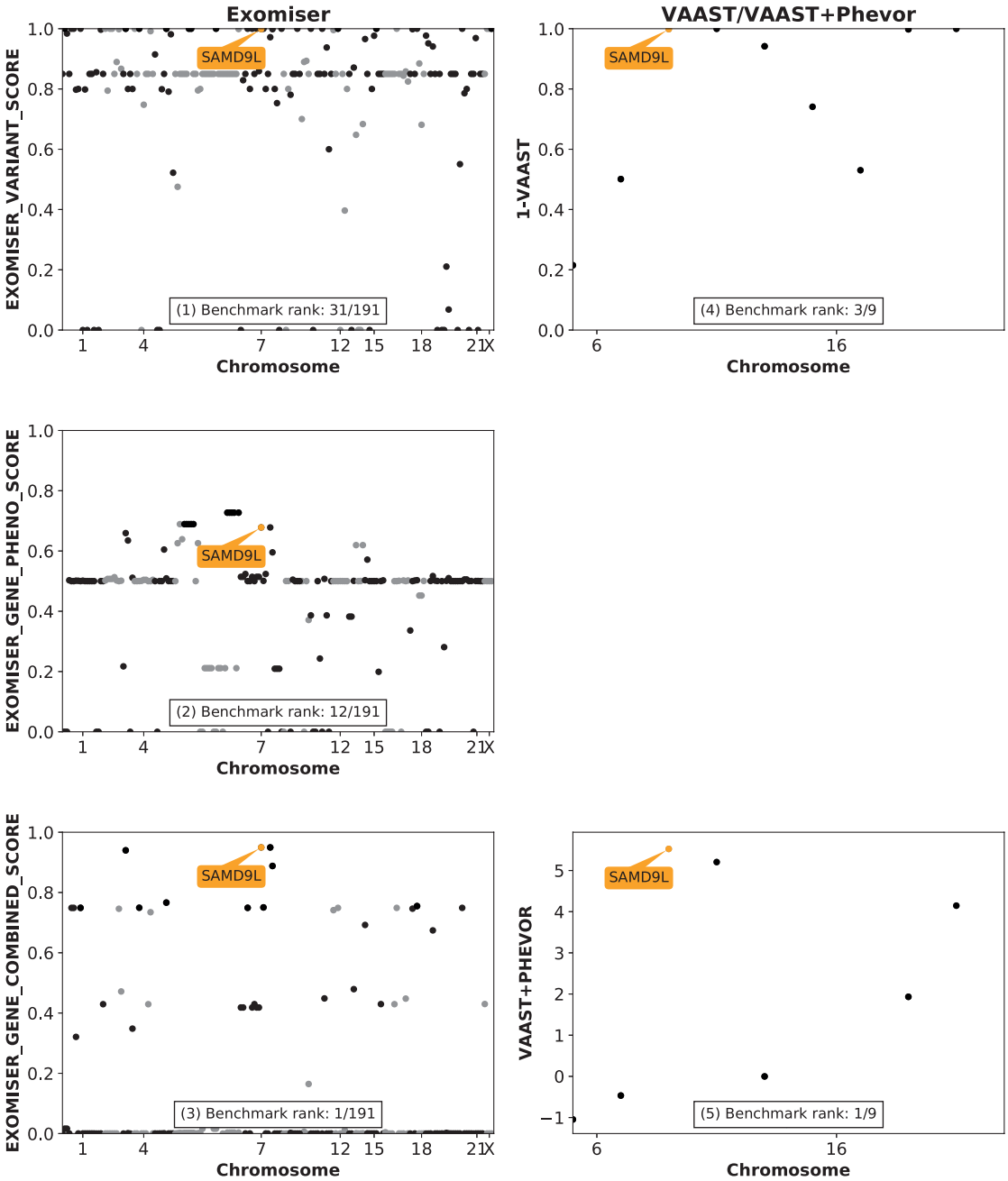


Fig. 3.15 Manhattan plots for case 6 (trio, *de novo*): undefined immunodysregulatory disorder for *de novo* missense candidate variants in *SAMD9L*. Genomic coordinates on x-axes, 'importance' of variants on y-axes (variant score and combined score for Exomiser, $1 - p\_value$ for VAAST and the Phevor score for VAAST+Phevor). The benchmark variant is highlighted orange. (1) Exomiser's variant score ranks *SAMD9L* in 31st place among a cluster of likely pathogenic variants. (2) Exomiser's phenotype score ranks *SAMD9L* in 12th place due to existing HPO annotations for the known gene. (3) Exomiser's combined score ranks *SAMD9L* 1st. (4) VAAST ranks the benchmark variant third, out of just seven candidate variants due to effective filtering. (5) VAAST+Phevor successfully ranks *SAMD9L* in first place.

3.3.2.1.7 Case 7: Fine-Lubinsky syndrome with missense variant in *POR*

3.3.2.1.7.1 Phenotype This patient was diagnosed with FLS by Dr Usha Kini, a consultant in Clinical Genetics at the Oxford University Hospitals, an autosomal recessive disorder with a phenotype including nonsynostotic plagiocephaly, megalocornea, cleft palate, digital abnormalities, including camptodactyly, brachydactyly, syndactyly and clinodactyly, as well as dysmorphic facial features. Presenting with a cleft palate, narrow mouth, micrognathia, short chin, shallow orbits, agenesis of permanent teeth, plagiocephaly, megalocornea, preauricular skin tag, cupped ears, arthrogryposis multiplex congenita, hypospadias, renal agenesis, moderate global developmental delay, and talipes, the patient's phenotype largely overlaps with FLS.

3.3.2.1.7.2 Candidate gene The proband is the only affected individual in a consanguineous family, of which both parents were sequenced. A homozygous missense variant in the Cytochrome P450 Oxidoreductase (*POR*) gene[8] is the lead candidate. *POR* is a known gene for Antley-Bixler syndrome [199], a condition with a high, but incomplete phenotypic overlap[9] with FLS [200]. The benchmark variant is predicted to be pathogenic by MutationTaster (score: 0.93) and SIFT (score: 0.006) and was not listed in gnomAD and TOPMed at the time of analysis. Analysing recessive conditions in consanguineous pedigrees is challenging due to the high number of homozygous variants that are difficult to distinguish from the benchmark variant [201].

3.3.2.1.7.3 Analysis The results in Figure 3.16 demonstrate the advantages of GPAs. Exomiser's variant score cannot distinguish between the benchmark and numerous homozygous variants. VAAST ranks the benchmark variant second, but clusters it with other homozygous but likely non-disease-causing variants. Ranked first

[8] c.1493G>C [NM_000941.2], p.Arg498Pro
[9] including craniosynostosis, brachycephaly and midface hypoplasia

by VAAST+Phevor, the benchmark variant is easily identifiable when using the GPA. Exomiser's combined score only brings the variant up to rank 27, which nonetheless represents a significant improvement compared to rank 426 with Exomiser's stand-alone variant score. This case demonstrates that GPAs can help to both improve the accuracy of WGS data analysis and save time for clinicians.

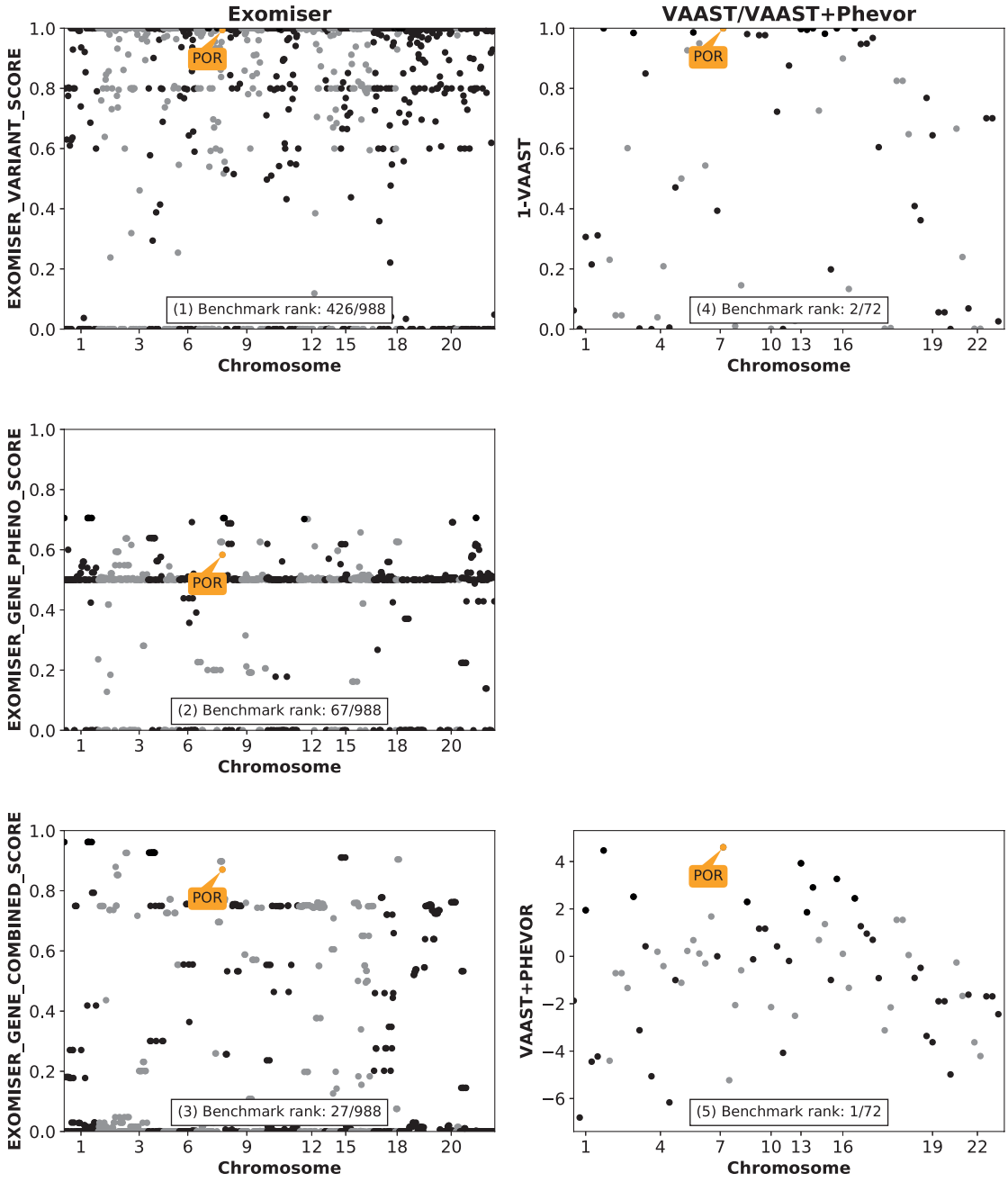


Fig. 3.16 Manhattan plots for case 7 (trio, AR): FLS for candidate homozygous missense variant in *POR*. Genomic coordinates on x-axes, 'importance' of variants on y-axes (variant score and combined score for Exomiser, $1 - p\_value$ for VAAST and the Phevor score for VAAST+Phevor). The benchmark variant is highlighted orange. (1) Exomiser's variant score ranks *POR* in 426th place among a cluster of likely pathogenic variants. (2) Exomiser's phenotype score ranks *POR* in 67th place. (3) Exomiser's combined score ranks *POR* 27th. (4) VAAST ranks the benchmark variant second. (5) VAAST+Phevor successfully ranks *POR* in first place.

3.3.2.1.8 Case 8: Congenital erythrocytosis with missense variant in *SLC30A10*

3.3.2.1.8.1 Phenotype Congenital erythrocytosis, the disease affecting this patient, is a condition where the red cell mass is greater than 125% than predicted given the patient's sex and body mass, usually correlated with elevated hemoglobin and/or hematocrit [202]. Affected by cerebral hemorrhage, hypotension, hemangioma, varicose veins, stroke, increased hemoglobin, peripheral thrombosis, increased hematocrit, increased red blood cell mass, stroke, headache, an abnormal integument, plethora, neoplasm, and hemangioma, the proband was diagnosed at age 4.

3.3.2.1.8.2 Candidate gene The proband and her parents were sequenced. Prior analysis identified a homozygous variant in the solute carrier family 30 member 10 (*SLC30A10*) gene[10]. *SLC30A10* is a well-established disease gene for hypermanganesemia with dystonia-1, an autosomal recessive metabolic disorder with motor neurodegeneration with extrapyramidal features, increased serum manganese, polycythemia and hepatic dysfunction, which can sometimes lead to cirrhosis with usually preserved intellectual function [203, 204]. In these patients, manganese excretion is severely impaired due the loss of function of *SLC30A10*, which leads to its accumulation in liver, brain and peripheral tissues. Hypermanganesemia is known to have a hypoxia-related effect on erythropoietin gene expression, causing polycythemia [205]. Following the genetic analysis, manganese levels in the patient's blood were tested and shown to be elevated. The benchmark variant was predicted to be pathogenic by SIFT (score: 0.001), Polyphen2 (score: 1.0), and MutationTaster (score: 1.0) and was not listed in gnomAD and TOPMed.

3.3.2.1.8.3 Analysis Both Exomiser's variant score and VAAST rank the deleterious benchmark variant in *SLC30A10* high, on rank 1 and 3 respectively. While

[10] c.823T>A [NM_018713.2], p.Trp275Arg

both algorithms produce large clusters of predicted-to-be-pathogenic variants at the top of the distribution, the use of phenotypic information with Exomiser's combined score and VAAST+Phevor significantly down-grades variants that are not related to the patient's phenotype. Exomiser's combined score and VAAST+Phevor both rank the benchmark variant first (see Figure 3.17).

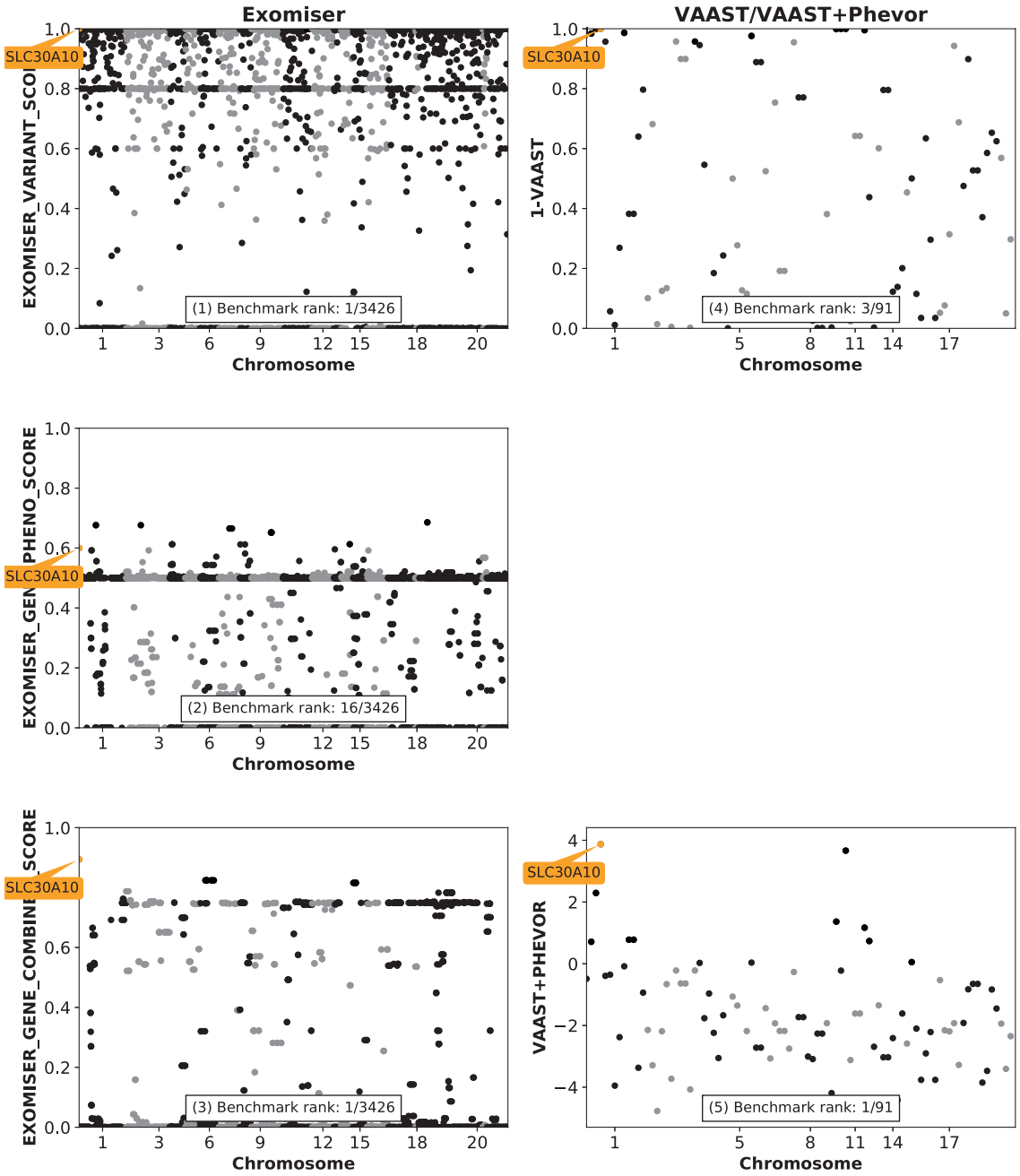


Fig. 3.17 Manhattan plots for case 8 (trio, AR): Congenital erythrocytosis for candidate homozygous missense variant in *SLC30A10*. Genomic coordinates on x-axes, 'importance' of variants on y-axes (variant score and combined score for Exomiser, $1 - p\_value$ for VAAST and the Phevor score for VAAST+Phevor). The benchmark variant is highlighted orange. (1) Exomiser's variant score successfully ranks *SLC30A10* in 1st place. (2) Exomiser's phenotype score ranks *SLC30A10* in 16th place, given holistic phenotypic annotation of the known gene. (3) Exomiser's combined score ranks *SLC30A10* 1st. (4) VAAST ranks the benchmark variant third. (5) VAAST+Phevor successfully ranks *SLC30A10* in first place.

3.3.2.2 Novel genes

3.3.2.2.1 Case 9: Atypical Klippel-Trenaunay syndrome with missense variant in *RBPJ*

3.3.2.2.1.1 Phenotype This patient has an atypical form of Klippel-Trenaunay syndrome (KTS). KTS is a vascular disease with combined vascular malformations of the capillary, venous, and lymphatic types, unusually distributed varicosities and limb enlargement [206]. The patient's phenotype is considered to be atypical due to superficial skin lesions that do not match those reported for KTS. The patient's phenotype further includes juvenile onset hemangioma, large hemagioendothelioma on the right buttock and right thing since the age of six, as well as splenomegaly and abnormal thrombosis.

3.3.2.2.1.2 Candidate gene The preceding analysis identified three lead candidates: a *de novo* splice site missense variant in the Recombination Signal Binding Protein For Immunoglobulin Kappa J Region (*RBPJ*) gene[11], a *de novo* missense variant in the TRNA Methyltransferase 1 (*TRMT1*) gene[12], and a compound heterozygous missense variant in the Patched 1 (*PTCH1*) gene[13]. *RBPJ*, a novel gene candidate, is involved in regulation of Notch signalling [207], which plays an important role in vascular development [208], and was considered to be the most likely candidate by the HICF2 analysts at the time this analysis was conducted, but had not been definitively confirmed as disease-causing. The benchmark variant in *RBPJ* is predicted to be pathogenic by MutationTaster (score: 1.0) and SIFT (score: 0.063) and was not reported in gnomAD and TOPMed.

[11] c.535T>G [NM_005349], p.Leu179Val
[12] c.1149G>C [NM_001142554], p.Glu383Asp
[13] c.1405G>A [NM_000264.3], p.Val469Met and c.404G>A [NM_000264.3], p.Arg135Gln

3.3.2.2.1.3 Analysis Splice site variants, by nature, are highly deleterious and therefore receive high scores in pathogenicity-based algorithms. Thus, Exomiser's variant score ranks the *RBPJ* variant high, with a score of 1.0, among ≈60 other predicted-to-be-pathogenic variants. VAAST ranks the *RBPJ* third, behind *TRMT1*. Once phenotypic information is taken into account, however, both GPAs effectively distinguish the *RBPJ* variant from the rest of the variants. *RBPJ* is proximate to *DTX1* in Exomiser's PPI and *DTX1* is annotated with 'spleen hyperplasia' in the MPO (MP:0000693), a phenotypic match for the patient's splenomegaly, resulting in a phenotype score of 0.5 for *RBPJ*. Consequently, Exomiser's combined score ranks *RBPJ* 13th, while VAAST+Phevor assigns the first rank to the candidate gene. Despite the lack of direct human phenotypic annotations for *RBPJ* as a novel gene candidate, the GPAs are able to distinguish between the *RBPJ* variant and other variants (see Figure 3.18).

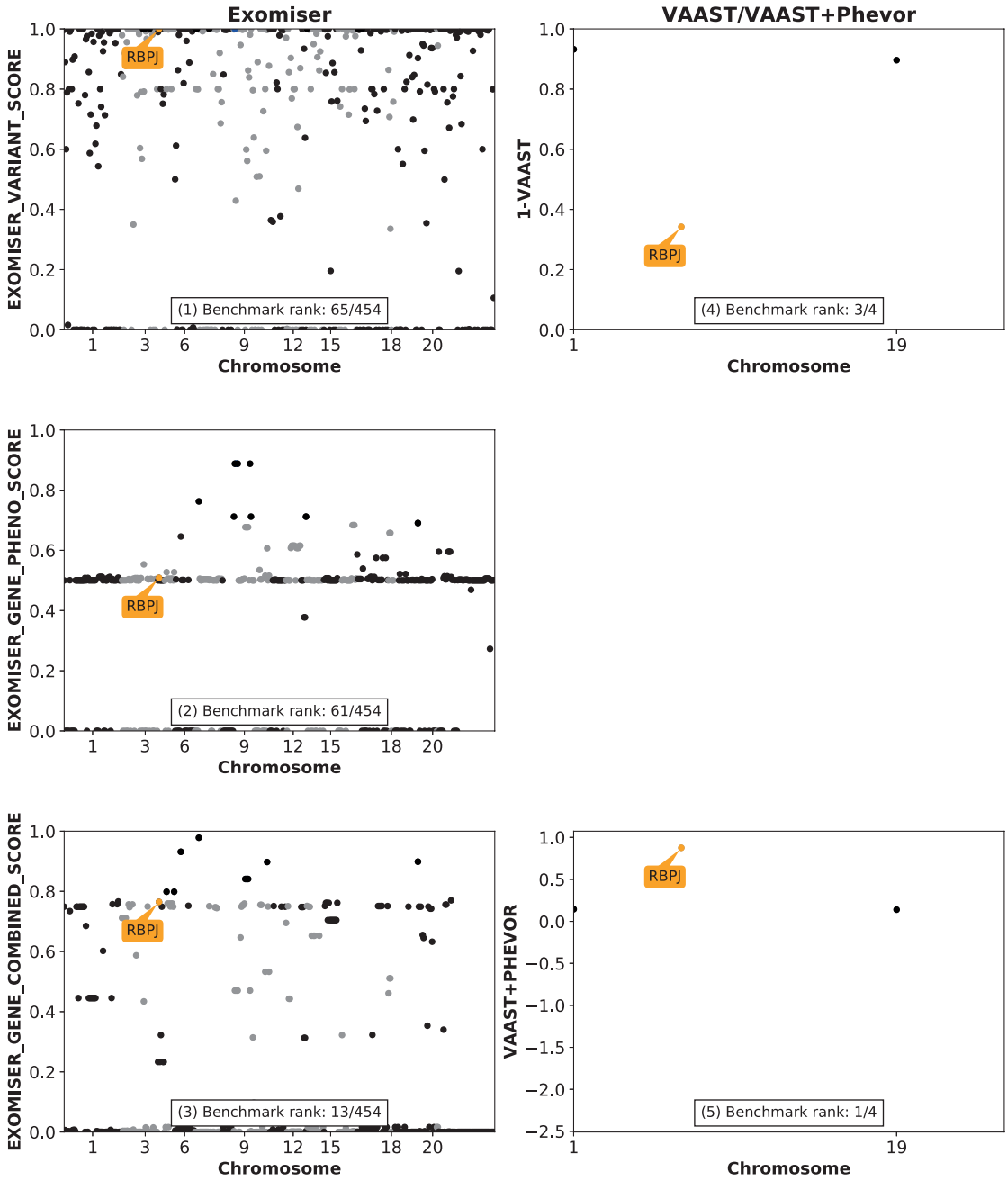


Fig. 3.18 Manhattan plots for case 9 (trio, *de novo*): Atypical KTS for *de novo* missense candidate variant in *RBPJ*. Genomic coordinates on x-axes, 'importance' of variants on y-axes (variant score and combined score for Exomiser, $1 - p\_value$ for VAAST and the Phevor score for VAAST+Phevor). The benchmark variant is highlighted orange. (1) Exomiser's variant score ranks *RBPJ* in 65th place. (2) Exomiser's phenotype score ranks *RBPJ* in 61st place, defaulting to a score of 0.5 via the PPI due to missing HPO and MPO annotations. (3) Exomiser's combined score ranks *RBPJ* 13th. (4) VAAST ranks the benchmark variant third out of just four candidate variants. (5) VAAST+Phevor successfully ranks *RBPJ* in first place.

3.3.2.2.2 Case 10: Fatal acute encephalitis with X-linked missense variant in *DOCK11*

3.3.2.2.2.1 Phenotype This patient suffered from fatal acute encephalitis, an inflammation of the brain thought to be rooted in an immunodeficiency due to the patient's medical history. The phenotype includes encephalitis, lymphadenopathy, and abnormalities of the spleen as well as the bone marrow cell morphology. The proband and both parents were sequenced.

3.3.2.2.2.2 Candidate gene The lead candidate is a hemizygous X-linked missense variant in the Dedicator Of Cytokinesis 11 (*DOCK11*)[14] gene, which is a novel candidate for this phenotype hypothesised to play a role in the immune system due to its high expression in peripheral blood leukocytes [209]. The benchmark variant was predicted to be pathogenic by Polyphen2 (score: 0.999), SIFT (score: 0.005), and MutationTaster (score: 1.0) and was rare in gnomAD (0.0094%) and TOPMed (0.0048%). At the time of the analysis, the variant had not been definitively confirmed as disease-causing. To test the performance of the algorithms on a novel gene combined with the X-linked recessive inheritance mode, this case was selected (see Figure 3.19).

3.3.2.2.2.3 Analysis Exomiser's variant score ranks *DOCK11* in 13th place with a fairly high pathogenicity score, while the benchmark variant ranks third for VAAST. Exomiser's phenotype score shows clustering at 0.5, suggesting that no HPO annotations were available for most variants on the X chromosome that passed filtering for this patient (see Section 3.3.1.2 for a discussion of Exomiser's phenotype score's clustering at 0.5). However, despite the clustering of the phenotype score, the information gained through the HPO helps with the identification of the disease-

[14] c.1679C>T [NM_144658.3], p.Ser560Leu

causing variant. *ATM*, which is proximate to *DOCK11* in Exomiser's PPI, is annotated with 'small lymph nodes' (MP:0002217) and 'small spleen' (MP:0000692) in the MPO, which are matches for the patient's phenotypic terms 'lymphadenopathy' and 'abnormality of the spleen'. Furthermore, *ATM* is annotated with Mantle cell lymphoma in the HPO, a condition with a partial phenotypic overlap with fatal acute encephalitis. The phenotype of Mantle cell lymphoma includes 'lymphadenopathy', 'splenomegaly', and 'abnormality of bone marrow cell morphology', which partially match the patient's phenotypic terms 'lymphadenopathy', 'abnormality of the spleen', and 'abnormality of bone marrow cell morphology'. The proximity to *ATM* in the PPI results in a phenotype score of 0.5 for *DOCK11* by Exomiser. Consequently, Exomiser's combined score ranks the *DOCK11* variant second. VAAST+Phevor benefits from a similar effect, ranking the benchmark variant first.

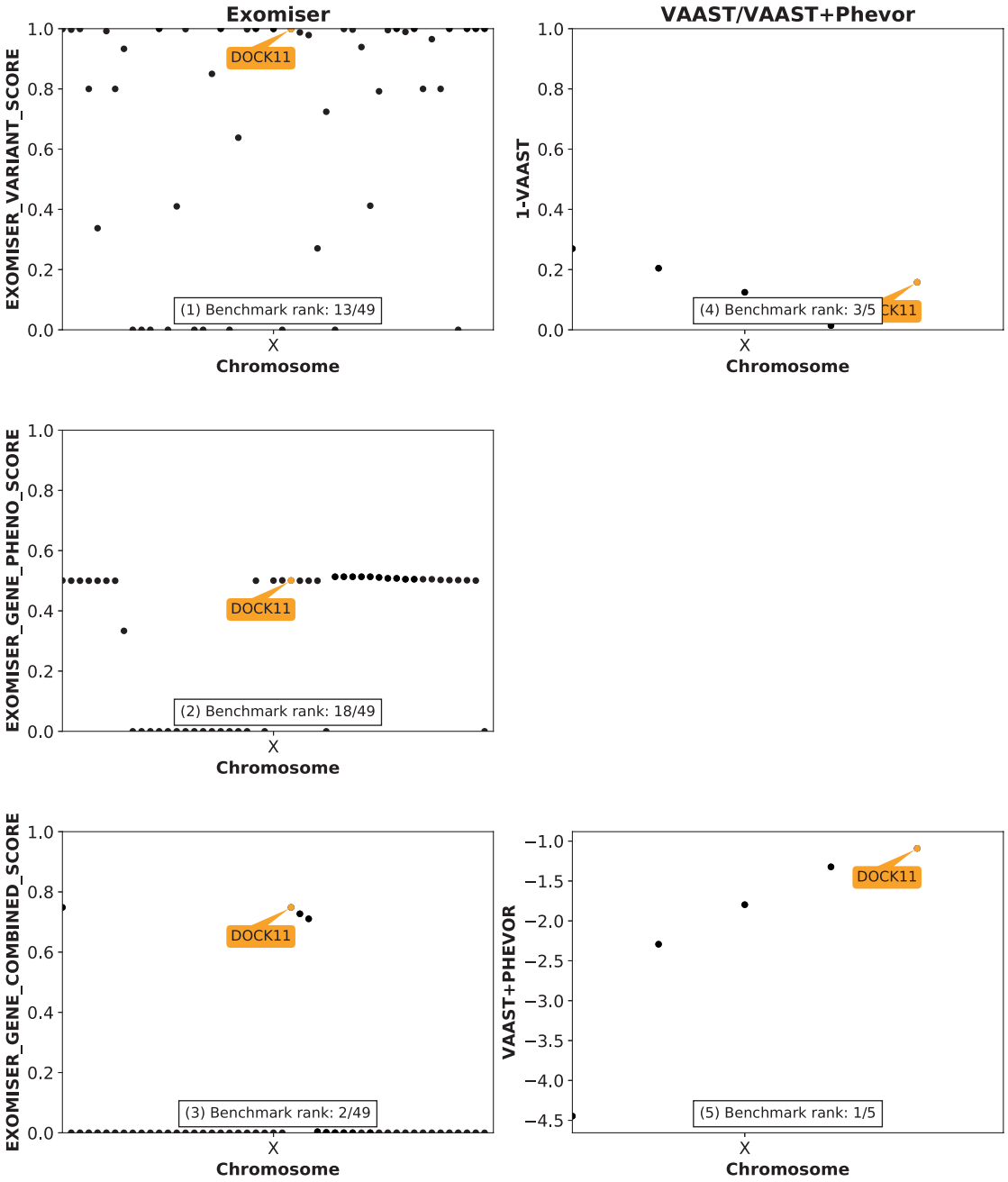


Fig. 3.19 Manhattan plots for case 10 (trio, X-linked): Fatal acute encephalitis for X-linked heterozygous variant in *DOCK11*. Genomic coordinates on x-axes, 'importance' of variants on y-axes (variant score and combined score for Exomiser, $1 - p\_value$ for VAAST and the Phevor score for VAAST+Phevor). The benchmark variant is highlighted orange. (1) Exomiser's variant score ranks *DOCK11* in 13th place. (2) Exomiser's phenotype score ranks *DOCK11* in 18th place, defaulting to a score of 0.5 via the PPI due to missing HPO and MPO annotations. (3) Exomiser's combined score ranks *DOCK11* 2nd. (4) VAAST ranks the benchmark variant third out of just five candidate variants. (5) VAAST+Phevor successfully ranks *DOCK11* in first place, just ahead of *TEX11*, a gene with a poor functional fit associated with infertility in men [210–212].

3.3.2.2.3 Case 11: FLS with splice-site LoF variant in *HDLBP*

3.3.2.2.3.1 Phenotype The patients in case 11 were diagnosed by Dr Usha Kini, a consultant in Clinical Genetics at the Oxford University Hospitals, with FLS, a rare dysmorphic condition previously described in Section 3.3.2.1.7. The family described here consists of five affected first degree cousins from two different branches of the same consanguineous family, two of which were sequenced. The two described patients are phenotypically similar to the described *POR* case, with some additional phenotypic features. In addition to having a narrow mouth and abnormally shaped - or in this case, posteriorly rotated - ears, as well as global developmental delay, an abnormal cornea morphology, a short chin, shallow orbits, plagiocephaly, they also presented with bilateral camptodactyly and a contracture of the proximal interphalangeal joint of the 5th finger.

FLS is very rare. On October 26, 2018, there were only three cases contained in the 100,000 Genomes Project with homozygous variants in *HDLBP*. Two cases presented with intellectual disability, while the third case presented with epilepsy. None of the cases had any dismorphic or abnormal skeletal features. Thus, the GEL database did not contain other matches for this case. Equally, the GEL database was searched for the most defining phenotypic features of FLS. The search did not produce any matches for our case.

3.3.2.2.3.2 Candidate gene All five affected patients were genotyped using cytoSNP12 arrays, which identified a single 5 Mb region on 2q37.3 shared in all five subjects. The previously mentioned cousin pair was subsequently whole genome sequenced and two candidate pathogenic variants were identified in the LOH region on chromosome two. The first candidate is a missense variant in Histone Deacetylase 4 (*HDAC4*)[15]. *HDAC4* is a known candidate gene for 2q37 deletion syndrome, a

[15] p.E374K

condition with a phenotype that partially, but incompletely, overlaps with FLS. A second candidate - a loss-of-function variant - was identified in *HDLBP*[16]. The benchmark variant is predicted to be pathogenic by MutationTaster (score: 1.0) and was not reported in gnomAD and TOPMed. The identified splice site variant leads to skipping of exon 14, which is part of the RNA-binding KH6 domain. Previous studies have showed that *HDLBP* lies in regions that are deleted in several patients affected by brachydactyly mental retardation syndrome [156]. Chapter 6 contains a detailed analysis of this case. At the time this benchmark study was conducted, *HDLBP* was considered as a novel candidate for FLS.

3.3.2.2.3.3 Analysis In addition to the standard filtering stages (see Section 3.2), the VCF was filtered to only include variants present in the genomes of both affected cousins. Due to that stringent filtering protocol, the final VCF only contains seven variants for Exomiser and five variants for VAAST+Phevor to rank. Since the benchmark variant causes skipping of exon 14, both Exomiser's variant score and VAAST predict the variant to be highly pathogenic and rank the variant first (see Figure 3.20). At the time of analysis, *HDLBP* had not been described in the context of a human phenotype and thus was not annotated in the HPO. In the absence of HPO annotations, Exomiser's phenotype score defaults to the PPI, in which *HDLBP* is connected with *PLOD3* via the interactome. Due to the PPI proximity, Exomiser's phenotype VPA assigns a score of 0.5 to the benchmark variant. *PLOD3* is annotated with lysyl hydroxylase 3 deficiency [202]. The condition's phenotype partially overlaps with the patients' phenotype. Importantly, patients affected by lysyl hydroxylase 3 deficiency have cataracts, while our patients presented with megalocornea, as well as low-set ears, similar to our patients. Patients described with deleterious variants in *HDAC4* do not share those features. Additionally, the patient's phenotypic terms 'uplifted earlobe', 'short nose', 'shallow orbits', 'severe global developmental delay', 'hypertelorism',

[16] c.1731+1G>A [NM_203346.4]

'contracture of the proximal interphalangeal joint of the 5th finger', and 'bilateral camptodactyly' are matched with the terms 'abnormality of the pinna', 'short nose', 'shallow orbits', 'global developmental delay', and 'elbow flexion contracture' for lysyl hydroxylase 3 deficiency. As a result of the two component scores, Exomiser's combined score ranks *HDLBP* first, ahead of *HDAC4*.

Similar to Exomiser's phenotype score, Phevor leverages additional ontologies to compensate for the lack of HPO annotations. The final VAAST+Phevor output ranks *HDLBP* first. The high phenotype-related ranking stems from *HDLBP*'s annotations in the gene ontology.

Importantly, *HDLBP* was not ranked high in a previous analysis of the case conducted with an older version of Exomiser (version 7.2.1, released on January 5th, 2016). Since the first analysis of this case, allele frequency databases were updated, resulting in only seven variants having to be ranked by Exomiser, as opposed to 77 previously. Additionally, the interactome link between *HDLBP* and *PLOD3* had not been established yet, which is why *HDLBP* was assigned a phenotype score of 0. This case thus highlights the importance of continuously re-analysing rare genetic disease patient cohorts, even in the absence of new findings related to the human phenotype in question.

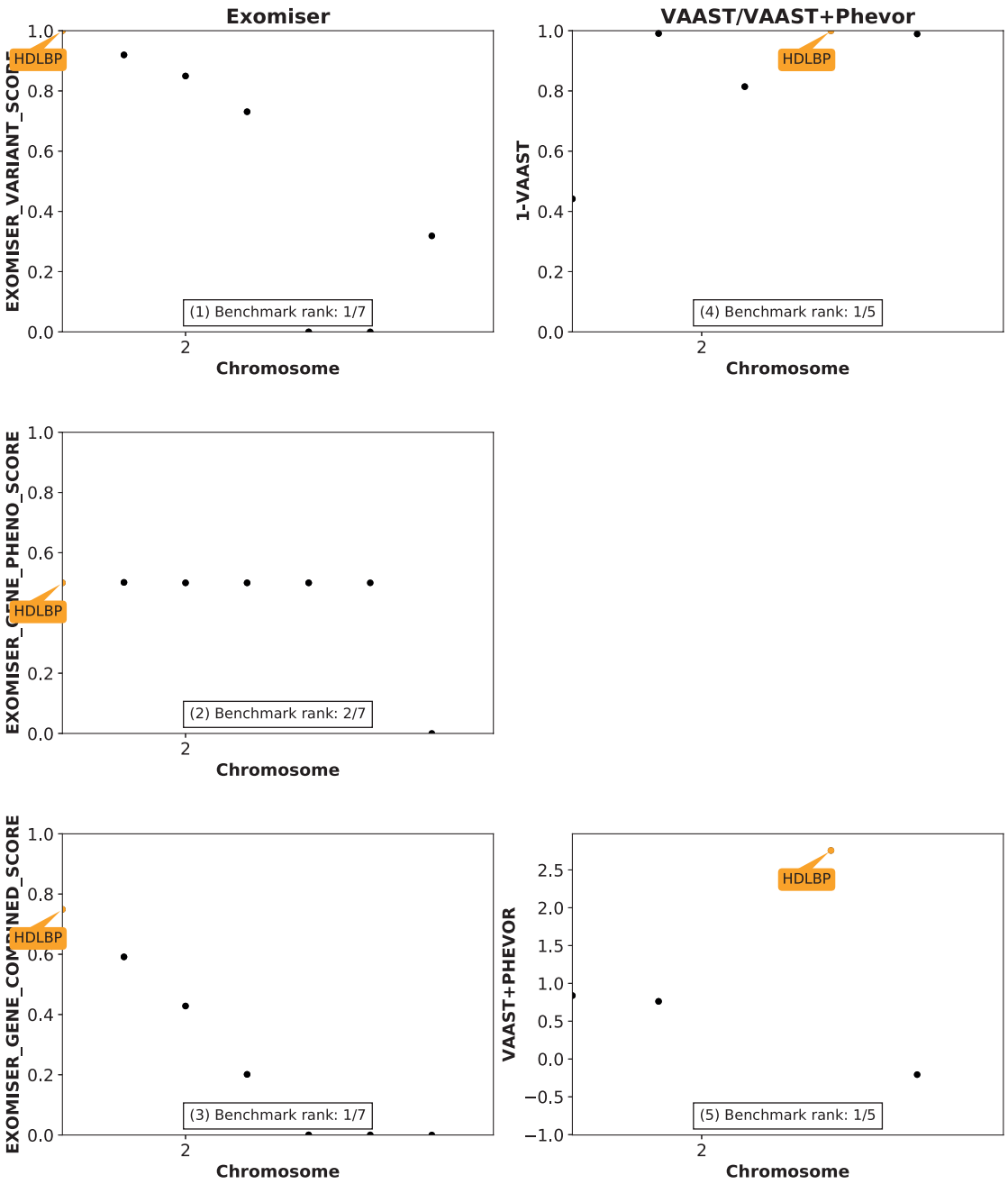


Fig. 3.20 Manhattan plots for case 11 (cousins, AR): FLS with homozygous LoF variant in *HDLBP*. Genomic coordinates on x-axes, 'importance' of variants on y-axes (variant score and combined score for Exomiser, $1 - p\_value$ for VAAST and the Phevor score for VAAST+Phevor). The benchmark variant is highlighted orange. (1) Exomiser's variant score ranks *HDLBP* in 1st place. (2) Exomiser's phenotype score ranks *HDLBP* in 2nd place, defaulting to a score of 0.5 via the PPI due to missing HPO and MPO annotations. (3) Exomiser's combined score ranks *HDLBP* 1st. (4) VAAST ranks the benchmark variant first out of just five candidate variants. (5) VAAST+Phevor successfully ranks *HDLBP* in first place.

3.4 Discussion

In this chapter, I analysed the performance of algorithms used to prioritise variants from whole genome sequencing data using genotypic and phenotypic information. The two VPA frameworks initially used by the 100,000 Genomes Project for their case analyses, Exomiser and VAAST+Phevor, were selected for the analysis. Both frameworks, at a high level, rely on genotypic and phenotypic data for their prioritisation. The goal of the analysis was two-fold: comparing the performance of Exomiser with VAAST+Phevor and comparing the performance of algorithms not using phenotypic data with algorithms that use phenotypic data. To that end, several analyses were conducted based on eleven rare genetic disease patient cases from the HICF2 study for which the disease-causing variant had previously been identified. For eight of the patient cases, the confirmed disease-causing variant lies in a gene known for the patient's phenotype and those genes were thus annotated in the HPO. The remaining three genes are novel candidates for the observed phenotypes. As such, this analysis not only provides an important independent benchmark analysis of two of the main analysis frameworks in the field for known gene candidates, it also assesses the frameworks' performance for novel gene discovery on non-simulated data.

3.4.1 Discussion of summary results

First, a summary analysis of all eleven cases was conducted to assess the ability of each algorithm to (1) rank the disease-causing benchmark variant at the top of a list of all variants and (2) decrease the number of non-disease-causing variants that receive a high score.

3.4.1.1 Ranking performance

All analysed algorithms successfully ranked a significant number of variants at the top of their respective distributions. Overall, the GPAs outperformed the GAs.

Exomiser's combined score ranked five variants first (≈45%) and six in the top five (≈45%), compared to four variants ranked first (≈36%) and four in the top five (≈36%) for Exomiser's variant score and one ranked first (≈9%) and two in the top five (≈18%) for Exomiser's phenotype score. This analysis falls short of other evaluations of Exomiser on real patient cohorts, including the 74% of candidate variants ranked first and 94% of variants ranked in the top five observed by Cipriani *et al.* [119], or the 80% top five-ranked variants seen by Koehler *et al.* [99], but is still aligned with findings of other groups showing that the best ranking performance is achieved by algorithms using genotypic and phenotypic information in their ranking process [173, 58, 113].

A similar trend is observed with VAAST and VAAST+Phevor. VAAST ranks one variant first (≈9%) and eight variants in the top five (≈73%), as compared to VAAST+Phevor with eight variants ranked first (≈73%) and nine in the top five (≈82%).

Overall, VAAST+Phevor ranks three more variants first and three more in the top five than Exomiser's combined score, but a larger sample size would be required to make definitive statements about how their ability to rank disease-causing variants at the top of a prioritised list differs.

In addition to comparing the most recent releases of Exomiser and VAAST+Phevor, I conducted a time series analysis of two different versions of Exomiser to contrast how algorithm performance changed over time. The newer version of Exomiser was able to rank more benchmark variants first, in the top five, and top ten and the older version of Exomiser only matched the total number of benchmark variants ranked in the top

20. Importantly, the older version of Exomiser successfully prioritised a known gene, *POR* in the top 20 that was missed by the newer version. Simultaneously, the newer release of Exomiser ranked a benchmark variant in *HDLBP*, a novel candidate gene for FLS, first that was still missed by the older version of the algorithm. Both of these effects are driven by the fact that databases drawn upon by the analysis frameworks are growing. *POR* was not ranked in the top 20 anymore because the specificity of its phenotypic annotations relative to other ranked candidate genes was less pronounced. This is also evident in the fact that, counterintuitively, the older version of Exomiser's phenotype score successfully ranks more benchmark variants first, in the top five and the top ten than the newer version of the same algorithm. The more genes are annotated with HPO terms, the more gene candidates will likely look relevant for certain phenotypes. Between the two versions of Exomiser used for this analysis, the number of annotations per term in the HPO has risen from ≈5.3 in 2010 (9,500 terms and 50,000 annotations [97]) to 12 in 2019 (13,000 terms and 156,000 annotations [179]). While an increasing number of annotations for each term means a more holistic phenotypic representation, it also presents new challenges for distinguishing signal from noise. At the same time, the expansion of the PPI used by Exomiser led to *HDLBP* being successfully ranked in Exomiser's latest release, in contrast to the older version.

Continuous reanalysis of RGD cohorts is thus highly important to surface new disease insights for patients, such as demonstrated here for the FLS case, as has been confirmed by several studies [213, 214]. In particular, large scale academic population studies, such as the 100,000 Genomes Project, that are funded by finite governmental and academic research grants, should plan for the continuous reanalysis of their data to ensure patients participating in trials benefit in the long run. WGS studies of RGDs still only accomplish diagnostic yields of up to 40% [38, 215] and thus, while not all of the remaining 60% of patients will get diagnostic results over time, some might.

3.4.1.2 Variant distribution

While GAs successfully rank a significant number of benchmark variants high, they do so for a large number of variants, making analyses time-consuming and difficult. To illustrate this effect, I analysed the percentage of variants ranked by Exomiser and VAAST+Phevor respectively that receive a significant score by the algorithms, thus necessitating further investigation by analysts. Both GPA frameworks meaningfully decrease the number of variants receiving a significant score, from 32.3% to 2.2% and from 25.4% to 11.2% for Exomiser and VAAST+Phevor respectively. Algorithms using phenotypic information are thus not only useful for ranking disease-causing variants at the top of a sorted variant list, they also meaningfully downgrade variants that are predicted to be pathogenic based on their genotype, but carry no relevance for the patient's phenotype. In practice, this is highly relevant as analysts spend the majority of their time closely evaluating all significantly highly ranked variants, first in silico with additional tools and literature research and later via *in vitro* or *in vivo* studies. Therefore, minimising the number of candidate variants is critical in time-sensitive clinical settings. It is worth noting that determining significance is easier for p-value based algorithms like VAAST and Bayesian approaches like Phevor than for Exomiser without a natural cut-off value. To my knowledge, this is the first study highlighting this effect.

3.4.2 Future directions

As demonstrated in this chapter, GPAs have become an increasingly powerful tool for the identification of disease-causing variants from whole genome sequencing data for rare genetic disease cases. Beyond academic research cohorts, these algorithms are already successfully being used in care settings where rapid, accurate analyses are crucial, as demonstrated by Clark *et al.* [216] in a study illustrating the use of

VAAST+Phevor to deliver molecular diagnoses for seriously ill children with genetic diseases based on WGS data and HPO terms in neonatal and pediatric intensive care units within just ≈20 hours. The increased adoption of GPAs in academic and non-academic settings will also drive the adoption of methods for the large-scale collection of deep phenotypic data, for example natural language processing approaches, as highlighted by Clark *et al* [216]. In addition to rapid variant ranking, GPAs also make result interpretation easier, as HPO annotations facilitate faster literature research by linking to relevant publications. Furthermore, the HPO enables automated differential diagnoses by prioritising phenotypically-relevant variants that would have gone unnoticed with GAs.

Despite the rapid improvement of GPAs and the growth of associated databases, the diagnostic yield for rare genetic diseases remains approximately 40%. Furthermore, as discussed in Section 3.4.1.1, the increasing number of genes being annotated with HPO terms poses a challenge for algorithms leveraging the HPO for gene prioritisation. This challenge will likely become more pronounced with the increase in scientific knowledge through projects like the 100,000 Genomes Project.

To continuously improve the diagnostic yield for rare genetic diseases, several developments will likely be observed in the future.

Resources like the HPO will have to be leveraged more effectively. Phenotypes in the HPO will be annotated with term-specific frequencies to illustrate the percentage of patients with a certain disease that present with a specific phenotype known for that disease. The Orphanet consortium has started to collect such data [114]. Consequently, GPAs will adapt to make use of phenotype frequencies. Moreover, phenotypic data collection tools such as Phenotips provide an option for analysts to indicate that a specific phenotype was not observed in a patient, but frameworks that make use of this information yet have to be developed. In addition to that, annotations in the HPO might become more granular than the gene-level. Instead, one could annotate specific exons

or functional blocks of a gene with phenotypic terms that are observed in patients with variants in those sections of the gene.

Separately from HPO-related developments, GPAs will evolve to better account for disease-causing intronic variation. While all eleven benchmark variants in this study lie in coding regions, novel GPAs to extend analyses to the non-coding genome have recently been published, the clinical validation of which remains difficult due to the lack of a patient "truth dataset" [120, 217].

In addition to better ways of leveraging the HPO and improved algorithms, GPAs will evolve to incorporate more independent data sources that can increase the accuracy of VPA. Exomiser and VAAST+Phevor already make use of the MPO, PPI, the GO and other databases, which proved particularly useful for novel gene discovery, as demonstrated for the case of *HDLBP*. Large-scale experimental databases, such as tissue-specific gene expression databases, present an opportunity to improve VPA, as illustrated in the following chapter.

Chapter 4

Value of tissue-specific gene expression data for variant prioritisation – a novel algorithm

4.1 Introduction

VPA are widely used for the analysis of WGS data from RD patients. Established analysis frameworks used by studies like the 100,000 Genomes Project combine genomic data and phenotypic terms in the HPO format to rank variants. One widely used framework, Exomiser, relies on an algorithm called hiPHIVE for variant prioritisation [119]. hiPHIVE calculates a combined score for each candidate variant based on a variant score and a phenotype score. The variant score is a multiplication of a minor allele frequency score and a pathogenicity score. For missense variants, the pathogenicity score is calculated as the maximum of a variant's MutationTaster [57], Polyphen-2 [82] or SIFT score [21], scaled from 0 to 1. For other classes of variants, fixed pathogenicity scores were assigned by the authors. The phenotype score, scaled from 0 to 1, is a combination of the information content of a patient's

phenotypic profile and the semantic similarity of the patient's phenotypic profile and each candidate gene's phenotypic annotations. The gene's annotations are based on human, mouse and zebrafish phenotypic data, or the phenotypic annotations of genes closely related to the candidate gene in the protein interactome. A logistic regression classifier was trained to calculate the combined hiPHIVE score based on the variant and pathogenicity score. For a detailed description of hiPHIVE, see Section 1.2.2.3.1.

While frameworks like Exomiser are effective for the prioritisation of genes that are known for the patient's phenotype, the algorithms perform less well for the identification of candidate variants in genes that are novel for the phenotype in question (see Section 3.3.1.1.1). As demonstrated in Chapter 3, when the databases hiPHIVE relies on for information on novel disease genes do not contain any annotations for the correct candidate gene, candidate variant prioritisation is challenging for the algorithms.

Other data sources exist that can be used for variant prioritisation when genotype- and phenotype-based annotations do not contain sufficient information to prioritise the correct variant. Expression data has previously been shown to be valuable for disease gene prioritisation [139–141] (see Section 1.2.3 for details).

In particular, Deelen *et al.* [142] showed with the GADO algorithm that gene expression data can be used to prioritise genes based on gene co-regulation. The authors compared GADO with Exomiser on a WES cohort of 83 patients with confirmed diagnoses. GADO achieved a lower median rank for the candidate variants than Exomiser (12.5 compared to 21), but Exomiser ranked more candidate variants in the top three (28 compared to 14). For a WES cohort of 61 patients without established molecular diagnoses, GADO yielded a potentially disease-causing candidate variant for ten cases (see Section 1.2.3.3.4 for a detailed description of GADO). GADO was published after the work described in this chapter was concluded.

While Deelen *et al.* [142] used gene expression data irrespective of the tissue of origin, Feiglin *et al.* [132] showed that genes carrying disease-causing variants are on

average more highly expressed in tissues that are likely affected based on the disease's HPO annotations (see Section 1.2.3.4). Feiglin *et al.*'s [132] findings suggest that gene expression data in tissues indicated as affected based on HPO terms could be an informative input for variant prioritisation, including for candidate variants in genes lacking HPO annotations.

The goal of this chapter is thus to assess the value of tissue-specific expression data combined with HPO terms for the ranking of likely disease-causing variants. Furthermore, I propose a novel machine learning VPA based on genotype, phenotype, expression, and tissue data (GPET). HPO terms are used to identify tissues that are likely affected by disease. Information on likely affected tissues is then used in conjunction with genomic, phenotypic and tissue-specific gene expression data to inform variant classification.

4.2 Materials and methods

In this section, I describe the GPET algorithm's concept (see Section 4.2.1), provide details about the construction of the training dataset (see Section 4.2.2), illustrate the underlying machine learning methods (see Section 4.2.3), and characterise different scenarios modelled for testing (see Section 4.2.4),

4.2.1 Concept of the algorithm

The algorithm presented here combines the two component features of the Exomiser hiPHIVE framework - the Exomiser variant score and the Exomiser phenotype score - with a set of tissue-specific expression features. Each feature class represents an independent type of information used for variant ranking: genoytpic information, phenotype, and tissue-specific expression. The individual features are combined into

one final output score for variant ranking using a machine learning classifier called a 'Random Forest'. Figure 4.1 illustrates that concept.

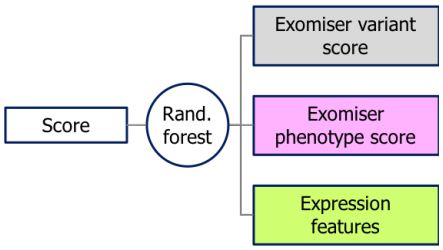


Fig. 4.1 **The GPET algorithm framework.** GPET uses genotypic, phenotypic, and expression data to prioritise variants from WGS data. Genotypic data is represented by Exomiser's variant score (grey), phenotypic data by Exomiser's phenotype score (pink), and expression data by a set of expression features (green) described in detail in Section 4.2.2.3. The algorithm employs a random forest classifier to combine the individual feature classes into one final output score.

4.2.2 Training dataset

To create the above described algorithm, I first created a labeled training dataset consisting of disease-causing and non-disease-causing variants annotated with Exomiser's two scores and a new concept I introduce here called 'tissue-specific expression features'. Section 4.2.2.1 describes how the disease-causing and non-disease-causing variants were curated. Section 4.2.2.2 explains how Exomiser's variant and phenotype scores were calculated for each variant in the dataset. In Section 4.2.2.3, the concept of tissue-specific expression features is introduced and the overall dataset architecture that reflects the framework shown in Figure 4.1 is described.

4.2.2.1 Curation of a large variant database

As a basis for the training dataset, a large number of variants that are known to be disease-causing as well as a large number of likely non-disease-causing variants were needed. To create the dataset, the SWISSProt database was used [218, 219], curated by Rogers *et al.* [220] for their FATHMM algorithm. The original SWISSProt database

only contained gene names, Swiss-Prot AC codes, FTIds, amino acid changes, a label for the type of variant, dbSNP IDs and OMIM IDs. VEP version 90 was used to annotate variants with variant coordinates. HPO terms were extracted from OMIM for each OMIM:ID. The database originally consisted of 24,646 variants labeled as disease-causing, 37,931 polymorphisms not classified as disease-causing, and 6,564 unclassified variants. In the SWISSProt database, variants that had previously been reported in OMIM as disease-causing for specific patient cases were labeled as disease-causing. Rogers *et al.* [220] selected variants from the 1,000 Genomes Project that were predicted to be benign as non-disease causing variants. After filtering out unclassified variants, variants without an associated OMIM:ID and variants that could not be annotated by VEP, 50,514 variants remained in the final dataset. Of those variants, 14,929 are labeled as disease-causing and 35,585 as non-disease-causing (see Figure 4.2).

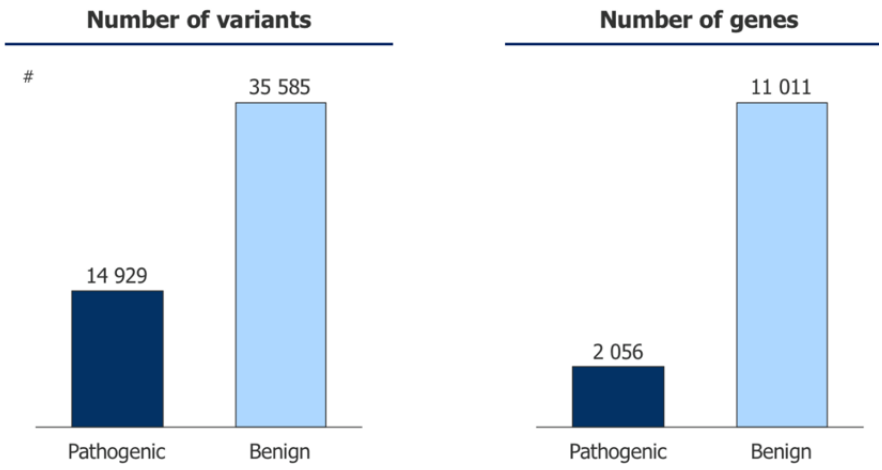


Fig. 4.2 **Overview of the curated list of variants used to create the training dataset.** The dataset used for training and testing of the machine learning algorithm consists of 50,514 variants, 14,929 of which are labeled as disease-causing according to OMIM and 35,585 of which are considered benign as drawn from the 1,000 Genomes project. The disease-causing variants are distributed across 2,056 different genes, while the benign variants sit in 11,011 different genes.

4.2.2.2 Calculation of Exomiser's hiPHIVE scores for variants in training dataset

To calculate Exomiser's hiPHIVE variant score and phenotype score for each variant in the dataset, 14,929 individual VCF files were created. Each VCF file contained one disease-causing variant and 3-4 randomly sampled non-disease-causing variants from the SWISSProt dataset. 5,727 VCF files contained four non-disease-causing variants and 9,202 VCF files contained three non-disease-causing variants. The number of non-disease-causing variants per VCF varied since the SWISSProt dataset contained 3-4 times more non-disease-causing variants than disease-causing variants (see Figure 4.2). Exomiser takes a VCF for a patient as well as the associated HPO terms for the patient as input. Thus, each simulated VCF was treated like a pseudo-patient, where the HPO terms associated with the contained disease-causing variant were used for the overall VCF, in order to be able to also assign a phenotype score to non-disease-causing variants, which otherwise would not have HPO terms associated with them and thus no phenotype score could be calculated. Exomiser's hiPHIVE (v7.2.1) was run on each VCF. The inheritance mode was set to autosomal dominant (AD). A similar approach was used by Bone *et al.* [118] to train Exomiser. After Exomiser was run, the variants from each individual VCF were combined into a large dataset used for further analysis.

4.2.2.3 Tissue-specific expression features

Next, expression features were created for each variant in the dataset. As described in Section 4.2.2.2, each variant was assigned HPO terms based on the VCF file the variant was allocated to. Those HPO terms are used here to calculate expression scores for every variant. The expression features consist of two different feature classes as illustrated in the framework in Figure 4.3, which was developed by Feiglin *et al.* [132]. The first expression feature class contains binary information on whether or not a tissue is predicted to be affected by the disease in question based on the supplied

phenotype profile in HPO format. Hereafter, those features are referred to as 'Binary Tissue Labels' (BTLs). For example, the term 'Abnormality of brain morphology' (HP:0012443) maps to the tissue 'Brain'. Thus, the BTL's value for the gene in question and for the tissue 'Brain' is '1'. If no terms map to 'Brain', the value '0' is assigned. Table 4.1 shows a mapping of 33 high-level HPO categories to 25 broad tissues created by Feiglin *et al.* [132].

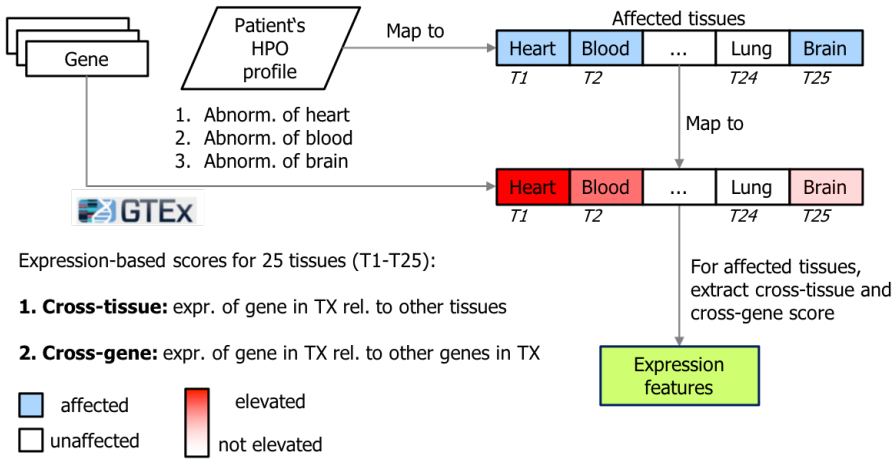


Fig. 4.3 **Mapping of tissues and HPO terms to expression values.** Affected tissues are inferred via HPO terms annotated to each gene in the dataset. For example, if a gene is annotated with the HPO term 'Abnormality of the heart' (HP:0001627), this suggests that the heart tissue is affected by that gene. Information on whether or not a tissue is affected by a gene is stored in one feature per tissue containing a binary tissue label (BTL) for each gene. In the figure, affected tissues are highlighted in light-blue and are assigned the value '1', whereas unaffected tissues are coloured white and carry the value '0'. Simultaneously, tissue-specific expression scores (later referred to as [Rawexpression]) are calculated for each tissue and gene using the GTEx database [132]. In this figure, elevated gene expression is indicated by a red highlight, whereas lower gene expression is indicated by a white background. Tissue-specific gene expression is captured in three different scores: a 'cross-gene' score, which represents the expression of a gene in a tissue relative to all other in genes in that tissue, as well as two distinct calculation methods to compute a 'cross-tissue' score, which indicates the relative expression of a gene in one specific tissue relative to that gene's expression in all other tissues. Collectively, this framework is used to generate a total of 100 expression features, highlighted in green in the figure: a total of four expression features per tissue (BTL, one cross-gene score and two cross-tissue scores) for each of 25 tissues. This figure is adapted from Feiglin *et al.* [132].

Broad tissue	HPO Term
Adipose Tissue	Abnormality of adipose tissue
Adrenal Gland	Abnormality of the adrenal glands
Blood	Abnormality of blood and blood-forming tissues
Blood Vessel (artery)	Abnormality of the systemic arterial tree
Brain	Abnormality of brain morphology
Brain	Abnormality of nervous system physiology
Breast	Abnormality of the breast
Colon	Abnormality of the large intestine
Esophagus	Abnormality of the esophagus
Heart	Abnormal heart morphology
Heart	Congestive heart failure
Heart	Arrhythmia
Heart	Heart murmur
Heart	Cardiac shunt
Heart	Abnormal EKG
Heart	Cardiogenic shock
Liver	Abnormality of the liver
Lung	Abnormality of the lung
Nerve	Abnormality of peripheral nerves
Ovary	Abnormality of the ovary
Pancreas	Abnormality of the pancreas
Pituitary gland	Abnormality of the pituitary gland
Prostate	Abnormality of the prostate
Muscle	Abnormality of muscle morphology
Muscle	Abnormality of muscle physiology
Skin	Abnormality of the skin
Small Intestine	Abnormality of the small intestine
Spleen	Abnormality of the spleen
Stomach	Abnormality of the stomach
Testis	Abnormality of the testis
Thyroid	Abnormality of the thyroid gland
Uterus	Abnormality of the uterus
Vagina	Abnormality of the vagina

Table 4.1 **Mapping of broad tissue categories to high-level HPO term categories** 25 broad tissue categories were mapped to 33 high-level HPO term categories. The mapping is used to create a new set of features called 'binary tissue labels', which indicate whether or not a specific tissue is suspected to be affected by a patient's disease based on the patient's HPO terms. The mapping was adapted from Feiglin *et al.* [132].

If a phenotypic term is not directly contained in the mapping showed in Table 4.1, the algorithm traverses up the branches of the Human Phenotype Ontology to the term's root in the HPO and maps it to the respective tissue. Feiglin *et al.*'s [132] mapping does not include all phenotype categories in the latest HPO release. Thus, terms from HPO branches that do not map to a broad tissue category receive BTLs of 0 for every tissue.

The second expression feature class contains information on the relative expression level of each gene in each tissue. To represent relative expression levels, the GPET algorithm uses two different scores calculated by Feiglin *et al.* [132] based on the GTEx database (v6). The first score represents the expression of a candidate gene in the tissue in question relative to the expression of that gene in all other tissues. The metric is called the 'cross-tissue' score. The second score represents the expression level of a candidate gene in a tissue in question relative to all other genes in that same tissue. The metric is called the 'cross-gene score'.

To calculate the cross-gene and cross-tissue scores, Feiglin *et al.* [132] used gene level reads per kilobase per million mapped (RPKM) data collected by the GTEx consortium using RNA sequencing. Feiglin *et al.* used version 6 files[1] containing data on 8,555 GTEx samples. RPKM values are indicative of gene expression. The analysts included a total of 7,051 samples in the analysis, which were also included in the expression Quantitative Trait Loci (eQTL) analysis[2] by the GTEx consortium. The files can be dowloaded from www.gtexportal.org. Feiglin *et al.* [132] further filtered out cell line samples, resulting in 6,665 samples across 51 specific tissue types.

The analysis was limited to 19,644 genes defined as protein coding in a patched version of GENCODE v19[3]. For each gene and tissue, Feiglin *et al.* [132] calcu-

[1] file: GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct
[2] file: GTEx_Analysis_V6_eQTLInputFiles_ geneLevelNormalizedExpressionMatrices.tar.gz
[3] file: gencode.v19.genes.patched_contigs.gtf.gz

lated the mean expression level of all samples available for that specific gene-tissue combination.

For the analysis, the authors transformed mean RPKM values, t, to a log scale as follows:

$$t = log_{10}(RPKM + 1) \tag{4.1}$$

The analysts subsequently performed quantile normalisation for each tissue using the R package *preprocessCore*.

The cross-gene score for a gene in a specific tissue is thus simply t:

$$s_{cg}(gene) = t \tag{4.2}$$

To calculate the cross-tissue score, for each gene Feiglin *et al.* [132] compared that gene's expression from a single tissue with all other tissues using the Wilcoxon rank sum test from the standard stats package in R. This process was repeated once for each tissue. To determine the direction of the difference, the analysts set the alternative hypothesis to 'greater' for upregulation and 'less' for downregulation. The negative log10 p-values from the test were used as the measure of cross-tissue expression.

To test the robustness of the Wilcoxon rank sum test, the authors also computed limma-voom p-values using the 'voom' function of the R 'limma' package and extracted the p-values. P-values of genes with negative fold changes were set to 1 since upregulation is being investigated. The GTEx raw counts matrix was used as input for the limma-voom calculation[4].

For my analysis, I used three separate output files I received from Feiglin *et al.* containing the final output results for all three scores: the cross-gene score ('median',

[4]file: (GTEx_Analysis_v6p_RNA-seq_RNA-SeQCv1.1.8_gene_reads.gct

file: gtex.ts.median.csv), the cross-tissue score according to the Wilcoxon rank sum test ('wilcoxp', file: gtex.ts.wilcoxp.csv) and the cross-tissue score according to the limma-voom calculation ('limma', file: gtex.ts.limma.csv).

The raw expression scores received from Feiglin *et al.* [132] contain scores for a total of 51 specific GTEx tissues. In their study however, the authors provide a manually curated mapping from HPO terms to only 25 broad tissue categories. Thus, the 51 specific GTEx tissues have to be mapped to 25 'broad' tissue categories. I used a mapping provided by the authors shown in Table 4.2, which leaves out the specific GTEx tissues 'Bladder', 'Brain - Amygdala', 'Brain - Spinal cord (cervical c-1)', 'Brain - Substantia nigra', 'Cervix - Ectocervix', 'Cervix - Endocervix', 'Fallopian Tube', and 'Minor Salivary Gland'.

Broad tissue	Specific GTEx tissue
Adipose Tissue	Adipose - Subcutaneous
Adipose Tissue	Adipose - Visceral (Omentum)
Adrenal Gland	Adrenal Gland
Blood Vessel	Artery - Aorta
Blood Vessel	Artery - Coronary
Blood Vessel	Artery - Tibial
Brain	Brain - Anterior cingulate cortex (BA24)
Brain	Brain - Caudate (basal ganglia)
Brain	Brain - Cerebellar Hemisphere
Brain	Brain - Cerebellum
Brain	Brain - Cortex
Brain	Brain - Frontal Cortex (BA9)
Brain	Brain - Hippocampus
Brain	Brain - Hypothalamus
Brain	Brain - Nucleus accumbens (basal ganglia)
Brain	Brain - Putamen (basal ganglia)
Breast	Breast - Mammary Tissue
Colon	Colon - Sigmoid
Colon	Colon - Transverse
Esophagus	Esophagus - Gastroesophageal Junction
Esophagus	Esophagus - Mucosa
Esophagus	Esophagus - Muscularis
Heart	Heart - Atrial Appendage
Heart	Heart - Left Ventricle
Liver	Liver
Lung	Lung
Muscle	Muscle - Skeletal
Nerve	Nerve - Tibial
Ovary	Ovary
Pancreas	Pancreas
Pituitary	Pituitary
Prostate	Prostate
Skin	Skin - Not Sun Exposed (Suprapubic)
Skin	Skin - Sun Exposed (Lower leg)
Small Intestine	Small Intestine - Terminal Ileum
Spleen	Spleen
Stomach	Stomach
Testis	Testis
Thyroid	Thyroid
Uterus	Uterus
Vagina	Vagina
Blood	Whole Blood

Table 4.2 **Mapping of specific GTEx tissues to broader tissue categories.** Version 6 of the GTEx expression database, which was used to train this algorithm, provides expression data for 51 specific tissue categories. In order to create the previously introduced tissue-specific expression features, these 51 specific tissue categories had to be mapped to a smaller number of 25 broader tissue categories. The broader tissue categories could in turn be mapped to HPO terms, as described in Table 4.1. This table shows the mapping of 51 specific GTEx tissue categories to 25 broad tissue categories by Feiglin *et al.* [132].

To generate the cross-gene or cross-tissue scores for a gene in a broad tissue, I calculated the mean of all specific tissue scores for that respective gene and tissue. For example, the broad tissue score for each gene expressed in 'Adipose Tissue' is calculated as the mean of the expression scores for the specific tissues 'Adipose - Subcutaneous' and 'Adipose - Visceral (Omentum)'.

Thus, I created four new features for each tissue encapsulating information on tissue-specific gene expression based on the HPO: a BTL, the cross-gene score and two versions of the cross-tissue score (wilcoxp and limma). In total, 25 tissues are contained in the mapping and therefore a total of 100 new features are created. Combined with the two Exomiser scores and a label indicating if a variant is disease-causing or not, each variant thus is assigned 103 features. Figure 4.4 shows the overall dataset architecture.

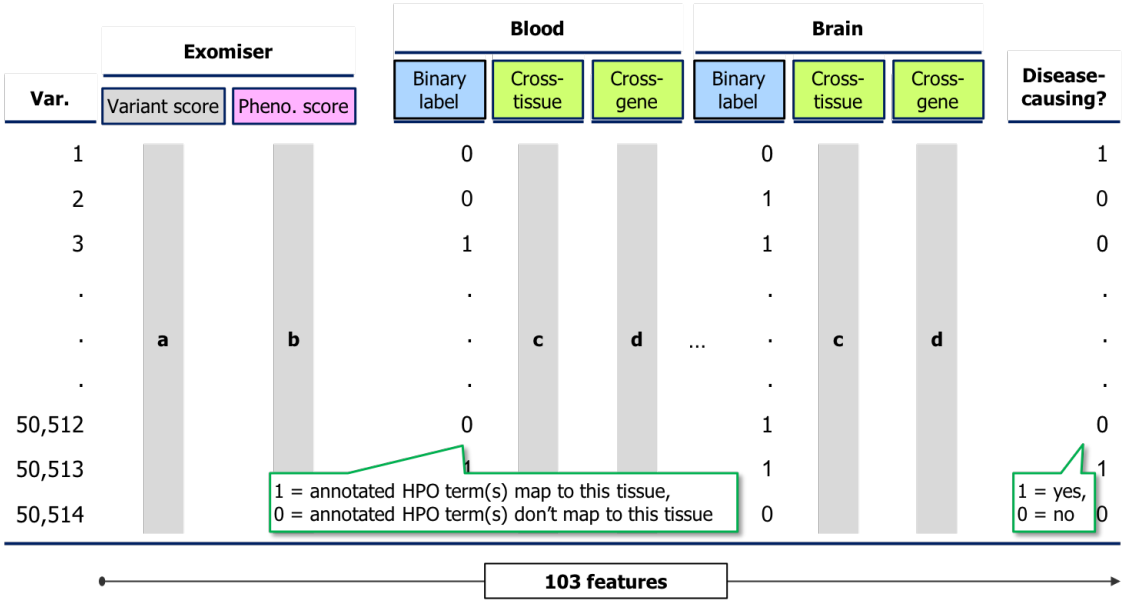


Fig. 4.4 **Overview of the dataset architecture used for training and testing of GPET.** The dataset used for training and testing consisted of 50,514 variants (approximately 30% disease-causing from OMIM with HPO terms, and roughly 70% benign from 1,000 Genomes project) with tissue-specific expression data from GTEx used for training. Each variant is annotated with a total of 103 features: two features for Exomiser's variant and phenotype scores, four expression-specific features (BTL, cross-gene score, and two cross-tissue scores) for 25 tissues respectively, and one feature to label each variant as disease-causing or benign.

4.2.3 Machine learning

The machine learning method was programmed in Python 3.6.4. The scikit-learn library (0.19.1) [221] was used for all machine learning functions. Table 4.3 contains a complete list of all programming libraries used in my python implementation.

Library (1/4)	Version (1/4)	Library (2/4)	Version (2/4)	Library (3/4)	Version (3/4)	Library (4/4)	Version (4/4)
alabaster	0.7.10	html5lib	1.0.1	ncurses	6	qtconsole	4.3.1
alembic	0.9.7	icu	58.2	networkx	2.1	qtpy	1.3.1
asn1crypto	0.24.0	idna	2.6	nodejs	6.12.2	readline	7
astroid	1.6.0	imagesize	0.7.1	numcodecs	0.5.2	requests	2.18.4
babel	2.5.3	intel-openmp	2018.0.0	numexpr	2.6.4	rope	0.10.7
backports	1.0	ipykernel	4.7.0	numpy	1.12.1	scikit-allel	1.1.9
backports.functools$_{lru}$*cache*	1.4	ipython	6.2.1	numpydoc	0.7.0	scikit-learn	0.19.1
bcolz	1.1.2	ipython$_{g}$*enutils*	0.2.0	obonet	0.2.2	scipy	1.0.0
beautifulsoup4	4.6.0	isort	4.2.15	openssl	1.0.2n	seaborn	0.8.1
bleach	2.1.2	jedi	0.11.1	pamela	0.3.0	setuptools	38.4.0
bokeh	0.12.13	jinja2	2.1	pandas	0.22.0	simplegeneric	0.8.1
bottleneck	1.2.1	jpeg	9b	pandoc	1.19.2.1	sip	4.18
bzip2	1.0.6	jsonschema	2.6.0	pandocfilters	1.4.2	six	1.11.0
ca-certificates	2017.08.26	jupyter$_{c}$*lient*	5.2.1	parso	0.1.1	snowballstemmer	1.2.1
certifi	2017.11.5	jupyter$_{c}$*ore*	4.4.0	partd	0.3.8	sortedcontainers	1.5.7
cffi	1.11.4	jupyterhub	0.8.1	patsy	0.5.0	sphinx	1.6.6
chardet	3.0.4	lazy-object-proxy	1.3.1	pcre	8.39	sphinxcontrib	1.0
click	6.7	libedit	3.1	pexpect	4.3.1	sphinxcontrib-websupport	1.0.1
cloudpickle	0.5.2	libffi	3.2.1	pickleshare	0.7.4	spyder	3.2.6
configurable-http-proxy	3.1.0	libgcc	7.2.0	pip	9.0.1	sqlalchemy	1.2.1
cryptography	2.1.4	libgcc-ng	7.2.0	prompt$_{t}$*oolkit*	1.0.15	sqlite	3.20.1
cycler	0.10.0	libgfortran	3.0.0	psutil	5.4.0	statsmodels	0.8.0
dask	0.16.1	libgfortran-ng	7.2.0	ptyprocess	0.5.2	tblib	1.3.2
dask-core	0.16.1	libiconv	1.15	pycodestyle	2.3.1	testpath	0.3.1
dbus	1.10.22	libpng	1.6.34	pycparser	2.18	tk	8.6.7
decorator	4.2.1	libsodium	1.0.15	pyflakes	1.6.0	toolz	0.8.2
decorator	4.2.1	libstdcxx-ng	7.2.0	pygments	2.2.0	tornado	4.5.3
distributed	1.20.2	libxcb	1.12	pylint	1.8.1	traitlets	4.3.2
docutils	0.14	libxml2	2.9.7	pyopenssl	17.5.0	typing	3.6.2
entrypoints	0.2.3	locket	0.2.0	pyparsing	2.2.0	urllib3	1.22
expat	2.2.5	lzo	2.1	pyqt	5.6.0	wcwidth	0.1.7
fasteners	0.14.1	mako	1.0.7	pysocks	1.6.7	webencodings	0.5.1
fontconfig	2.12.6	markupsafe	1.0	pytables	3.4.2	wheel	0.30.0
freetype	2.8.1	matplotlib	2.1.2	python	3.6.4	wrapt	1.10.11
gettext	0.19.7	matplotlib-venn	0.11.5	python-dateutil	2.6.1	xorg-libxau	1.0.8
glib	2.55.0	mccabe	0.6.1	python-editor	1.0.3	xorg-libxdmcp	1.1.2
gmp	6.1.2	mistune	0.8.3	python-oauth2	1.0.1	xz	5.2.3
gst-plugins-base	1.8.0	mkl	2018.0.1	pytz	2017.3	yaml	0.1.6
gstreamer	1.8.0	monotonic	1.3	pyyaml	3.12	zarr	2.1.4
h5py	2.7.1	msgpack-python	0.4.8	pyzmq	16.0.3	zeromq	4.2.2
hdf5	1.10.1	nbconvert	5.3.1	qt	5.6.2	zict	0.1.3
heapdict	1.0.0	nbformat	4.4.0	qtawesome	0.4.4	zlib	1.2.11

Table 4.3 **Programming libraries used in Python.** This table contains a list of all programming libraries and their version numbers used in my Python implementation.

4.2.3.1 Choice of the ExtraTreesClassifier

The proposed algorithm is a classification algorithm. Given that the training dataset consists of both binary (BTLs) and continuous features of different magnitudes, I chose a classifier called ExtraTreesClassifier, supplied by scikit-learn, that is known to perform well for this type of classification task.

The ExtraTrees method is a variation of the random forest algorithm, which is based on decision trees. In this case, the goal of the random forest algorithm is to classify different data points into categories. The algorithm needs to differentiate disease-causing from non-disease-causing variants. To do that, the random forest algorithm asks a series of questions of the dataset based on the dataset's features to determine how to best classify variants. For example, it could ask if the Exomiser variant score for a specific variant is greater than or equal to a certain cut-off, called a split point. If the answer is yes, there is a high chance the variant is disease-causing. In turn, if the answer was no, there is a high chance the variant is not disease-causing. Merely asking that question gives the model a powerful tool to split the dataset into likely disease-causing and likely non-disease-causing variants.

However, classification purely based on predicted pathogenicity is not sufficiently accurate, since a large number of non-disease-causing variants still receive a high Exomiser variant score, as demonstrated in Chapter 3.

To further distinguish variants from each other, the random forest model asks another question of the already split dataset. For example, the next question could be if the variant has an Exomiser phenotype score greater than or equal to 0.6. If the answer is yes, the likelihood of the variant being disease-causing is slightly higher. A next question could ask if the cross-tissue score in a specific tissue is greater than or equal to a certain value, and so on.

The random forest algorithm first creates a question with a split point defined based on the distribution of data points for each feature. The split point is set as the value with the highest accuracy for distinguishing disease-causing from non-disease causing variants. To add another split point, the algorithm asks another question of the split subsets of data based on another feature. Each question thus becomes a yes/no node in a decision tree.

Each decision tree uses a defined number of features and only gets access to a subset of the available data to build the tree. The features and subset of the data selected for each tree are chosen at random, hence the name 'random forest'.

The final classification prediction of the random forest for a new datapoint is the average of the predictions of all individual trees.

The ExtraTrees classifier is based on the random forest algorithm but reduces the variance of the model. The variance is the average of the squared differences from the mean of a distribution. As such, it represents how far a set of data points is spread out from their mean. In machine learning, a high variance means an algorithm is modelling the random noise in a training dataset, rather than the intended result. This phenomenon is called overfitting. Overfitting is to be avoided so that a machine learning algorithm works well on all datasets it has to process, not just the one it was trained on.

The ExtraTrees classifier reduces variance through two steps. First, the algorithm samples data points from the training dataset without replacement. Sampling without replacement means that a datapoint used in one random subset of the data used to train one tree cannot also be used in a second subset of the training data for a second tree. Second, splits for nodes are chosen at random among a random subset of features, and not as the most optimal split point for each node. The ExtraTrees classifier was chosen to reduce overfitting.

4.2.3.2 Training of the algorithm and tuning of hyperparameters to prevent overfitting

I used the scikitlearn 'auto' setting for the ExtraTreesClassifier and only varied the number of trees used in the model for testing [222]. For the auto setting, the ExtraTreesClassifier uses the entire available dataset to train each tree, and not a random

subset. Furthermore, the number of randomly sampled features used per tree is determined as the square root of the number of available features.

The model is trained on a dataset of disease-causing and non-disease-causing variants in order to be able to distinguish the variants based on a number of different characteristics. In this case, those characteristics are the features described in Section 4.2.2.1. Once the algorithm is trained, given a variant it previously has not seen, it is supposed to be able to classify that new variant as either disease-causing or non-disease-causing based on the features of that new variant. In order to ensure the algorithm is trained properly, several steps are necessary.

First, features are scaled. While Exomiser's output scores range from 0 to 1, the cross-tissue and cross-gene scores display a wide distribution of values. To train machine learning models, it is important that all features in a training dataset display values in the same order of magnitude. Therefore, the training dataset's features were scaled using the scitkit-learn preprocessing imputer.

Thereafter, so-called 'overfitting' has to be prevented. Overfitting means that an algorithm works well on the dataset it is trained on, but will perform poorly on any other dataset. A good algorithm should work on any given dataset. In other words, the algorithm is not only supposed to classify variants from the SWISSProt database well, but has to be able to classify any variant.

Overfitting is controlled for through a mechanism called stratified n-fold cross validation [223]. The idea behind cross validation is to split the dataset into n equally sized portions and then train the algorithm on a subset of those portions, while testing its performance on the remaining portions that the algorithm was not trained on. In this case, I split the dataset into ten separate portions. The algorithm is trained on the combination of nine random portions and tested on the remaining tenth portion. That process is repeated n times on continuously randomly differing splits of the dataset. Once the process is completed, the average outcome from the n iterations is calculated

to determine how well the model performs for the classification task. The key measure to determine a model's performance for a classification task is the so-called area under the curve (AUC) in a receiver-operator characteristic (ROC) plot. In an ROC diagram, the sensitivity of a classification algorithm is plotted as a function of the algorithm's specificity. The AUC is a measure for the algorithm's accuracy. The higher the AUC, the more accurate the algorithm. Thus, the measure for how well a specific algorithm works for the classification task is the average AUC of the n folds.

One variable to control the performance of a classification algorithm are its hyperparameters. In the case of the ExtraTreesClassifier, the hyperparameter to control is the number of decision trees allowed in the model. To create the best possible algorithm, I tested the algorithm for 10, 50, 100 and 500 trees.

To control for overfitting, the model trained and tested as described above is run on yet another portion of the training dataset that the model previously has not seen. The mean model created via the n-fold process should perform the same way on the previously unseen dataset as it did on data it was trained for (see Section 4.3.2 for results).

4.2.4 Modeling of different scenarios

To test if tissue-specific expression features improve variant classification performance, I benchmarked GPET against Exomiser. To compare the algorithms under fair conditions, Exomiser was retrained on my training dataset. Thereafter, the retrained version of Exomiser was compared to GPET under two different scenarios.

In the first scenario, the training and testing data are perfectly annotated with HPO terms to simulate an ideal situation for variant classification. In the second scenario, to demonstrate the usefulness of tissue-specific expression features, the percentage of variants in my dataset for which Exomiser's phenotype score performs well is varied

to illustrate how the algorithms perform on real-life patient cases, where the diagnostic yield of WGS studies still suggests that no diagnostic answer is found for ≈60% of cases.

4.2.4.1 Retraining of Exomiser and benchmarking on a perfectly annotated dataset

To fairly compare GPET with Exomiser, Exomiser was retrained on my dataset using the same hyperparameters as GPET. Retraining is done to avoid unfair biases in algorithm comparisons. Exomiser's hiPHIVE algorithm was originally trained on a different dataset and is therefore optimised for a different classification task. To retrain Exomiser, I trained and tested the same ExtraTreesClassifier as described in Section 4.2.3.1 on the training dataset described in Section 4.2.2, but only using the Exomiser variant score and phenotype score as features.

Separately, I also trained a third model on the dataset, which uses Exomiser's scores, as well as the cross-gene and cross-tissue scores, but omitting the BTLs. That way, the performance of the retrained Exomiser hiPHIVE algorithm could be compared with the performance of GPET using just Exomiser's features with cross-gene and cross-tissue scores, as well as GPET using all features (i.e. Exomiser's scores, the cross-gene and cross-tissue scores, as well as the BTLs). The results were plotted in a ROC diagramme.

4.2.4.2 Benchmarking on an imperfectly annotated dataset

In the previous comparison, the three models (Exomiser_retrained, Exomiser+Expression Scores and Exomiser+Expression Scores+BTLs) were trained and tested on a dataset with a perfectly performing phenotype score. The results in Section 3.3.2.2 however underscored that a significant portion of disease genes are yet to be discovered

and therefore are not annotated with HPO terms. The motivation behind using tissue-specific expression data and BTLs as an additional data source was to supply additional evidence that can be used for variant prioritisation and is independent of genotype and phenotype data. Therefore, I tested how the three models perform when the phenotype score does not perform perfectly, thus simulating scenarios in which a portion of the variants to be classified are not perfectly annotated with HPO terms. To do this, I created three different scenarios.

In the first scenario, the Exomiser phenotype scores for all variants in the training dataset were not changed, but the Exomiser phenotype score of an increasing percentage of the variants in the testing dataset was set to zero. I chose to set the phenotype score to zero to simulate scenarios in which all three stages of the hiPHIVE algorithm (human phenotype, mouse phenotype and PPI) do not contain annotations for the candidate gene and thus produce an output phenotype score of zero.

In the second scenario, the Exomiser phenotype score is set to zero for the same percentage of variants in both the training and the testing dataset.

The third scenario reflects the status quo of the HPO's annotations most closely. The Exomiser phenotype score for 50% of the variants in the training dataset was set to zero and the model was tested on a dataset with an increasing percentage of the Exomiser phenotype scores of variants set to zero.

For all three tests, the resulting AUC of the ROC was plotted as a function of the percentage of variants in the training and/or test dataset for which the Exomiser phenotype score was set to zero. For the model trained and tested on datasets with 50% of the variants' Exomiser phenotype scores set to 0, an ROC was plotted.

4.3 Results

In this section, I present results from the validation of GPET, including a training dataset characterisation (see Section 4.3.1), data from the hyperparameter tuning of the machine learning model (see Section 4.3.2), and a benchmark study comparing GPET with Exomiser for various scenarios (see Section 4.3.3).

4.3.1 Characterisation of the training dataset

All variants in the dataset were mapped to tissues they are suspected to affect based on their HPO annotations as described in Section 4.2.2.3. Figure 4.5 shows the number of variants mapped to each tissue category. Variants were mapped to more than one tissue if their HPO terms matched more than one tissue category. If no HPO terms annotated to a gene in question could be mapped to one of the 25 tissues, the variants were not annotated with a tissue.

The largest number of variants is mapped to brain tissue (≈11,000 variants), followed by muscle, skin, heart, blood and liver. Approximately 3,800 variants could not be mapped to a tissue.

Furthermore, I assessed how well each feature of the dataset helps distinguish non-disease-causing from disease-causing variants. If a feature is powerful for a binary classification task, for example distinguishing disease-causing from non-disease-causing variants, a bimodal result distribution is expected. In this case, the majority of disease-causing variants would be expected to cluster at one end of the distribution for the respective feature, while the vast majority of non-disease-causing variants should cluster at the other end of the distribution. Figure 4.6 shows box plots for the following metrics: Exomiser's variant, phenotype, and combined scores, as well as the mean cross-gene and cross-tissue scores (for both Wilcoxonp and limma) across all tissues for each variant.

All six plots show that each respective feature is sufficiently powerful to distinguish disease-causing from non-disease-causing variants. Exomiser's variant score shows a clear separation between the two categories, with minimal overlap between the two categories. Exomiser's phenotype score achieves a less pronounced, but still visible separation of the values. These results are in line with the observations from Chapter 3. The histogram in Chapter 3 plotting all variants scored by Exomiser's variant and phenotype scores for the HICF2 patients showed that the two scores produce a largely binary distribution (see Figure 3.9).

The cross-gene and two cross-tissue scores distinguish between non-disease-causing and disease-causing variants from the dataset. This is a unique finding that, to my knowledge, has not been shown before. The values plotted for each variant are the mean of the variant's scores for all tissues. Thus, for all three scores plotted, disease-causing variants are generally more highly expressed than non-disease-causing variants. Importantly, that effect is true for both the cross-gene comparison, as well as the cross-tissue analyses. These results suggest that tissue-specific expression holds predictive power for distinguishing disease-causing from non-disease-causing variants.

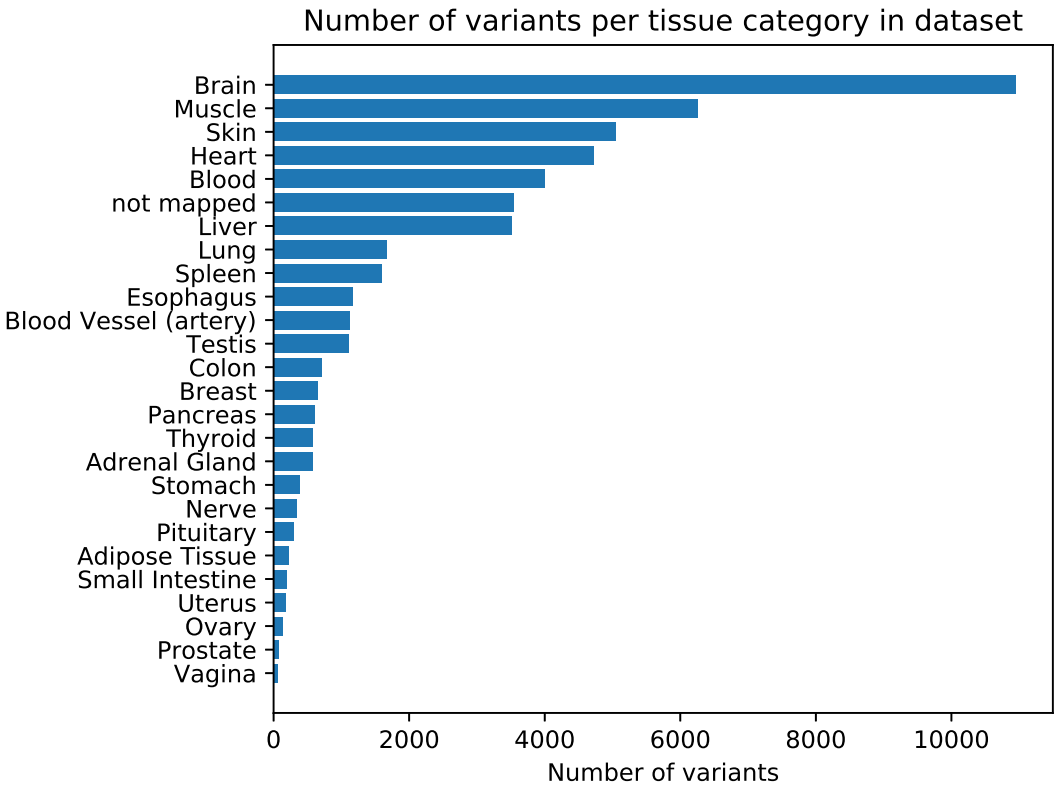


Fig. 4.5 **Overview of tissue distribution of variants in the GPET dataset.** A total of 50,514 variants were mapped to 25 broad tissue categories. This plot shows the number of variants mapped to each tissue category. The largest number of variants, roughly 11,000, were mapped to 'Brain' tissue. One variant can be mapped to multiple tissue categories depending on the term-to-tissue mapping of the associated HPO terms.

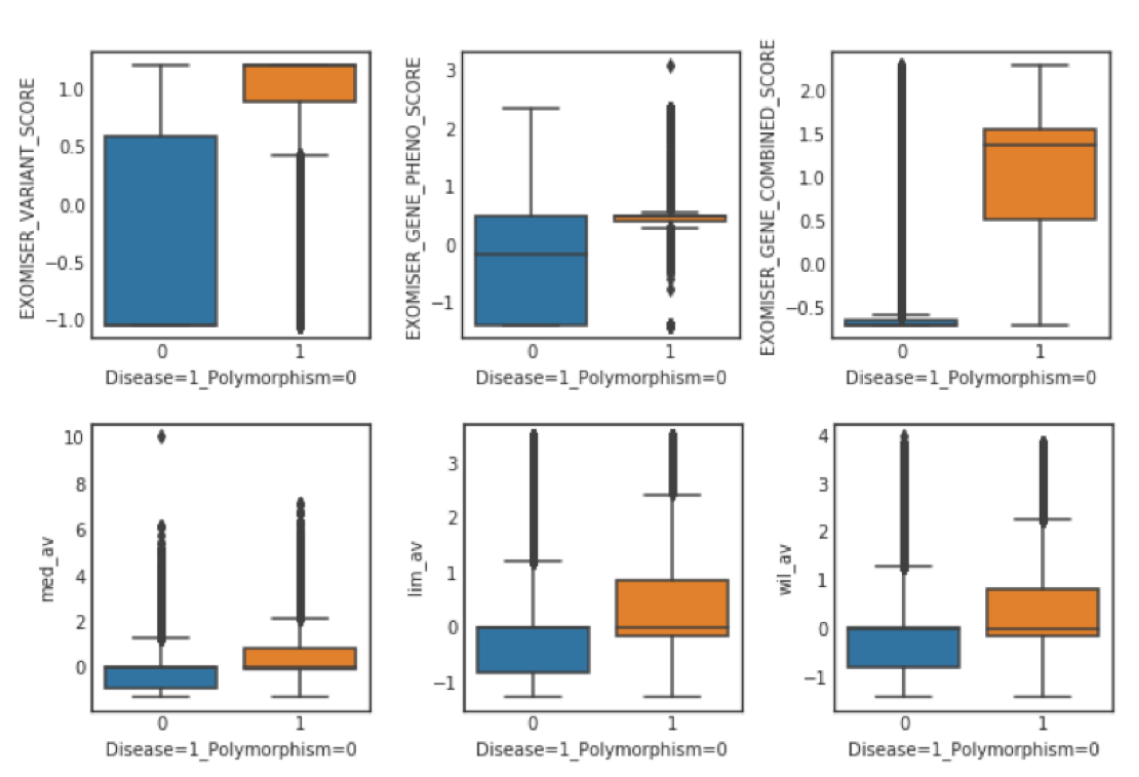


Fig. 4.6 **Distribution of variants for different GPET features.** Distribution of disease-causing and non-disease-causing variants for Exomiser's variant score, Exomiser's phenotype score, Exomiser's combined score, the cross-gene score ('med_av') and the cross-tissue scores ('lim_av' for the limma and 'wil_av' for the Wilcoxonp method). The rectangle in each box plot captures the second and third quartile, with the horizontal line in the rectangle indicating the median. The upper and lower quartiles are shown as horizontal lines above and below the rectangle. The data for disease-causing variants is shown in orange and the data for non-disease-causing variants in blue. Each score separates disease-causing from non-disease-causing variants, with the Exomiser combined score showing the best performance

4.3.2 Optimisation of the machine learning model

To find the best-suited settings for the machine learning algorithm, the number of trees used for the chosen ExtraTreesClassifier had to be optimised. The 10-fold-cross-validation was run for 10, 50, 100, 500, and 1000 trees. The AUC of the mean model for 500 trees reached its maximum at 0.98 +/- 0.01. Using 1000 trees did not improve the performance further. Thus, 500 trees were chosen as the standard setting for all subsequent analyses. Figure 4.7 shows data of the 10-fold cross-validation for 10, 50, 100, and 500 trees.

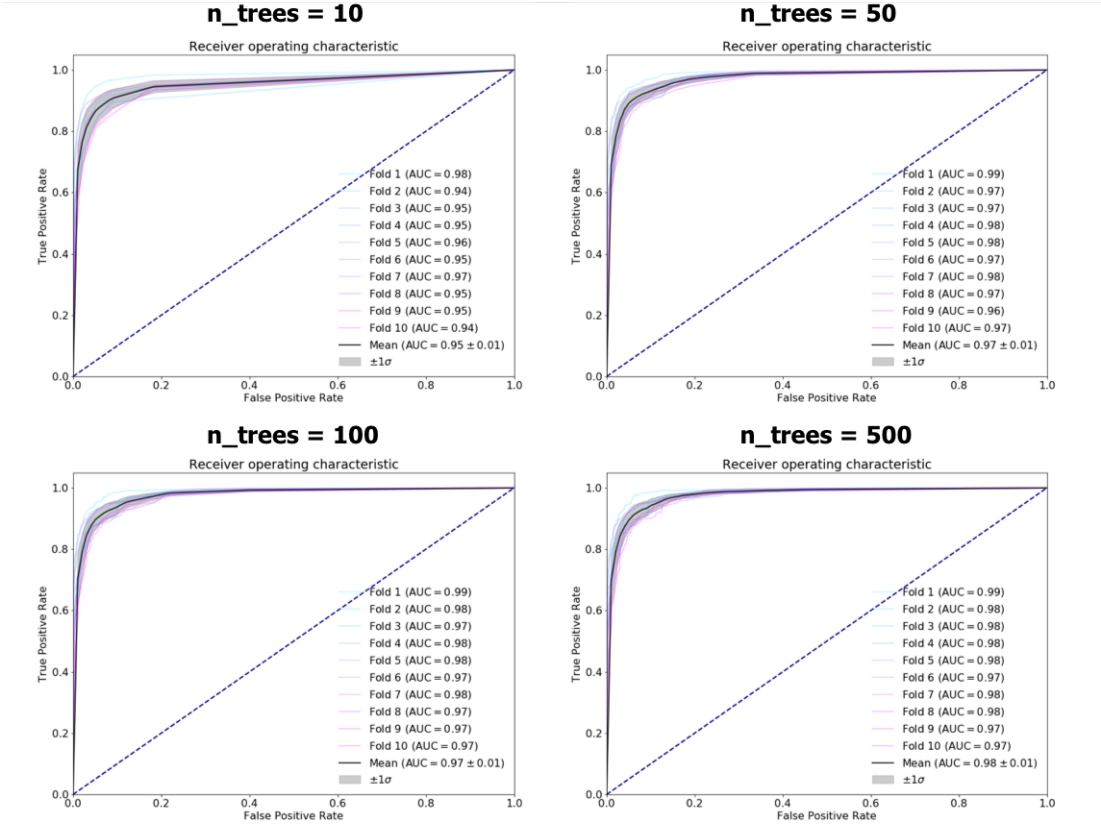


Fig. 4.7 **Hyperparameter tuning for GPET via 10-fold cross-validation.** To improve the accuracy of GPET and achieve the highest AUC, the machine learning model had to be optimised. For that purpose, the number of trees, a specific hyperparameter of random forest models, was varied (10, 50, 100, and 500) in a process called 10-fold cross-validation (see Section 4.2.3.2 for details) and the number of trees for which the highest AUC is achieved was determined. n_trees = 500 was found to have the best performance at AUC = 0.98 +/- 0.01

4.3.3 Benchmarking of the novel algorithm against Exomiser

To assess GPET's classification performance, I benchmarked the algorithm against a retrained version of Exomiser (see Section 4.2.4) under two different scenarios: perfect phenotypic annotations (see Section 4.3.3.1) and real-world phenotypic annotations (see Section 4.3.3.2).

4.3.3.1 Scenario 1: perfect annotations

First, the algorithms were trained and tested on a perfectly annotated dataset as described in Section 4.2.4. As shown in the ROC curve in Figure 4.8, GPET using all available features achieves an AUC of 0.99, only marginally outperforming the retrained version of Exomiser with an AUC of 0.98.

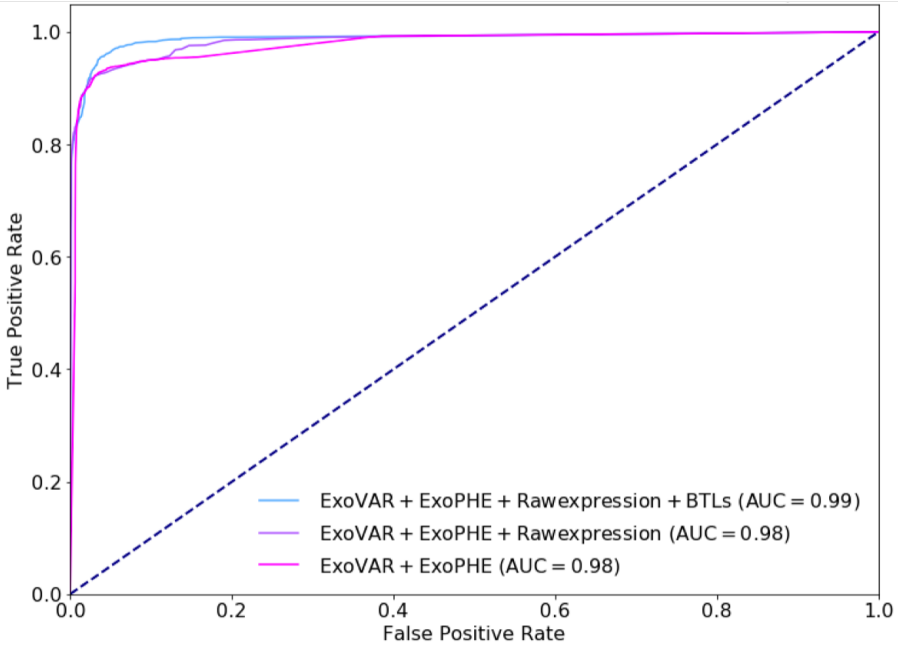


Fig. 4.8 **Benchmarking of GPET against Exomiser with perfect annotations.** GPET was benchmarked against Exomiser's performance on a perfectly annotated dataset. In the figure, the different algorithms are described by the combination of their component features. Exomiser (represented as ExoVAR + ExoPHE) achieves an AUC of 0.98. The additional use of gene expression features (referred to as 'Rawexpression') and BTLs only shows marginal improvements over Exomiser's variant and phenotype scores, with AUCs of 0.99 and 0.98 respectively.

Table 4.4 shows the performance of additional combinations of the individual features used to train GPET on the perfectly annotated dataset. GPET achieves the best performance with an AUC of 0.99, while Exomiser's phenotype score on its own produces the lowest AUC at 0.68.

Features	AUC
ExoVAR + ExoPHE + Rawexpression + BTLs	0.99
ExoVAR + ExoPHE + Rawexpression	0.98
ExoVAR + ExoPHE	0.98
ExoCombined	0.91
ExoVAR + Rawexpression	0.91
Rawexpression + BTLs	0.90
ExoVAR + BTLs	0.90
ExoVAR	0.89
Rawexpression	0.89
BTLs	0.84
ExoPHE	0.68

Table 4.4 **Performance comparison of different feature combinations used for GPET with perfect annotations.** GPET was trained using Exomiser's variant and phenotype scores (ExoVAR and ExoPHE), as well as gene expression features ('Rawexpression') and BTLs. In this table, the AUC for different combinations of those features for the classification task on a perfectly annotated dataset is shown, with GPET achieving the highest AUC (0.99) and ExoPHE achieving the lowest AUC (0.68). For reference, the performance of Exomiser's combined score (ExoCombined) is also shown (AUC=0.91). For the benchmarking of GPET against Exomiser, a retrained version of Exomiser's combined score (ExoVAR+ExoPHE) was used to avoid unfair training biases (see Section 4.2.4.1 for details).

4.3.3.2 Scenario 2: real-world annotations

The second benchmarking test represents a scenario closer to the practice of clinical genetics. Large WGS and WES cohort studies regularly achieve maximum diagnostic yields of 35-40%, suggesting that a significant share of cases likely harbour disease-causing variants in genes for which the link to a disease phenotype yet has to be established. Simultaneously, only 3,526 genes are annotated in the version of the HPO used for Exomiser v7.2.1 (HPO version 20160125), while another ≈21,500 genes remain unannotated. Thus, it is likely that a significant portion of disease genes have not been annotated with HPO terms. To simulate that effect, I set the Exomiser phenotype score of a varying portion of variants to zero and reran the ROC analysis to determine how the algorithms' performance would change as a function of decreasing phenotype scoring performance.

First, the three algorithms (Exomiser, Exomiser with tissue-specific expression features, and GPET) were trained on perfectly annotated data and tested on a dataset for which an increasing percentage of variants had their Exomiser phenotype scores set to 0. The result is shown in Figure 4.9, plot 1. The more variants lose their Exomiser phenotype score, the more the AUC of each algorithm decreases. The curves start with the AUCs shown in Figure 4.8 at ExoPHE%=0, with Exomiser's AUC at 0.98, Exomiser with tissue-specific expression features' AUC at 0.98, and GPET's AUC at 0.99. The AUC of GPET is consistently the highest, including the final step (ExoPHE% = 100%), at which point GPET achieves an AUC of 0.86, compared to Exomiser's 0.8. In the absence of reliable phenotypic annotations, tissue-specific expression and BTLs contribute significantly to the AUC.

As a control, I plotted the feature importance of the Exomiser phenotype score for all three algorithms as a function of ExoPHE% (see Figure 4.10, plot 1). The

phenotype score feature importance stays constant for all tested ExoPHE% since the dataset the model is trained on is not changed.

Thereafter, the three algorithms were trained and tested on datasets with a varying ExoPHE%. The results are shown in Figure 4.9, plot 2. The AUC of all three algorithms drops as the phenotype score becomes less reliable. Since the algorithms are always trained under the same conditions as they are tested (i.e. the same ExoPHE%), the AUC does not drop as significantly as for the previous analysis (0.84, 0.91 and 0.92 respectively for Exomiser, Exomiser with tissue-specific expression features, GPET). The tissue-specific expression features and BTLs enable GPET to rank the variants in the absence of a reliable phenotype score. The feature importance of Exomiser's phenotype score drops with every increase in ExoPHE%, as shown in Figure 4.10, plot 2.

Finally, to create a scenario that is closest to the real-life clinical genetics setting, I trained the algorithms on a dataset with ExoPHE%=0.5 and again tested the algorithms on a dataset with varying ExoPHE%. Figure 4.9, plot 3 shows the results. All three algorithms achieve an AUC of 0.98 for an ExoPHE% of 0.0, but the progression differs significantly. At an ExoPHE% of 1.00, GPET achieves an AUC of 0.91, compared to Exomiser with 0.83. The feature importance of Exomiser's phenotype score remains constant for all tested iterations (see Figure 4.10, plot 3).

Since the diagnostic yield of large WGS and WES studies is rarely higher than 50%, ExoPHE%=0.5 is chosen as the most realistic scenario. At this cut-off, the AUC of Exomiser is 0.91, while GPET achieves an AUC of 0.95. The corresponding ROC is shown in Figure 4.11. The ROC shows that GPET achieves a low false positive rate at a relatively high true positive rate. At a false positive rate of $\approx$0.05, Exomiser's true positive rate is $\approx$0.5. For the same false positive rate, GPET's true positive rate reaches $\approx$0.8. Hence, expression features can be useful to maximise the number of

true positives and reduce the time analysts have to spend on excluding false positives.

For all three testing scenarios, GPET achieves the highest overall AUC.

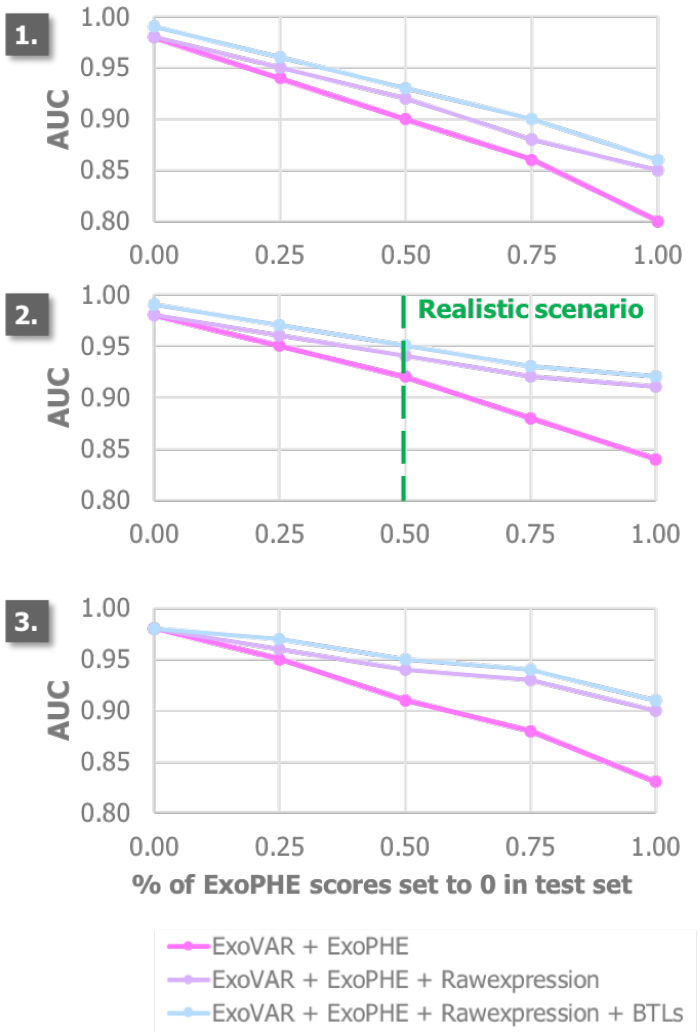


Fig. 4.9 **AUC as a function of HPO annotation disturbance.** The three algorithms characterised in the benchmark analysis (Exomiser = ExoVAR + ExoPHE (pink), Exomiser with tissue-specific expression features = ExoVAR + ExoPHE + Rawexpression (purple), and GPET = ExoVAR + ExoPHE + Rawexpression + BTLs (blue)) were trained and tested on a dataset in which the percentage of variants for which the Exomiser phenotype score (ExoPHE) was set to 0 was varied to simulate missing HPO annotations. Plot 1 shows the AUC for algorithms trained on a perfectly annotated dataset and tested on a dataset with varying ExoPHE=0 percentages. Plot 2 shows the AUC for algorithms trained and tested on a dataset with varying ExoPHE=0 percentages. Plot 3 shows the AUC for algorithms trained on a 50% annotated dataset and tested on a dataset with varying ExoPHE=0 percentages. Increasing ExoPHE lowers the AUC for all three algorithms. Using tissue-specific expression features and BTLs improves the AUC with an increasing ExoPHE=0 percentage.

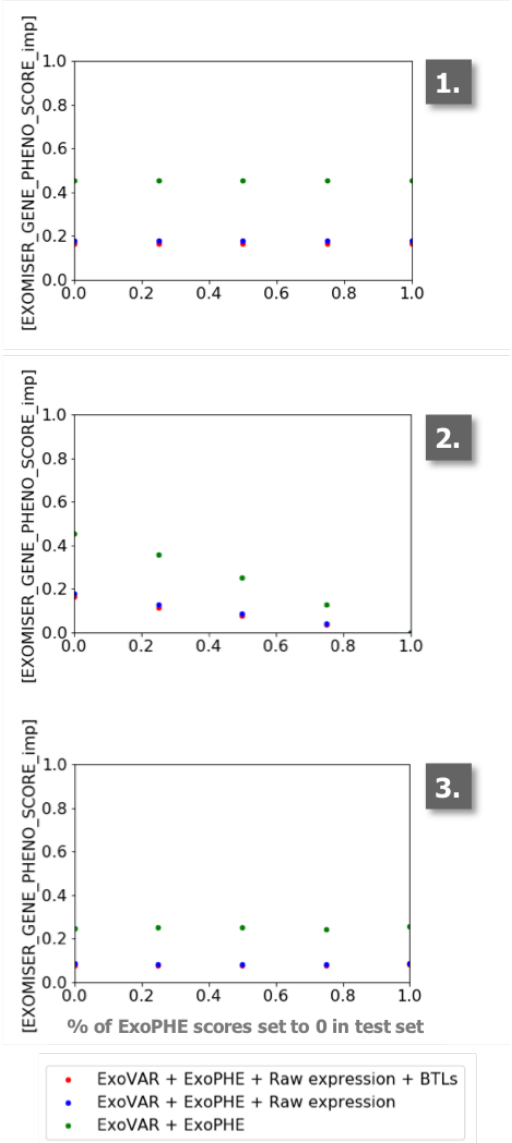


Fig. 4.10 **Feature importance as a function of HPO annotation disturbance.** Corresponding to Figure 4.9, this figure shows the feature importance of the Exomiser phenotype score (EXOMISER_GENE_PHENO_SCORE_imp) for all three algorithms as a function of the percentage of variants for which the Exomiser phenotype score (ExoPHE) was set to 0 for the training and test dataset. Plot 1 shows EXOMISER_GENE_PHENO_SCORE_imp for algorithms trained on a perfectly annotated dataset and tested on a dataset with varying ExoPHE=0 percentages. Plot 2 shows EXOMISER_GENE_PHENO_SCORE_imp for algorithms trained and tested on a dataset with varying ExoPHE=0 percentages. Plot 3 shows EXOMISER_GENE_PHENO_SCORE_imp for algorithms trained on a 50% annotated dataset and tested on a dataset with varying ExoPHE=0 percentages. When the ExoPHE=0 percentage is varied for both the training and test dataset (see Plot 2), EXOMISER_GENE_PHENO_SCORE_imp decreases as a function of an increasing ExoPHE=0 percentage. When the ExoPHE=0 percentage is not varied for the training dataset, EXOMISER_GENE_PHENO_SCORE_imp stays constant (plot 1 and plot 3). EXOMISER_GENE_PHENO_SCORE_imp for plot 3, where the ExoPHE=0 percentage of the training dataset is higher than for plot 1, is lower than EXOMISER_GENE_PHENO_SCORE_imp for plot 1.

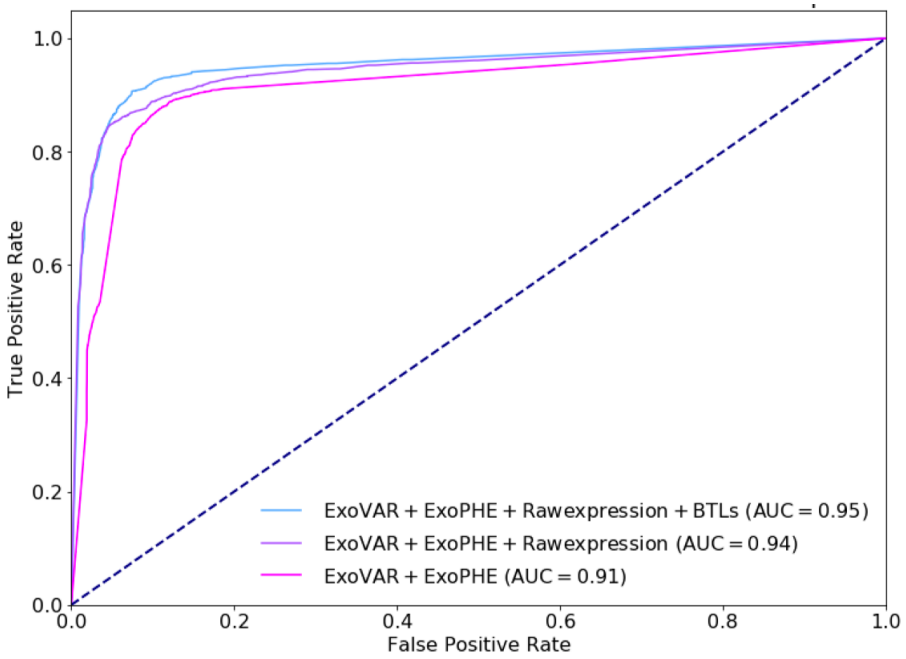


Fig. 4.11 **Benchmarking of GPET against Exomiser with realistic annotations.** GPET was benchmarked against Exomiser using a dataset with annotations that are deemed to represent a realistic scenario, where 50% of the training and test datasets are annotated with the correct phenotype score, corresponding to Figure 4.9, plot 3. GPET (shown here as ExoVAR + ExoPHE + Rawexpression + BTLs) achieves an AUC of 0.95, as compared to Exomiser (shown here as ExoVAR + ExoPHE) with an AUC of 0.91.

4.4 Discussion

The proof of concept described in this chapter illustrates two points: first, tissue-specific gene expression data contains valuable information for rare disease variant prioritisation. Second, the HPO carries implicit information indicating which tissues are likely affected by a disease phenotype. In turn, that information can be used in conjunction with tissue-specific expression data to improve rare genetic disease variant prioritisation. To my knowledge, both of these findings are novel and have not been previously reported (see Section 4.3.1).

Based on these findings, I developed GPET, a machine learning algorithm that distinguishes disease-causing from non-disease-causing variants using Exomiser's variant and phenotype scores, as well as tissue-specific gene expression data and BTLs. When all variants in the training and test dataset are perfectly annotated with phenotypic information, GPET achieves a marginally higher AUC than a retrained version of Exomiser's hiPHIVE algorithm (0.99 and 0.98, see Section 4.3.3.1), the current state of the art used by many geneticists, including the 100,000 Genomes Project. If, however, a scenario closer to the current status quo of databases is modelled and 50% of the variants in the dataset have no phenotypic annotations, GPET shows improved performance compared to Exomiser's hiPHIVE (AUC of 0.95 and 0.91, see Section 4.3.3.2).

The fact that elevated gene expression can be leveraged for the identification of disease-causing genes is in line with work done by Feiglin *et al.* [132].

However, several factors limit GPET's performance. Feiglin *et al.* [132] demonstrated that the link between elevated gene expression in a specific tissue and the disease phenotype varies in strength depending on the tissue, as represented by a log odds ratio calculated by the authors. A log odds ratio greater than 0.5 indicates a strong correlation of gene expression levels and the disease phenotype, for example for 'Brain'

tissue. For vaginal tissue however, for example, the log odds ratio for the cross-gene score is less than 0.5, indicating that no correlation can be established between gene expression and the relevance of the gene for the disease phenotype. Importantly, log odds ratios are stronger for tissues where GTEx includes more samples. Low log odds ratios are observed for tissues with small sample numbers. This effect might thus be mitigated as the GTEx database grows.

Furthermore, the age of onset of the disease in question is important. GTEx samples are collected from adults between 20 and 70 years of age and are hence not representative of paediatric conditions. Additionally, gene expression profiles of developing organs and tissues are not included in the database. This limitation, too, will be mitigated over time as gene expression databases grow.

In addition to sampling issues, the mapping of HPO terms to GTEx tissues created by Feiglin *et al.* [132] and used for GPET is incomplete and should be expanded. In this analysis, HPO terms only map to 25 high-level GTEx tissue categories (see Table 4.1). In GTEx, however, more than 70 specific tissue categories are available, almost all of which can be accurately mapped to HPO terms. A more granular HPO-to-tissue mapping would likely improve algorithm performance.

Additional advancements of variant classification algorithms can be achieved with larger databases and an improved generation of training dataset features.

The performance of VPA is further limited by the size of available real-world training datasets. The publication of studies such as the 100,000 Genomes Project will make a training dataset of significant size available to scientists to train algorithms on real-world data produced by one consistent pipeline. This represents an improvement over algorithms like GPET or Exomiser, which were trained on artificially assembled datasets, often created by seeding genomes from the 1,000 Genomes Project with variants that have previously been confirmed to be disease-causing. Problematically, several studies have demonstrated that not all variants that are labeled to be disease-

causing in databases such as ClinVar or HGMD could reproducibly be proven to actually cause disease in humans. While some variants labeled as disease-causing in these reference databases were identified in real human genomes of patients with the correct associated phenotypes, other variants were merely predicted to be pathogenic by algorithms such as CADD, SIFT, or Polyphen, and subsequently submitted to reference databases. If new algorithms are trained on variants that are labeled as disease-causing based on the prediction of another *in silico* algorithm, those new algorithms simply learn to mimic the behaviour of the original classification algorithm, instead of learning to identify disease-causing variants. By combining new real-world datasets with advanced machine learning models, rare genetic disease diagnostics will be significantly accelerated.

Beyond the approach demonstrated here, more opportunities exist to leverage genomic, phenotypic, and other types of data to identify disease-causing variants. Existing studies suggest that disease-causing genes have elevated transcript levels and an increased number of tissue-specific protein-protein interactions in relevant disease tissues compared to unaffected tissues [130]. To improve the algorithm, features capturing tissue-specific protein-protein interaction (PPI) could be included. Similarly, data on gene co-expression can be leveraged to establish further links between gene expression across different tissues. Deelen *et al.* [142] published GADO, an algorithm that uses co-regulation patterns of genes based on expression data and HPO terms to rank candidate genes. GADO was published after the work presented in this chapter was concluded and should be included in future benchmark studies.

Another alternative was presented by Mosley *et al.* [134]. Instead of identifying tissues in which gene expression levels are likely elevated based on the phenotype, Mosley *et al.* [134] inferred likely elevated protein levels based on the genotype, thus creating a 'virtual proteome'. The authors subsequently tested correlations between the virtually inferred proteomic markers and the individuals' phenotypes. Using that

methodology on over 40,000 genotyped individuals, they were able to demonstrate that 55 proteins were associated with 89 distinct diagnoses. In a future study, Mosley *et al.*'s [134] approach could be combined with GPET to test if correlations exist between the genotype-inferred virtual proteome and phenotype-inferred tissue-specific gene expression levels. If successful, existing genotype and phenotype data could be used to study the proteome of participants in studies where proteomic data is not available.

It is possible that tissue-level expression data will eventually be replaced by single cell sequencing data, mapping phenotypes to cell-level rather than tissue-level gene expression.

In addition to expression-based features, new conservation-based features are being introduced. Algorithms such as SIFT and Polyphen are based cross-species conservation. With growing databases on humans, intra-species conservation can be measured in addition to cross-species conservation.

Moreover, large datasets like the 100,000 Genomes Project and the UK BioBank open up new avenues for combining rare disease genetics with common disease genetics. Instead of relying on gene-level phenotypic annotations, the UK BioBank could be used to recall Mendelian phenotypes from GWAS hits produced by studies of common diseases. For example, microcephaly is a rare disease phenotype that exists as an extreme end of the spectrum of head circumference. GWAS results linked to Mendelian phenotypes could be used to annotate loci with phenotypic terms, creating phenotypic annotations for non-coding regions of the genome, which can be used for variant prioritisation.

Data presented in this chapter was limited to demonstrating that tissue-specific gene expression data and BTLs are useful for creating a prioritisation algorithm that performs well on *in silico*. In the next chapter, GPET is applied to the real patient cases introduced in Chapter 3 to test if this approach is useful for novel gene discovery.

Chapter 5

Evaluation of a new genotype-, phenotype-, and tissue-specific expression-based variant prioritisation algorithm on rare disease patient cases

5.1 Introduction

There are approximately 7,000 rare genetic diseases and molecular causes have been identified for only around 4,000, or ≈60% [5]. Resources like the HPO annotate approximately 4,300 genes with disease phenotypes [179]. The diagnostic yield of rare genetic disease population cohorts rarely exceeds 50% [24, 25]. At the same time, ≈100 new disease-causing genes are discovered every year [172].

To solve more rare genetic disease cases faster, improved methods are required to aid with the discovery of novel disease gene candidates. Various approaches exist to facilitate novel disease gene discovery, including ranking algorithms based on genotypic and phenotypic data such as Exomiser's hiPHIVE algorithm, or VAAST+Phevor [89, 113] (see Chapter 3). In addition to genotypic and phenotypic data, the use of tissue-specific expression data for variant prioritisation has shown promise. In Chapter 4, I introduce a novel analysis framework called GPET, named for the genotype, phenotype, and tissue-specific expression data it draws on with the aim of improving variant prioritisation for novel disease genes.

GPET showed improved performance compared to Exomiser's hiPHIVE algorithm for variant classification in the absence of perfect phenotypic annotations on a hand-curated *in silico* dataset (AUC of 0.95 compared to 0.91, see Section 4.3.3.2). The dataset used for the benchmarking analysis was based on the SWISSProt database [218, 219], which consists of disease-causing variants from OMIM and variants presumed to be benign from the 1,000 Genomes project. However, simulated datasets commonly used to train VPA suffer from several limitations: labels to distinguish disease-causing from non-disease-causing variants are not always reliable and simulated datasets do not perfectly reflect the challenges of analysing all variants in a real patient's exome or genome. **Therefore, the goal of this chapter is to benchmark GPET against Exomiser's hiPHIVE algorithm on the real patient cases introduced in Chapter 3. Specifically, the performance of GPET on genes that have not previously been linked to a disease phenotype and are thus not annotated in the HPO is assessed.**

5.2 Materials and methods

In this section, I provide details on the patient case data used for analysis (see Section 5.2.1), give an overview of the GPET framework (see Section 5.2.2), describe

a method used to assess if expression scores of candidate genes are higher in HPO-predicted tissues (see Section 5.2.3), and describe how GPET was benchmarked against Exomiser (see Section 5.2.4).

5.2.1 Patient case selection, sequencing, and benchmark variant identification

The analysis conducted in this chapter builds on the analysis of eleven HICF2 cases presented in Chapter 3. Table 5.1 gives details on the HICF2 cases, including which candidate genes were annotated in the December 9th, 2013 version of the HPO used by Exomiser version 7.2.1.

Importantly, a gene being annotated with HPO terms does not necessarily mean that it is a known disease gene for a specific patient's disease. For example, *RBPJ* was annotated in the December 9th, 2013 version of the HPO, but was not published as a disease gene for Atypical Klippel-Trenaunay syndrome, the disease patient nine in Table 5.1 is diagnosed with.

Patient cases consist of a variety of inheritance patterns (autosomal dominant, *de novo*, autosomal recessive and X-linked recessive), pedigree structures (singletons, trios, and cousins), and phenotypes. WGS was conducted, sequencing data was analysed, and benchmark variants for each case were identified by senior post-doctoral geneticists using the HICF2 project's rare disease research bioinformatics pipeline (see Chapter 2 for details).

	Diagnosis	Gene	Variant	# samples	Inheritance	HPO terms	HPO-annotated
1	Distal arthrogryposis	*TNNI2*	c.466C>T [NM_001145829.1], p.Arg156*	Singleton	AD or *de novo*	Micrognathia, downslanted palpebral fissures, high palate, malar flattening, limited shoulder movement, hip dislocation, ulnar deviation of finger, distal arthrogryposis, tapered finger, flexion contracture of finger, abnormality of the hand, talipes	yes
2	Bilateral hippocampal sclerosis	*CACNA1E*	c.5702G>A [NM_001205293.1], p.Arg1901His	Singleton	AD or *de novo*	Seizures, dysgenesis of the hippocampus	no
3	Severe epileptic encephalopathy	*WWOX*	c.705dupG [NM_016373.2], p.His236AlafsTer34	Singleton	AR	Coarse facial features, deep palmar crease, atrial septal defect, hypertonia, profound global developmental delay, epileptic encephalopathy, hypsarrhythmia, infantile spasms, abnormal hand morphology	yes
4	Dilated cardiomyopathy	*ACTC1*	c.664G>A [NM_005159.4], p.Ala222Thr	Trio	*de novo*	Weight for age (decreased body weight (<-2SD)), endocardial fibroelastosis, dilated cardiomyopathy, cardiomegaly, respiratory tract infection	yes
5	Majeed syndrome	*PSTPIP1*	c.748G>A [NM_003978.4], p.Glu250Lys	Trio	*de novo*	Recurrent skin infections, bone marrow hypocellularity, episodic fever, splenomegaly, recurrent infections	yes
6	Undefined immunodysregulatory disorder	*SAMD9L*	c.3353A>G [NM_152703.2], p.Tyr1118Cys	Trio	*de novo*	Nystagmus, psoriasis, respiratory tract infection, arthropathy, colitis, abnormality of the intestine, clumsiness, increased CSF protein, abnormality of the cerebral white matter, gait ataxia, cerebellar atrophy, thrombocytopenia, decreased antibody level in blood	no
7	Fine-Lubinsky syndrome	*POR*	c.1493G>C [NM_000941.2], p.Arg498Pro	Trio	AR	Cleft palate, Narrow mouth, micrognathia, short chin, shallow orbits, agenesis of permanent teeth, plagiocephaly, megalocornea, preauricular skin tag, cupped ear, arthrogryposis multiplex congenita, hypospadias, renal agenesis, moderate global developmental delay, talipes	yes
8	Congenital erythrocytosis	*SLC30A10*	c.823T>A [NM_018713.2], p.Trp275Arg	Trio	AR	Abnormality of the cardiovascular system, cerebral hemorrhage, hypotension, hemangioma, varicose veins, stroke, abnormality of blood and blood-forming tissues, increased hemoglobin, peripheral thrombosis, increased hematocrit, increased red blood cell mass, abnormality of the nervous system, headache, abnormality of the integument, plethora, neoplasm, constitutional symptom	yes
9	Atypical Klippel-Trenaunay syndrome	*RBPJ*	c.535T>G [NM_005349], p.Leu179Val	Trio	*de novo*	Localised skin lesion, hemangioma, juvenile onset, large hemangioendothelioma involving right buttock, right thigh since the age of 6, splenomegaly, abnormal thrombosis	yes
10	Fatal acute encephalitis	*DOCK11*	c.1679C>T [NM_144658.3], p.Ser560Leu	Trio	X	Encephalitis, abnormality of the spleen, abnormality of bone marrow cell morphology, lymphadenopathy	no
11	Fine-Lubinsky syndrome	*HDLBP*	c.1731+1G>A [NM_203346.4], p.Val540_Leu577del	Two cousins	AR	Uplifted earlobe, short nose, short chin, shallow orbits, severe global developmental delay, plagiocephaly, narrow mouth, low-set, posteriorly rotated ears, hypertelorism, contracture of the proximal interphalangeal joint of the 5th finger, brain atrophy, bilateral camptodactyly, abnormal cornea morphology	no

Table 5.1 **Characteristics of the HICF2 patient cases used in this chapter.** This table summarises the diagnoses, candidate genes, specific variants, number of samples used per analysis, and suspected inheritance patterns (AD = Autosomal dominant, AR = Autosomal recessive, X = X-linked recessive) for each HICF2 patient case analysed in this chapter. Furthermore, each patient's phenotype, described in HPO terms, is listed. Finally, it is indicated whether or not a candidate gene was annotated by the HPO at the time of analysis.

5.2.2 GPET analysis framework

Figure 5.1 shows the full GPET analysis pipeline. Non-coding regions were removed from the input VCF using a BED file downloaded from UCSC (see Sectin 3.2.3 for details). Exomiser's hiPHIVE algorithm (version 7.2.1) was run on each VCF to annotate variants via Jannovar, filter out variants with an allele frequency >1%, and assign Exomiser's variant score and phenotype score to each variant. Importantly, this analysis was conducted at a time when the most recent version of the HPO used by Exomiser had been published on December 9th, 2013.

Subsequently, expression-based features are annotated to the dataset as described in Section 4.2.2.3. Each variant in each patient VCF was annotated with tissue-specific expression scores for the gene harbouring the variant. The HPO terms assigned to each patient were used to infer which tissues were likely affected for each patient. Based on that, binary tissue labels (BTLs) were created for each patient. For example, if a patient's phenotypic profile includes the HPO term 'Abnormality of the heart' (HP:0001627), this suggests that the heart tissue is affected by that gene. That HPO term is mapped to the tissue 'Heart' in the HPO-tissue map described in Section 4.2.2.3 and subsequently each variant in the patient's VCF file is annotated with a '1' for the BTL 'Heart'.

Thereafter, the GPET classifier trained and tested in Chapter 4 on a dataset in which 50% of variants were scored with the correct Exomiser phenotype score, while 50% had their phenotype score set to 0, was run on each variant in each patient's VCF to produce the GPET score (see Section 4.2.4.2). Each variant is now annotated with a total of 104 features for genotype, phenotype, tissue-specific expression scores, BTLs, and the combined GPET score.

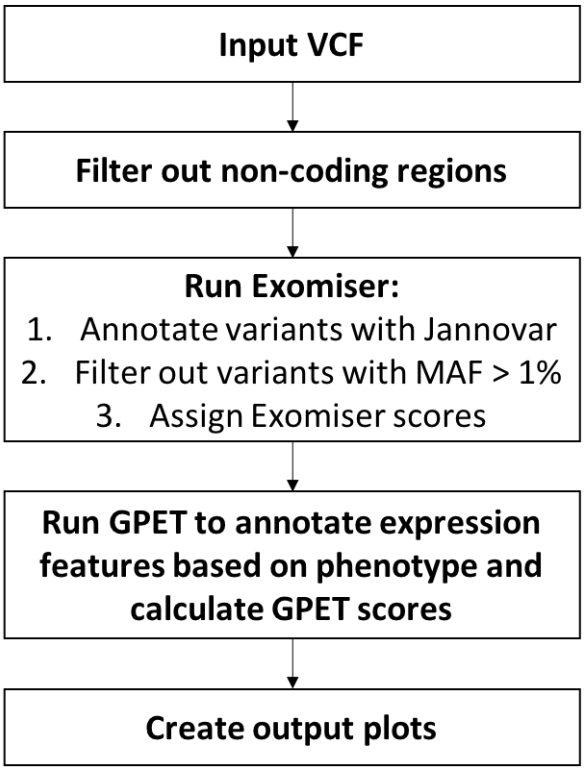


Fig. 5.1 **GPET analysis framework.** To analyse patient WGS data with GPET, variants in non-coding regions are first removed from the input VCF file. GPET is not designed for the ranking of non-coding variants. Subsequently, Exomiser is run on the VCF to annotate variants with Jannovar, remove variants with MAF > 1%, and assign the variant scores and phenotype scores of Exomiser's hiPHIVE algorithm. Thereafter, GPET is run to annotate each variant with four expression-specific features (BTL, cross-gene score, and two cross-tissue scores) for 25 tissues respectively, or a total of 100 features (see Section 4.2.2.3 for details). Finally, the GPET score is calculated for each variant.

5.2.3 Randomisation of HPO terms to assess expression scores per tissue

Feiglin *et al.* [132] and others have shown on *in silico* data that, for many genes, expression is elevated in tissues relevant for the HPO-based phenotype of diseases. To validate if a similar effect can be observed in real patient cases, I compared the expression of disease-causing genes for each patient in the tissues likely affected based on the patient's HPO profile with the candidate gene's expression in all other tissues. For that purpose, I calculated the average cross-gene and cross-tissue scores (as described in Chapter 4) for each candidate gene in the patient's HPO-implicated tissues and in all other tissues. For example, if a patient's HPO terms indicate that 'Blood', 'Lungs', and 'Liver' are affected, the mean of the candidate gene's cross-gene and cross-tissue scores respectively were calculated for those tissues. Similarly, the candidate gene's mean cross-gene and cross-tissue scores were calculated for the remaining non-HPO-indicated tissues, in this case 'Adipose Tissue', 'Adrenal Gland', 'Blood', 'Blood Vessel (artery)', 'Brain', 'Breast', 'Colon', 'Esophagus', 'Heart', 'Liver', 'Lung', 'Nerve', 'Ovary', 'Pancreas', 'Pituitary gland', 'Prostate', 'Muscle', 'Skin', 'Small Intestine', 'Spleen', 'Stomach', 'Testis', 'Thyroid', 'Uterus', and 'Vagina'.

To ensure that any observed effects were not coincidental, I ran the same analysis again, but this time randomly selected which tissues were affected, instead of inferring affected tissues using the patient's HPO terms. For each patient profile, the same number of tissues was used as was indicated by their HPO terms. For example, if the patient's HPO terms indicated that 'Blood', 'Lung', and 'Liver' were affected, the algorithm randomly selects three tissues different from those deemed 'affected' and calculates the mean cross-gene and cross-tissue scores for those pseudo-affected tissues. At the same time, all remaining tissues were used to calculate

the background. This process was repeated 25 times, once for each unique tissue, for each patient, and the mean 'simulated_affected_tissue_expression_score' and 'simulated_unaffected_tissue_expression_score' was plotted for the cross-gene and cross-tissue scores respectively.

5.2.4 Benchmarking of GPET against Exomiser

Subsequently, the performance of the GPET classifier to rank disease-causing variants was compared to the Exomiser hiPHIVE algorithm's variant and phenotype scores, similar to the algorithm comparison presented in Chapter 3. Ranking performance was assessed for all cases collectively, as well as individually for cases where the disease-causing gene was or was not HPO-annotated, respectively.

5.3 Results

In this section, the results of applying GPET to WGS data from real patient cases are presented. I showcase data on the ability of a patient's phenotypic terms to indicate in which tissues a disease-causing gene is likely more highly expressed (see Section 5.3.1). Furthermore, I apply GPET to well-characterised real patient cases to assess the algorithm's ability to rank disease-causing variants for real human genomes in comparison to Exomiser (see Section 5.3.2).

5.3.1 Randomisation of HPO terms to assess expression scores per tissue

Figure 5.2 shows that genes harbouring disease-causing variants for patient cases are more highly expressed in tissues indicated to be affected by the patients' HPO

terms ('HPO-tissue') than in tissues not implicated by the patients' HPO terms ('Background'). This relationship holds true for all the tissue-specific expression scores: cross-gene (median) (Figure 5.2, panel A), cross-tissue (wilcoxp) (Figure 5.2, panel B), and cross-tissue (limma) (Figure 5.2, panel C). The median tissue-specific expression for the HPO-tissue is higher than the background's median by a factor of $\approx$2 for all scoring methods. Importantly, as the plot shows, this relationship does not hold up when HPO terms are randomised, since the box plots for the randomised HPO-implicated tissues ('HPO-tissue_rand') and background ('Background_rand') almost perfectly overlap.

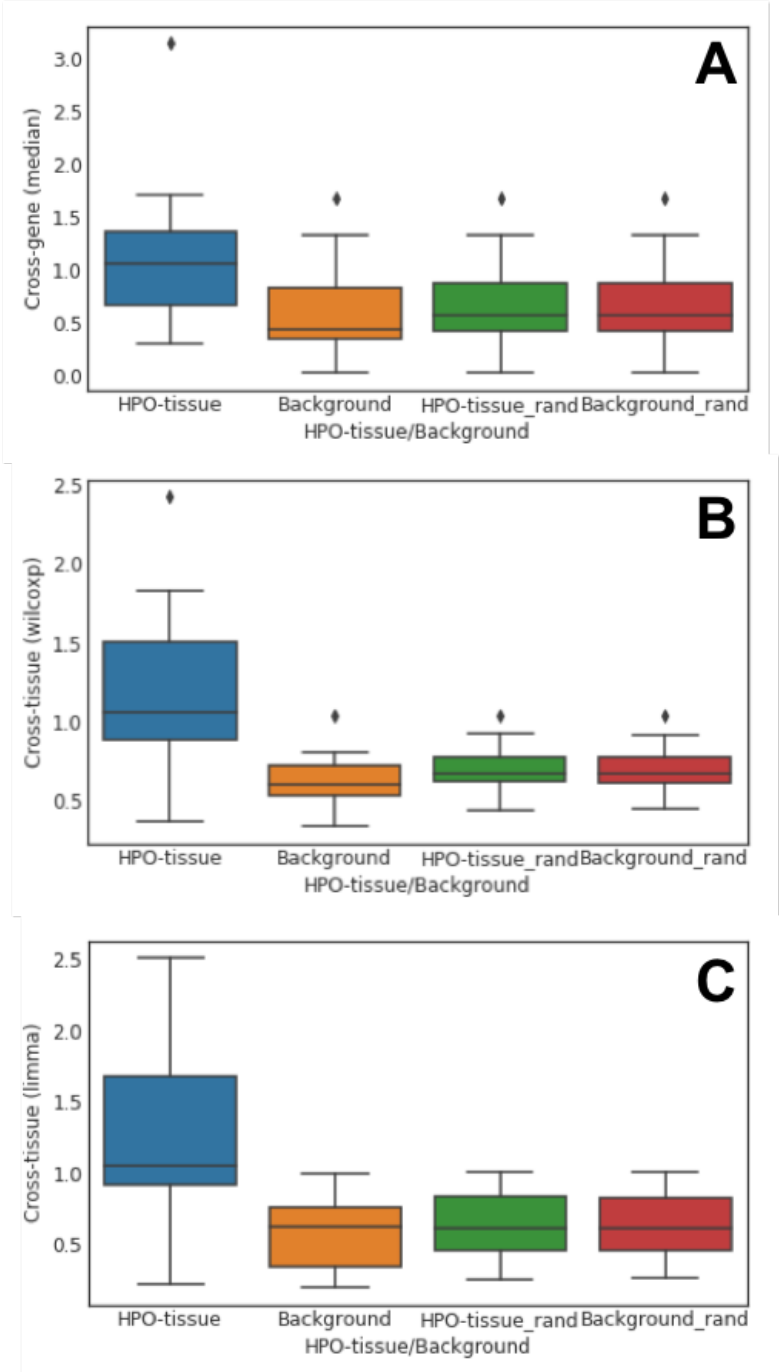


Fig. 5.2 **Expression scores of disease-causing genes confirmed in HICF2 cases.** The box plots show the expression level of a confirmed disease-causing gene for the HICF2 cases in tissues indicated by the patient's phenotype, compared to tissues not indicated by the patient's phenotype. Panel A shows the cross-gene score (median), while panel B and panel C show the cross-tissue scores, calculated with wilcoxp and limma respectively (see Section 4.2.2.3 for details). In each plot, the blue box represents the distribution of the respective expression score for the disease-causing genes confirmed in each HICF2 case for the tissues indicated by the patient's HPO terms. The yellow box shows the respective mean expression score in all tissues not implicated by the patient's HPO terms. The green box shows the respective expression scores for randomly assigned tissues replacing the patient's real HPO-implicated tissues. The red box shows the background signal for gene-specific expression in tissues that were assumed to not be affected in the random tissue allocation.

5.3.2 Case results

5.3.2.1 Summary analysis

To test if the results from Section 5.3.1 translate into improved variant rankings, I applied GPET to the eleven HICF2 patient cases. Figure 5.3 shows the summary results. Exomiser's hiPHIVE ranked more variants first, in the top 5, and top 20 than GPET for cases where the disease-causing genes were annotated in the HPO at the time of analysis. GPET performs poorly on cases with HPO-annotated benchmark genes due to the large number of variants contained in real patient VCFs that are up-ranked by GPET's expression features, despite the promising results presented in Chapter 4. However, the inverse is true for non-HPO-annotated candidate genes. For the four cases in question, GPET ranks one variant first, two in the top 5 and three in the top 20, as opposed to Exomiser with zero variants ranked first, one ranked in the top 5, and two in the top 20.

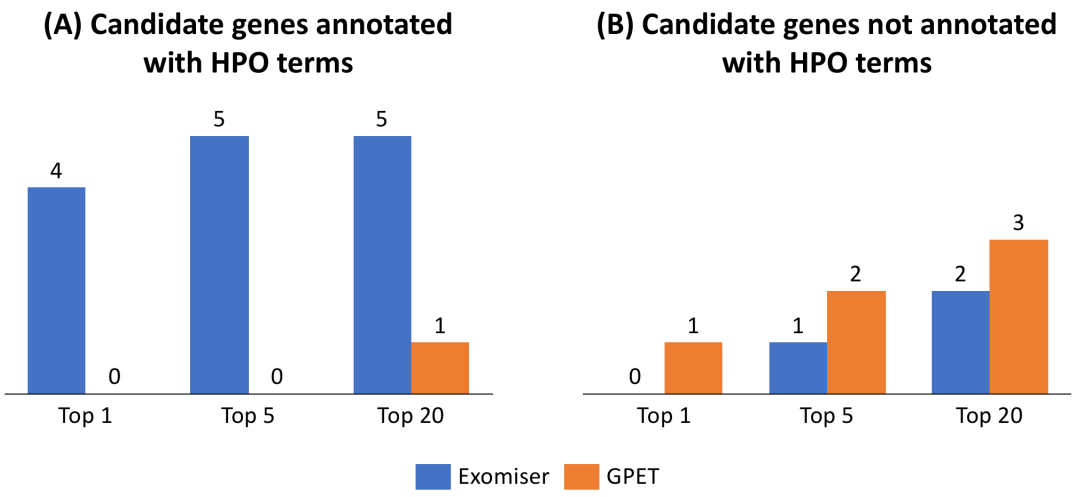


Fig. 5.3 **Summary comparison of GPET and Exomiser on HICF2 patient cases.** The bar charts represent the number of disease-causing variant ranked first, in the top 5 and top 20 respectively by the two ranking algorithms Exomiser (hiPHIVE) and GPET. Plot A shows the results for all genes that were annotated with HPO terms at the time of analysis, while plot B shows the results for genes that were not annotated with HPO terms at the time of analysis. Plot A shows that Exomiser ranked more variants highly for genes that were annotated in the HPO, while the inverse is true for plot B.

Table 5.2 shows a detailed summary of all case results. Exomiser ranks all variants in HPO-annotated genes higher, while GPET improves ranking results for all variants in non-HPO-annotated genes.

Gene	HPO-annotated	Exomiser ranking	GPET ranking
TNNI2	yes	1	19
WWOX	yes	1	571
ACTC1	yes	3	76
PSTPIP1	yes	1	229
POR	yes	23	64
SLC30A10	yes	35	1572
RBPJ	yes	1	33
CACNA1E	no	12387	5154
SAMD9L	no	5	1
DOCK11	no	12	4
HDLBP	no	52	20

Table 5.2 **Comparison of Exomiser's and GPET's performance on HICF2 cases.** Exomiser ranks all candidate variants in genes that were annotated in the December 9th, 2013 version of the HPO higher, while GPET improves ranking results for all non-annotated genes.

5.3.2.2 Individual case results

5.3.2.2.1 Cases with HPO-annotated candidate genes Figures 5.4, 5.5, 5.6, 5.7, 5.8, 5.9 and 5.10 show the ranking results for all cases with known candidate genes that were previously annotated with HPO terms. Each plot shows a manhattan representation of Exomiser's variant score, phenotype score, and combined score, as well as the combined GPET score. As explained in Chapter 3, the variant score serves to create an almost binary split of variants into pathogenic and benign variants. All disease-causing variants sit at the or close to the top of that distribution, receiving consistently high pathogenicity scores. For these HPO-annotated candidate genes, the phenotype score works well to separate the candidate gene from the two primary clusters of candidate variants at a score of 0.0 and 0.5. A phenotype score of 0 is assigned when no annotations are available for a gene of interest in the HPO, MPO, a

zebfrafish-specific phenotype database or the PPI network. A score of 0.5 is assigned as a function of the gene's annotation in the PPI network. The analysis in Chapter 3 showed that ≈62% of all variants receive a score of 0.5. Chapter 3 contains a detailed explanation of how the phenotype score is calculated. The combination of the two scores in Exomiser's combined score produces a well-defined separation of variants that are both pathogenic and likely relevant for the phenotype, and those that are not. As a result, Exomiser's hiPHIVE works very well for HPO-annotated candidate genes, ranking the candidate variants consistently high. In those cases, GPET performs worse than Exomiser and down-ranks variants compared to Exomiser's hiPHIVE.

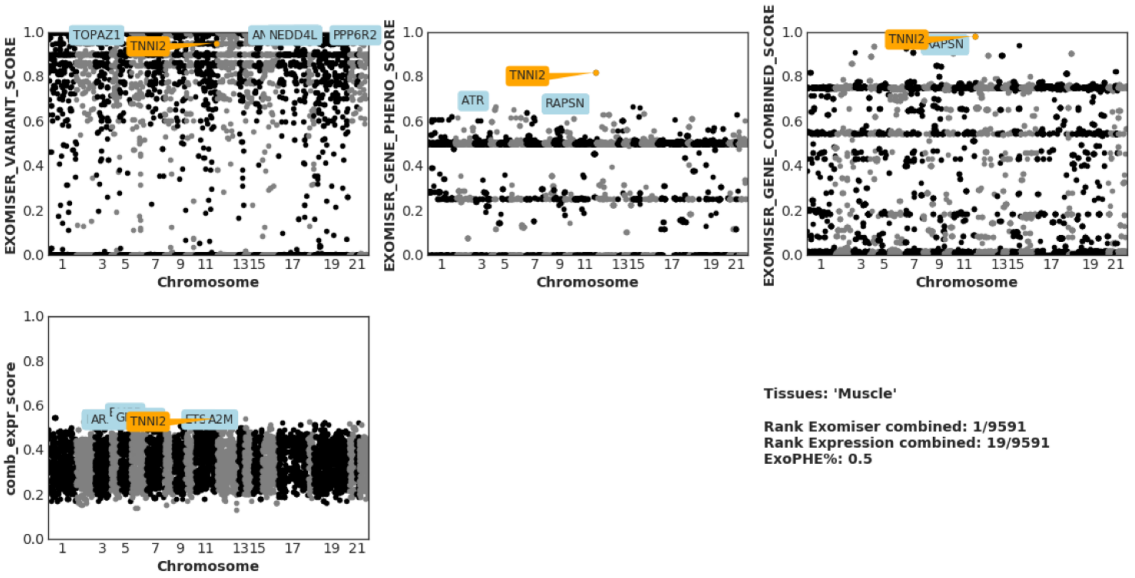


Fig. 5.4 ***TNNI2* results for Exomiser's hiPHIVE and GPET.** The plot shows Exomiser's hiPHIVE algorithm's variant score ('EXOMISER_VARIANT_SCORE'), phenotype score ('EXOMISER_GENE_PHENO_SCORE'), and combined score ('EXOMISER_GENE_COMBINED_SCORE'), as well as the GPET score ('comb_expr_score). Exomiser's hiPHIVE algorithm's combined score ranks the candidate variant in *TNNI2* first, compared to rank 19 with GPET. Based on the patient's HPO profile, the tissue 'Muscle' is relevant for the patient's condition, for which expression scores and BTLs were generated.

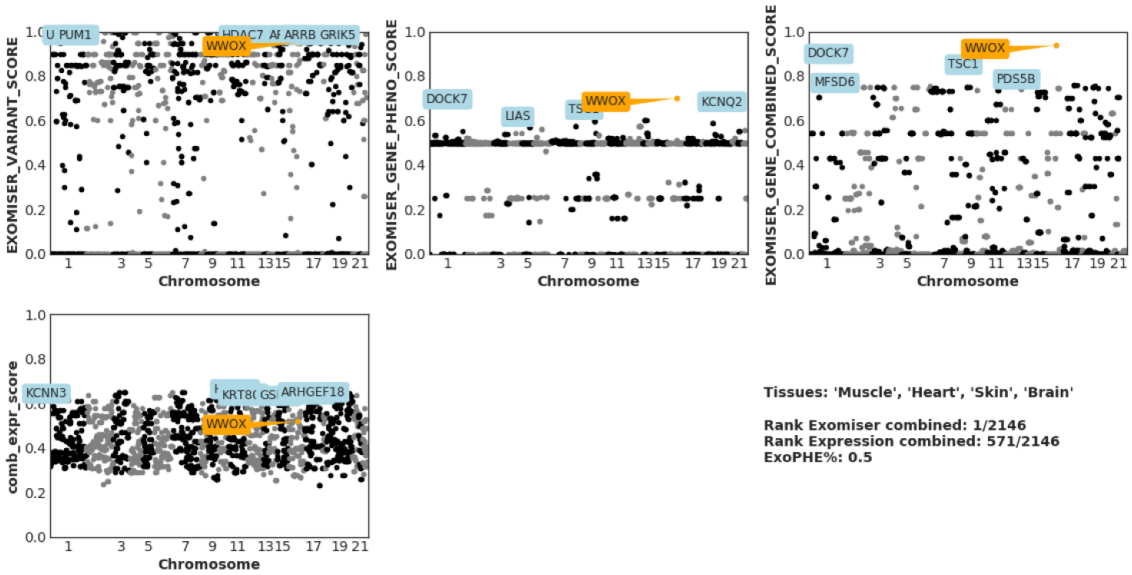


Fig. 5.5 ***WWOX* results for Exomiser's hiPHIVE and GPET.** The plot shows Exomiser's hiPHIVE algorithm's variant score ('EXOMISER_VARIANT_SCORE'), phenotype score ('EXOMISER_GENE_PHENO_SCORE'), and combined score ('EXOMISER_GENE_COMBINED_SCORE'), as well as the GPET score ('comb_expr_score). Exomiser's hiPHIVE algorithm's combined score ranks the candidate variant in *WWOX* first, compared to rank 571 with GPET. Based on the patient's HPO profile, the tissues 'Muscle', 'Heart', 'Skin', and 'Brain' are relevant for the patient's condition, for which expression scores and BTLs were generated.

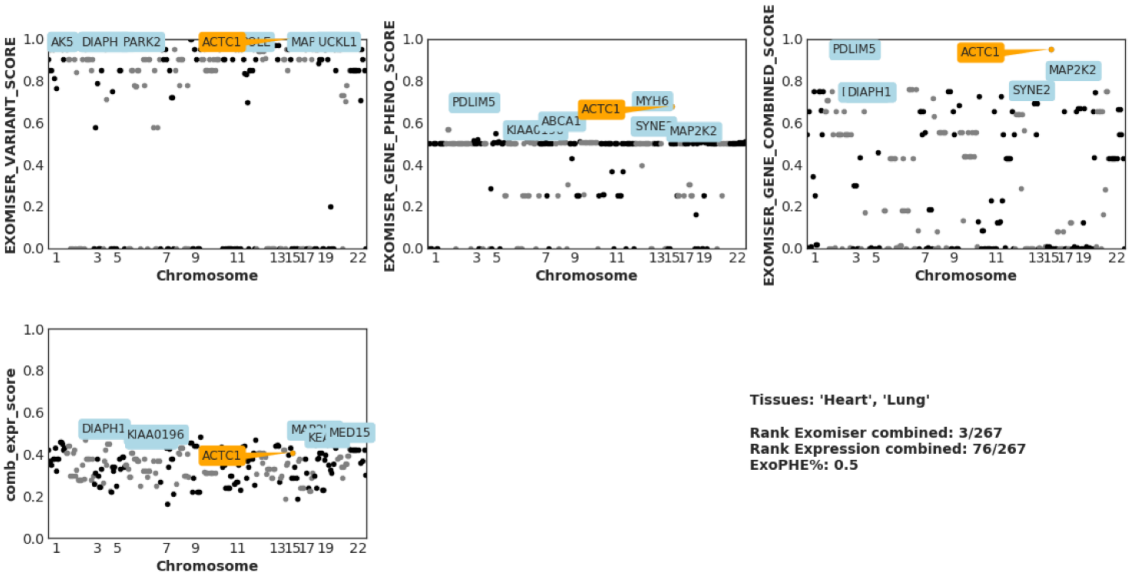


Fig. 5.6 ***ACTC1* results for Exomiser's hiPHIVE and GPET.** The plot shows Exomiser's hiPHIVE algorithm's variant score ('EXOMISER_VARIANT_SCORE'), phenotype score ('EXOMISER_GENE_PHENO_SCORE'), and combined score ('EXOMISER_GENE_COMBINED_SCORE'), as well as the GPET score ('comb_expr_score). Exomiser's hiPHIVE algorithm's combined score ranks the candidate variant in *ACTC1* third, compared to rank 76 with GPET. Based on the patient's HPO profile, the tissues 'Heart' and 'Lung' are relevant for the patient's condition, for which expression scores and BTLs were generated.

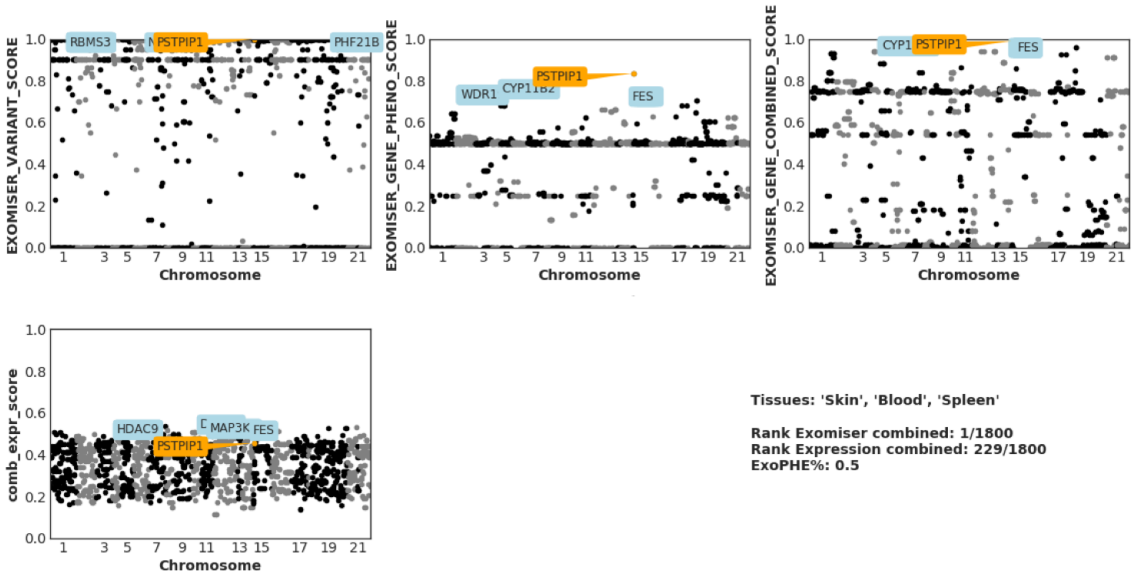


Fig. 5.7 ***PSTPIP1* results for Exomiser's hiPHIVE and GPET.** The plot shows Exomiser's hiPHIVE algorithm's variant score ('EXOMISER_VARIANT_SCORE'), phenotype score ('EXOMISER_GENE_PHENO_SCORE'), and combined score ('EXOMISER_GENE_COMBINED_SCORE'), as well as the GPET score ('comb_expr_score). Exomiser's hiPHIVE algorithm's combined score ranks the candidate variant in *PSTPIP1* first, compared to rank 229 with GPET. Based on the patient's HPO profile, the tissues 'Skin, 'Blood', and 'Spleen' are relevant for the patient's condition, for which expression scores and BTLs were generated.

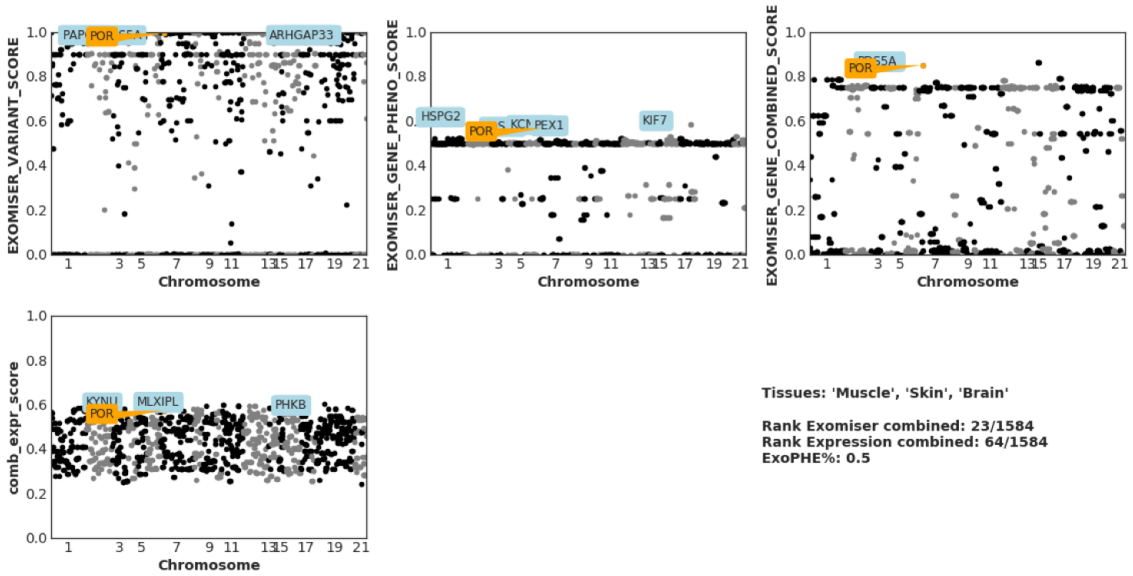


Fig. 5.8 ***POR* results for Exomiser's hiPHIVE and GPET.** The plot shows Exomiser's hiPHIVE algorithm's variant score ('EXOMISER_VARIANT_SCORE'), phenotype score ('EXOMISER_GENE_PHENO_SCORE'), and combined score ('EXOMISER_GENE_COMBINED_SCORE'), as well as the GPET score ('comb_expr_score). Exomiser's hiPHIVE algorithm's combined score ranks the candidate variant in *POR* 23rd, compared to rank 64 with GPET. Based on the patient's HPO profile, the tissues 'Muscle', 'Skin', and 'Brain' are relevant for the patient's condition, for which expression scores and BTLs were generated.

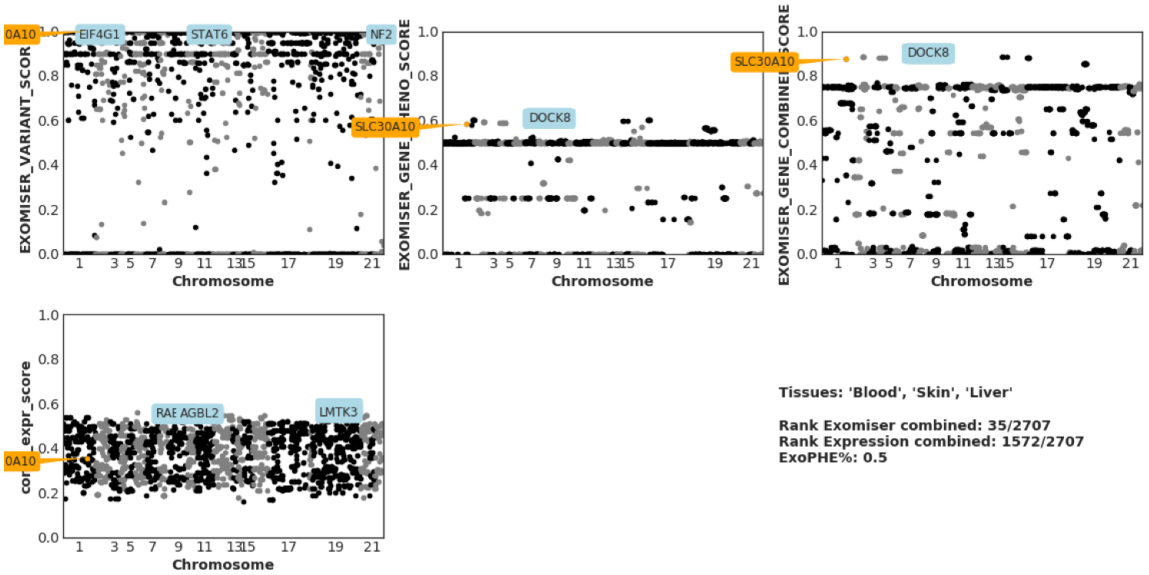


Fig. 5.9 ***SLC30A10* results for Exomiser's hiPHIVE and GPET.** The plot shows Exomiser's hiPHIVE algorithm's variant score ('EXOMISER_VARIANT_SCORE'), phenotype score ('EXOMISER_GENE_PHENO_SCORE'), and combined score ('EXOMISER_GENE_COMBINED_SCORE'), as well as the GPET score ('comb_expr_score). Exomiser's hiPHIVE algorithm's combined score ranks the candidate variant in *SLC30A10* 35th, compared to rank 1572 with GPET. Based on the patient's HPO profile, the tissues 'Blood', 'Skin', and 'Liver' are relevant for the patient's condition, for which expression scores and BTLs were generated.

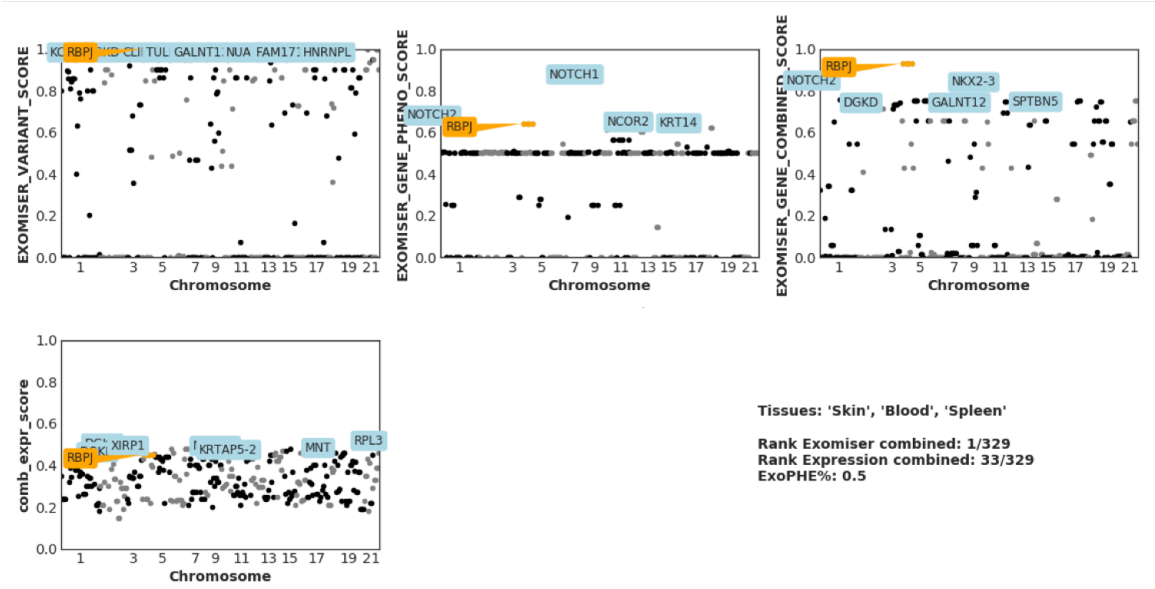


Fig. 5.10 ***RBPJ* results for Exomiser's hiPHIVE and GPET.** The plot shows Exomiser's hiPHIVE algorithm's variant score ('EXOMISER_VARIANT_SCORE'), phenotype score ('EXOMISER_GENE_PHENO_SCORE'), and combined score ('EXOMISER_GENE_COMBINED_SCORE'), as well as the GPET score ('comb_expr_score). Exomiser's hiPHIVE algorithm's combined score ranks the candidate variant in *RBPJ* first, compared to rank 33 with GPET. Based on the patient's HPO profile, the tissues 'Skin', 'Blood', and 'Spleen' are relevant for the patient's condition, for which expression scores and BTLs were generated.

5.3.2.2.2 Cases with non-HPO-annotated candidate genes Exomiser's hiPHIVE's variant score ranks all disease-causing variants high. However, the phenotype score cannot differentiate the disease-causing variant from all other variants. For *CACNA1E*, *SAMD9L*, and *DOCK11*, Exomiser's hiPHIVE cannot find any phenotypic annotations in the HPO, MPO, and a zebrafish phenotype database, but finds an entry in the PPI and thus assigns a phenotype score of 0.5. *HDLBP*, however, is not annotated by the PPI, and thus the candidate variant receives a phenotype score of 0. In these cases, the phenotype score does not provide as much support for variant classification as it did for the HPO-annotated candidate genes. Here, GPET is helpful for improving the candidate variants' ranking. hiPHIVE ranks the candidate variants 12387th, 5th, 12th, and 52nd for *CACNA1E*, *SAMD9L*, and *DOCK11*, and *HDLBP* respectively, in contrast to 5154th, 1st, 4th and 20th for GPET (see Figures 5.11, 5.12, 5.13, and 5.14).

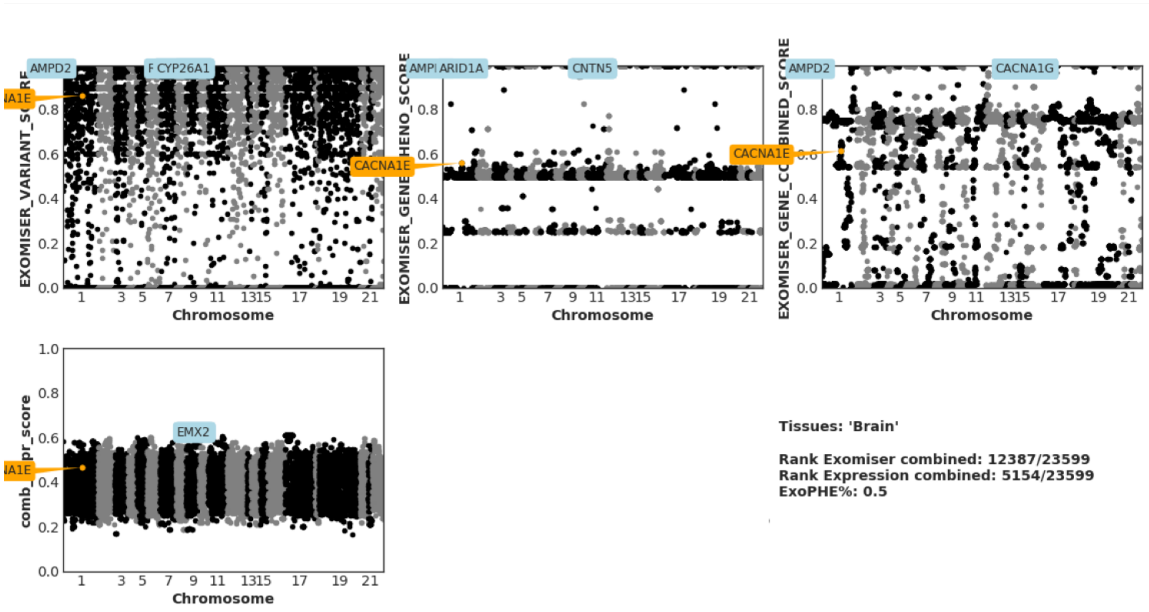


Fig. 5.11 ***CACNA1E* results for Exomiser's hiPHIVE and GPET.** The plot shows Exomiser's hiPHIVE algorithm's variant score ('EXOMISER_VARIANT_SCORE'), phenotype score ('EXOMISER_GENE_PHENO_SCORE'), and combined score ('EXOMISER_GENE_COMBINED_SCORE'), as well as the GPET score ('comb_expr_score). Exomiser's hiPHIVE algorithm's combined score ranks the candidate variant in *CACNA1E* 12387th, compared to rank 5154 with GPET. Based on the patient's HPO profile, the tissue 'Brain' is relevant for the patient's condition, for which expression scores and BTLs were generated.

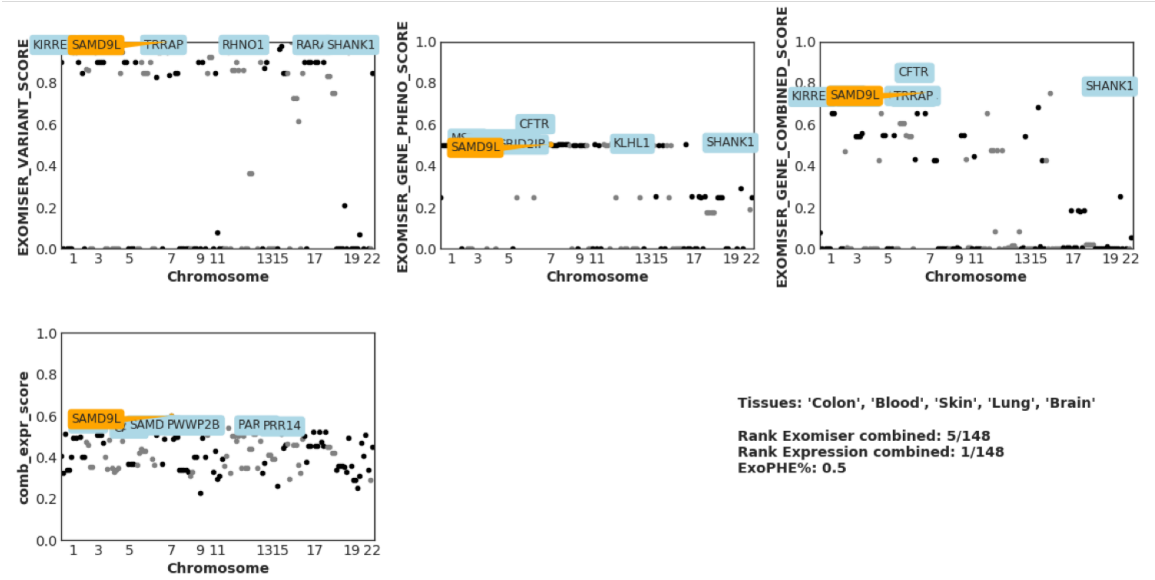


Fig. 5.12 ***SAMD9L* results for Exomiser's hiPHIVE and GPET.** The plot shows Exomiser's hiPHIVE algorithm's variant score ('EXOMISER_VARIANT_SCORE'), phenotype score ('EXOMISER_GENE_PHENO_SCORE'), and combined score ('EXOMISER_GENE_COMBINED_SCORE'), as well as the GPET score ('comb_expr_score). Exomiser's hiPHIVE algorithm's combined score ranks the candidate variant in *SAMD9L* 5th, compared to rank 1 with GPET. Based on the patient's HPO profile, the tissues 'Colon', 'Blood', 'Skin', 'Lung', and 'Brain' are relevant for the patient's condition, for which expression scores and BTLs were generated.

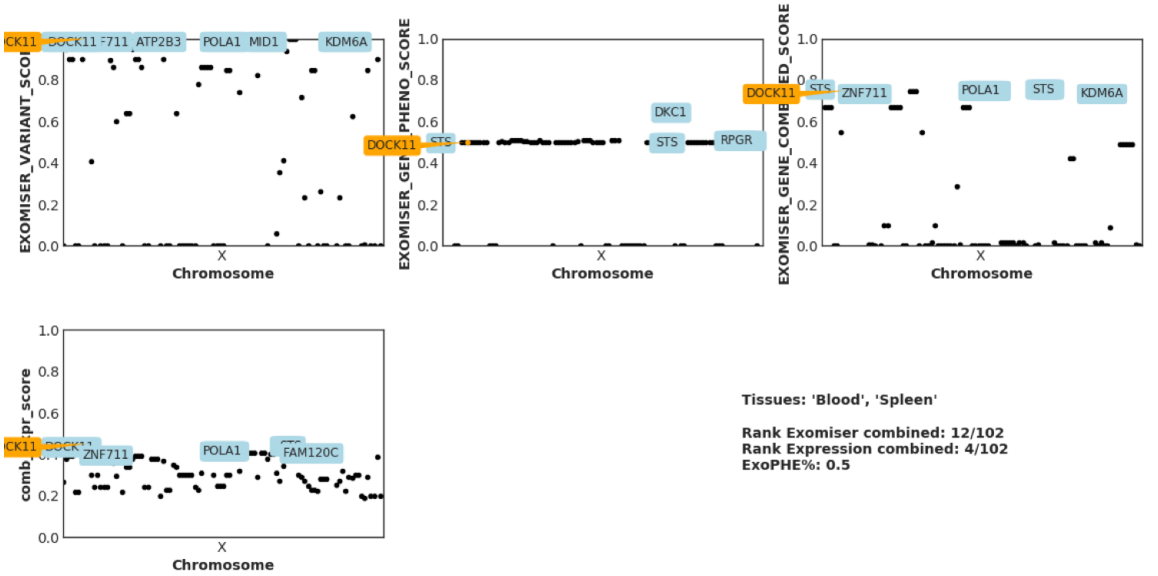


Fig. 5.13 ***DOCK11* results for Exomiser's hiPHIVE and GPET.** The plot shows Exomiser's hiPHIVE algorithm's variant score ('EXOMISER_VARIANT_SCORE'), phenotype score ('EXOMISER_GENE_PHENO_SCORE'), and combined score ('EXOMISER_GENE_COMBINED_SCORE'), as well as the GPET score ('comb_expr_score). Exomiser's hiPHIVE algorithm's combined score ranks the candidate variant in *DOCK11* 12th, compared to rank 4 with GPET. Based on the patient's HPO profile, the tissues 'Blood' and 'Spleen' are relevant for the patient's condition, for which expression scores and BTLs were generated.

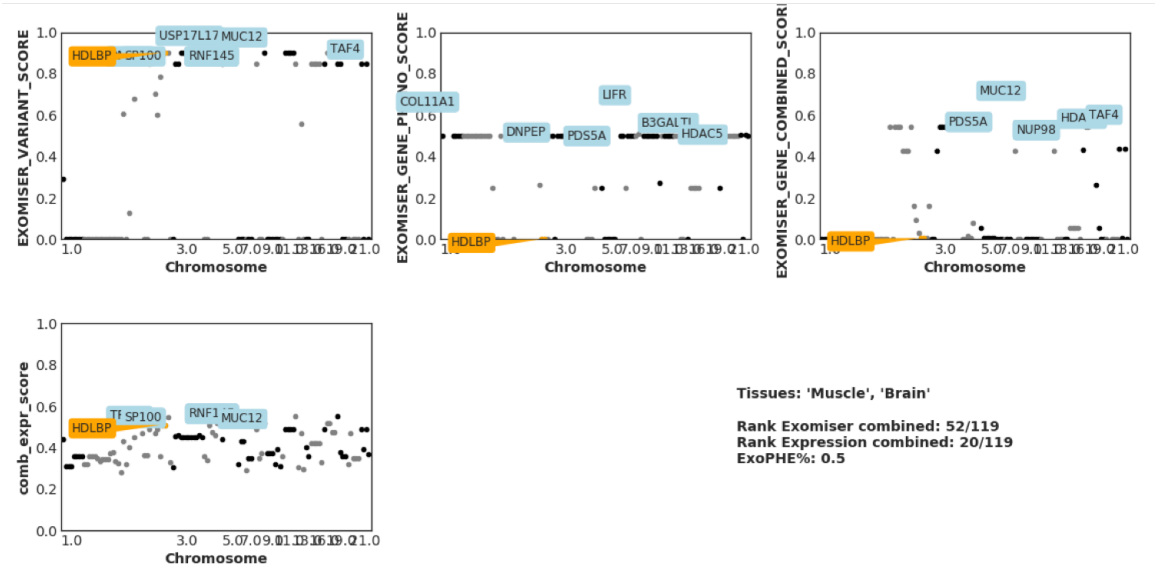


Fig. 5.14 ***HDLBP* results for Exomiser's hiPHIVE and GPET.** The plot shows Exomiser's hiPHIVE algorithm's variant score ('EXOMISER_VARIANT_SCORE'), phenotype score ('EXOMISER_GENE_PHENO_SCORE'), and combined score ('EXOMISER_GENE_COMBINED_SCORE'), as well as the GPET score ('comb_expr_score). Exomiser's hiPHIVE algorithm's combined score ranks the candidate variant in *HDLBP* 52nd, compared to rank 20 with GPET. Based on the patient's HPO profile, the tissues 'Muscle' and 'Brain' are relevant for the patient's condition, for which expression scores and BTLs were generated.

5.4 Discussion

The motivation for the GPET algorithm was to use expression data to improve variant prioritisation for rare genetic disease cases, in particular for novel genes without phenotypic annotations. A benchmarking analysis of GPET against Exomiser conducted in Chapter 4 on an *in silico* dataset showed that using expression data can improve upon the AUC achieved by Exomiser for variant prioritisation (AUC of 0.95 for GPET compared to 0.91 for Exomiser for imperfect phenotypic annotations, see Section 4.3.3.2 for details). In this chapter, I tested if the algorithm's performance on *in silico* data translates to real patient cases.

I first examined if the patients' HPO profiles can be used to infer which tissues the disease-causing genes are likely highly expressed in. The results illustrated in Section 5.3.1 demonstrate that genes harbouring disease-causing variants achieve approximately two times higher expression scores in tissues assumed-to-be-affected based on the patients' HPO terms than in tissues that are likely not affected by the disease. These results are meaningful as they, to my knowledge, for the first time confirm a link between the HPO and tissue-specific gene expression in real patient cases. Furthermore, this data suggests that BTLs and tissue-specific expression data can be used to indicate which genes are likely disease-causing.

To confirm if these results scale to larger patient populations, the algorithm should be tested on a cohort like the 100,000 Genomes Project or the DDD study. Both studies have built a database of confirmed disease-causing variants and linked HPO terms that can be used for this purpose. Additionally, results should be validated by analysing the expression of confirmed disease-causing genes in tissues predicted to be affected by a patient's HPO terms compared to likely non-affected tissues.

To test if the use of tissue-specific expression data results in an improved ranking for disease-causing variants in real patient data, the GPET framework was compared

to Exomiser's hiPHIVE algorithm on eleven well-characterised HICF2 cases. GPET worsened the results for candidate genes that were already HPO-annotated at the time of analysis compared to Exomiser (Exomiser vs. GPET, ranked first: 4 vs. 0, top five: 5 vs. 0, top 20: 5 vs. 1). However, GPET improved ranking results for non-HPO-annotated candidate genes (Exomiser vs. GPET, ranked first: 0 vs. 1, top five: 1 vs. 2, top 20: 2 vs. 3).

Of the approximately 7,000 rare genetic diseases, molecular diagnoses only exist for approximately 4,000 conditions [5]. GPET could be a helpful tool for novel disease gene discovery. To simulate a test for more previously solved patient cases, one could manually delete phenotypic annotations for the candidate genes in the version of the HPO used by Exomiser's hiPHIVE algorithm prior to the analysis. That way, it's possible to simulate the case of a non-HPO-annotated gene.

To implement GPET in bioinformatics labs, the algorithm could be used for cases where tools like Exomiser do not identify the disease-causing variant. Similar to PHIVE and PhenIX, GPET could be used as an additional analysis option in frameworks like Exomiser.

In addition to providing interesting data to support the GPET hypothesis, these results also provide further evidence supporting *HDLBP* as a candidate gene for FLS, as described in Chapter 3. *HDLBP* is further functionally examined in the following chapter.

Chapter 6

Functional validation of *HDLBP* for Fine-Lubinsky syndrome

6.1 Introduction

In vitro and in vivo functional studies are an important tool to determine if a variant has a damaging effect on a gene or the gene product. The ACMG guidelines published in 2015 categorise data from well-established functional studies as strong evidence of pathogenicity used to classify pathogenic and likely pathogenic variants [67]. Particularly for cases with genes that are novel for the observed phenotype, functional data is relevant to establish a link between the variant and the phenotype. I previously discussed a case from the HICF2 study consisting of five consanguineous Pakistani patients (Figure 6.1 describes the pedigree) affected by FLS (see Section 3.3.2.2.3). FLS is a rare developmental disorder with a complex phenotype, including plagiocephaly, megalocornea, digital abnormalities, cleft palate, facial dysmorphism, structural brain abnormalities, and sometimes deafness [147].

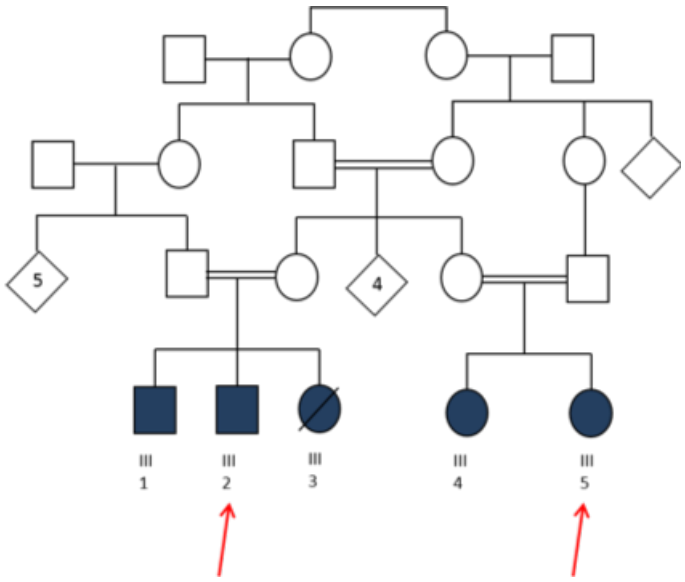


Fig. 6.1 **Pedigree of Pakistani family affected by Fine-Lubinsky syndrome.** Five consanguineous patients (shown in blue) from two different branches of a Pakistani family are affected by FLS. DNA from all five patients was analysed with a cytoSNP12 array and WGS was conducted for two of the five affected individuals (highlighted here with red arrows).

The five patients, from two different branches of the pedigree, are affected by a subset of the FLS phenotype, including plagiocephaly, facial dysmorphism, camptodactyly, moderate developmental delay, and megalocornea (see Section 3.3.2.2.3 for a detailed phenotypic description). Samples from the affected individuals were previously run on a cytoSNP12 array by Dr Alistair Pagnamenta, a post-doctoral researcher in the Taylor group, resulting in the identification of a single 5 Mb region on 2q37.3 shared in all 5 subjects (≈1/600 of the genome) (see Figure 6.2).

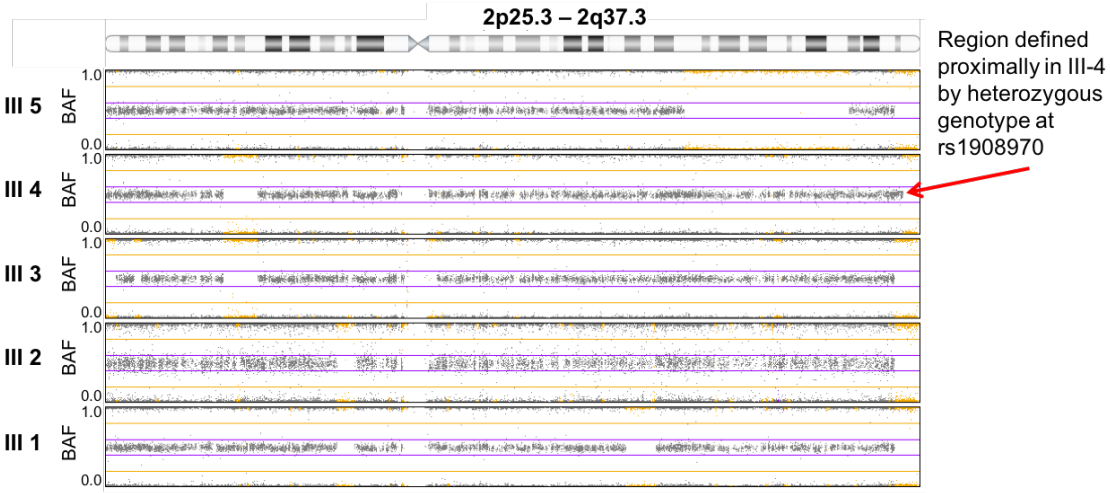


Fig. 6.2 **cytoSNP12 array data of individuals in Pakistani family affected by FLS.** DNA from all five patients was analysed with a cytoSNP12 array by Dr Alistair Pagnamenta, which identified a single 5 Mb region on 2q37.3 with loss of heterozygosity shared in all affected individuals. The plot shows B-allele frequencies (BAF) for all five affected individual from the family pedigree. Each dot represents one SNP. Homozygous SNP have a BAF of 0 or 1, while heterozygous SNP have a BAF of 0.5. This figure was created by Dr Alistair Pagnamenta.

WGS was conducted for two of the five affected individuals and two candidate variants were identified by Dr Alistair Pagnamenta in the 78 RefSeq genes in the shared region: a homozygous missense variant (p.E374K) in *HDAC4* and a homozygous variant (c.1731+1G>A) in *HDLBP* that was suspected to lead to an alternatively spliced gene with in-frame skipping of exon 14. At the time of analysis, loss of function variants in *HDLBP* were rare in ExAC and no homozygotes had been reported [224]. The identified *HDAC4* missense variant occured in the South Asian ExAC cohort with a MAF of 0.47%, which is considerably higher than the population frequency of the very rare FLS. Both variants were predicted to be pathogenic by MutationTaster (score = 1.0) [225].

To investigate if the *HDLBP* variant would lead to alternative splicing, Dr Alistair Pagnamenta used the reverse transcription polymerase chain reaction (RT-PCR) technique combined with Sanger sequencing on the sample of one affected family member, one parent, and an independent control. The results demonstrated that the c.1731+1G>A variant leads to in-frame skipping of exon 14 in *HDLBP* (see Figure 6.3).

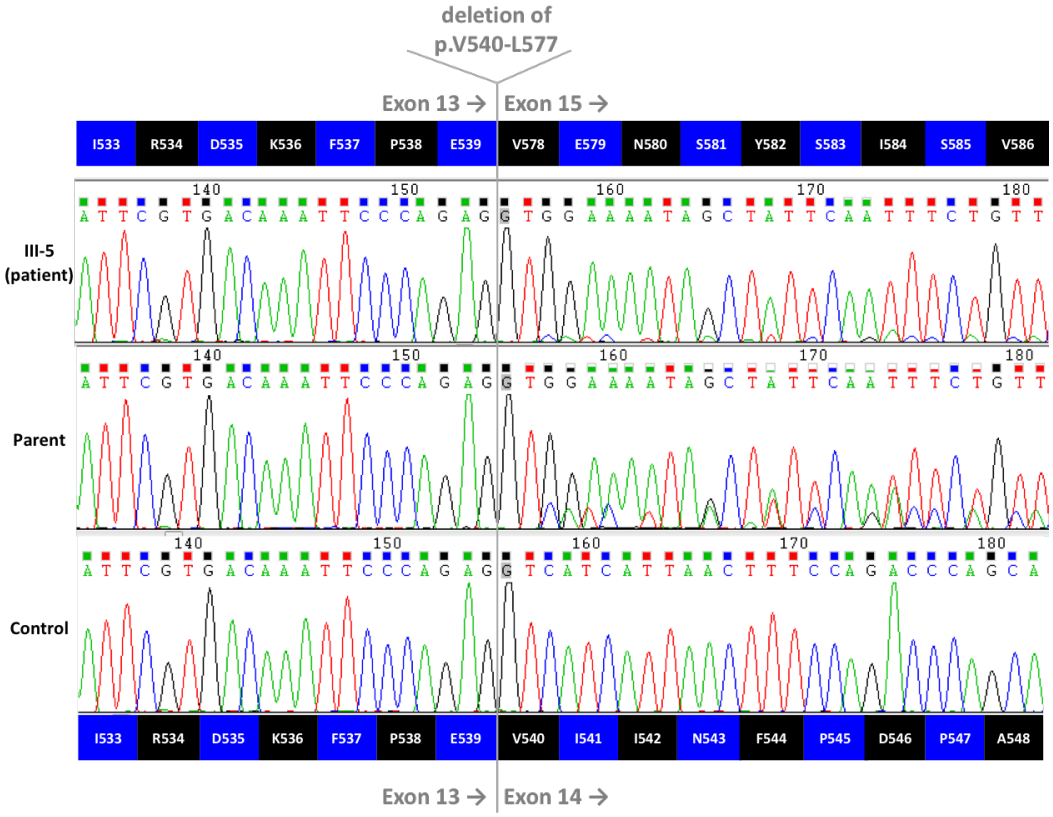


Fig. 6.3 **RT-PCR and Sanger sequencing result of *HDLBP* variant.** Dr Alistair Pagnamenta used RT-PCR and Sanger sequencing to test if the c.1731+1G>A variant in *HDLBP* would lead to alternative splicing. The electropherograms in this figure demonstrate that exon 14 is spliced out, leading to a deletion of p.V540-L577 in one of the affected family members (III 5) and one of the patients' parents (Parent) as compared to an independent control (Control). This figure was created by Dr Alistair Pagnamenta.

2q37 deletions have been shown to be associated with brachydactyly mental retardation syndrome (BDMR), which has a partially, but not completely overlapping phenotype with FLS, including brachydactyly, facial dysmorphism, low-set ears, intellectual deficiency, seizures, obesity, and short stature [155]. Williams *et al.* [226] suggested haploinsufficiency of *HDAC4* as the cause of BDMR, based on seven patients out of which two carry *de novo* missense variants in *HDAC4* and five have overlapping deletions which only overlap for *HDAC4*. A three-generation family reported on by Villavicencio-Lorini *et al.* [227, 228] in which the proband, her mother, and her maternal grandmother were affected by mild developmental delay and dysmorphic facial features, but did not present with brachydactyly type E, however, calls Williams *et al.*'s findings into question. The phenotype in the three patients was associated with an inherited heterozygous interstitial deletion of chromosome 2q37.3 and included the genes *HDAC4*, *FLJ43879*, and *TWIST2*. Three related patients with *HDAC4* haploinsufficiency described by Wheeler *et al.* [229] further contradict Wiliams' findings, since the patients are only affected by brachydactyly type E, but have non-dysmorphic facial features and normal intelligence. While previously viewed as a strong candidate for BDMR, *HDAC4* variants have been reclassified as variants of unknown significance by OMIM [228].

As stated previously, the support in the literature for *HDAC4* as a candidate for BDMR-related phenotypic features is inconclusive and only one of the five patients (patient 2) presented with brachydactyly, a key feature of BDMR. That patient also carries the homozygous missense variant in *HDAC4*. In contrast to that, patient five, also shown to be a homozygous carrier for the *HDAC4* missense variant, did not present with brachydactyly. The patients, however, are homozygous for both the *HDAC4* and the *HDLBP* variants, which would be expected to result in a more severe phenotype than heterozygous variants. In addition to the pathogenicity evidence supporting the *HDLBP* variant as a candidate, *HDLBP* lies in deleted regions in several patients

affected by BDMR. Felder *et al.* [156] showed that *HDLBP*, together with *FARP2* and *PASK*, is significantly down-regulated in lymphoblastoid cell lines of a patient with autism and BDMR and his parents. Given the evidence supporting the *HDLBP* splice-site variant as a candidate for FLS, the LoF variant in *HDLBP* was considered to be the strongest candidate.

Exon 14 of *HDLBP*, in-frame skipping of which is investigated as a potential cause of FLS in this chapter, makes up ≈51% of the RNA-binding KH6 domain of vigilin, the protein encoded by *HDLBP*, and vigilin's RNA-binding activity has been widely described [149]. Figure 6.4 shows the gene structure of *HDLBP*, including the 14 KH domains and exon 14. In particular, vigilin plays an essential role in tRNA transport from the nucleus to the cytoplasm as part of the nuclear and cytoplasmic vigilin-containing complexes (VCCn and VCCc) [149], which contain tRNA. Vigilin is also a key regulator in proliferating cells, which play a key role in embryonic development [149]. Since previous reports supporting *HDAC4* as a causative gene for the FLS-related phenotype have proven to be inconclusive, solving this case could not only provide the affected family with a diagnosis, it could also aid in delineating the root cause of the complex FLS phenotype. To my knowledge, no patients with deleterious variants in *HDLBP* and an associated phenotype have previously been described. **The goal of this chapter is thus to investigate the impact of the patient's splice-site variant on vigilin's RNA-binding activity and to characterise the link between *HDLBP* and FLS further.**

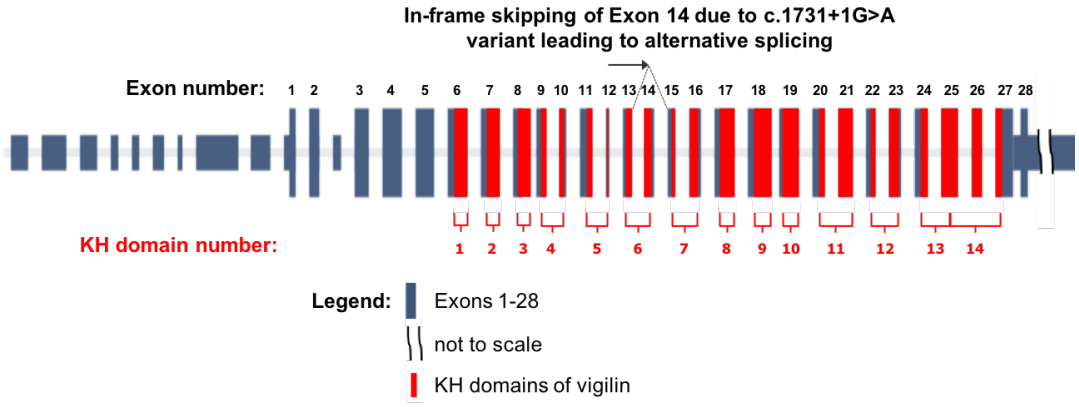


Fig. 6.4 **Gene structure of *HDLBP* with splice-site variant in relation to functional domains.** *HDLBP* consists of 28 exons (blue) and encodes for the protein vigilin, an RNA-binding protein with 14 KH domains (red). A homozygous variant, c.1731+1G>A, was identified in the patient family described in this chapter, leading to alternative splicing with in-frame skipping of exon 14.

6.2 Materials and methods

To test the effect of the candidate variant on vigilin's RNA-binding activity and exclude other potential effects on the gene product, several assays were conducted. I examined effects on vigilin's protein stability and intra-cellular protein localisation, investigated the simulated 3D structure of vigilin, and measured vigilin's RNA-binding activity.

First, I examined if the candidate variant impacted the stability of the *HDLBP* gene or the vigilin protein. To do this, cDNA fragments of the *HDLBP* wildtype and *HDLBP* mutant were cloned into an eGFP-containing plasmid and transfected into HeLa and HEK293 cells (see Section 2.3.1 for details). Figure 6.5 shows the cDNA fragments of *HDLBP* wildtype and *HDLBP* mutant. The corresponding plasmid maps for *HDLBP* wildtype and mutant are shown in Figure 6.6 and Figure 6.7. PCR using primers to amplify exon twelve to 16 of *HDLBP* was performed by Dr Pamela Kaisaki (see Section 2.3.2 for details), a post-doctoral researcher in the Taylor group, to assess if cloning and transfection were successful.

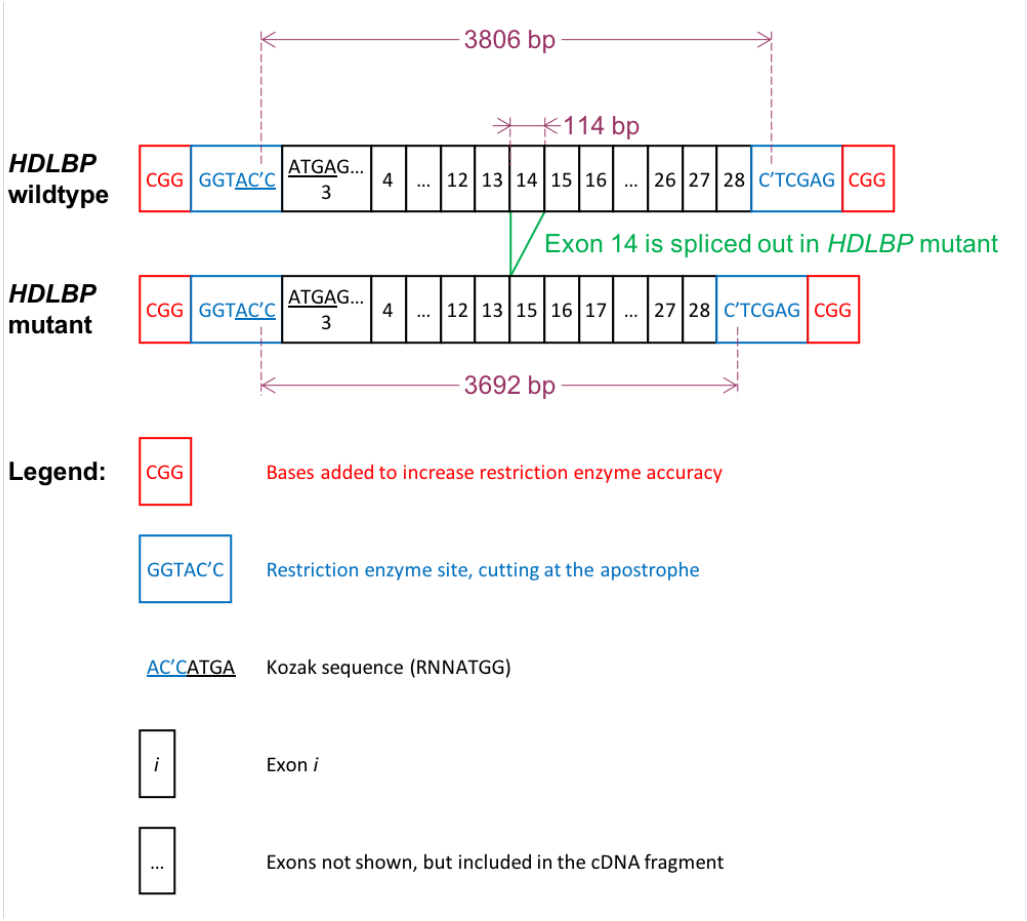


Fig. 6.5 **cDNA fragments of *HDLBP* wildtype and *HDLBP* mutant**. cDNA fragments were created for *HDLBP* wildtype and *HDLBP* mutant. Both constructs spanned from the start of exon three to the end of exon 28, with the *HDLBP* mutant construct skipping exon 14 to account for alternative splicing. Restriction enzyme sites were added for the restriction enzymes XhoI and KpnI following the Kozak sequence. Additional bases were added on both ends of each construct to increase restriction enzyme accuracy (red). The length of the cDNA fragments from restriction enzyme site to restriction enzyme site is 3806 bp and 3692 bp respectively for *HDLBP* wildtype and *HDLBP* mutant. Exon 14 is 114 bp long. cDNA fragment sequences are included in the Appendix.

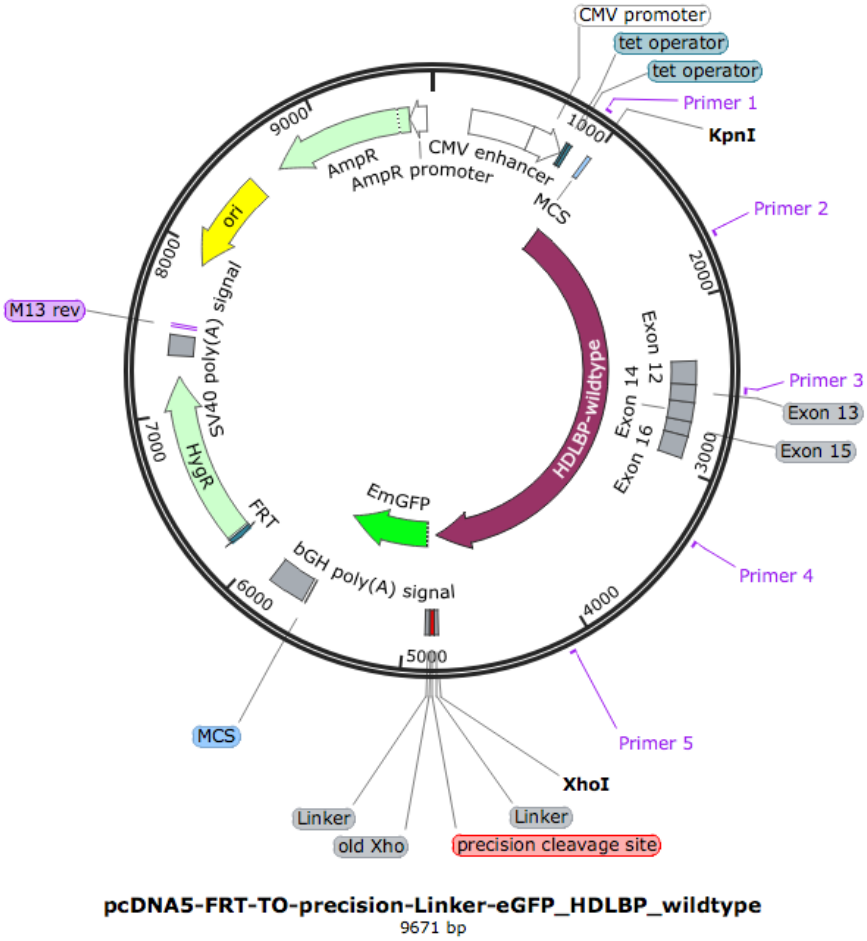


Fig. 6.6 **Plasmid map for *HDLBP* wildtype**. This map shows the eGFP-containing plasmid that the *HDLBP* wildtype cDNA fragment was cloned into. The restriction enzyme sites for KpnI and XhoI are highlighted, as well as the primers ('HDLBP-12F' and 'HDLBP-16R') used for the PCR amplification of exons twelve to 16 (grey) of *HDLBP* and primers used for quality control ('Primer 1' through 'Primer 5'). The expected PCR amplicon size is 423 bp.

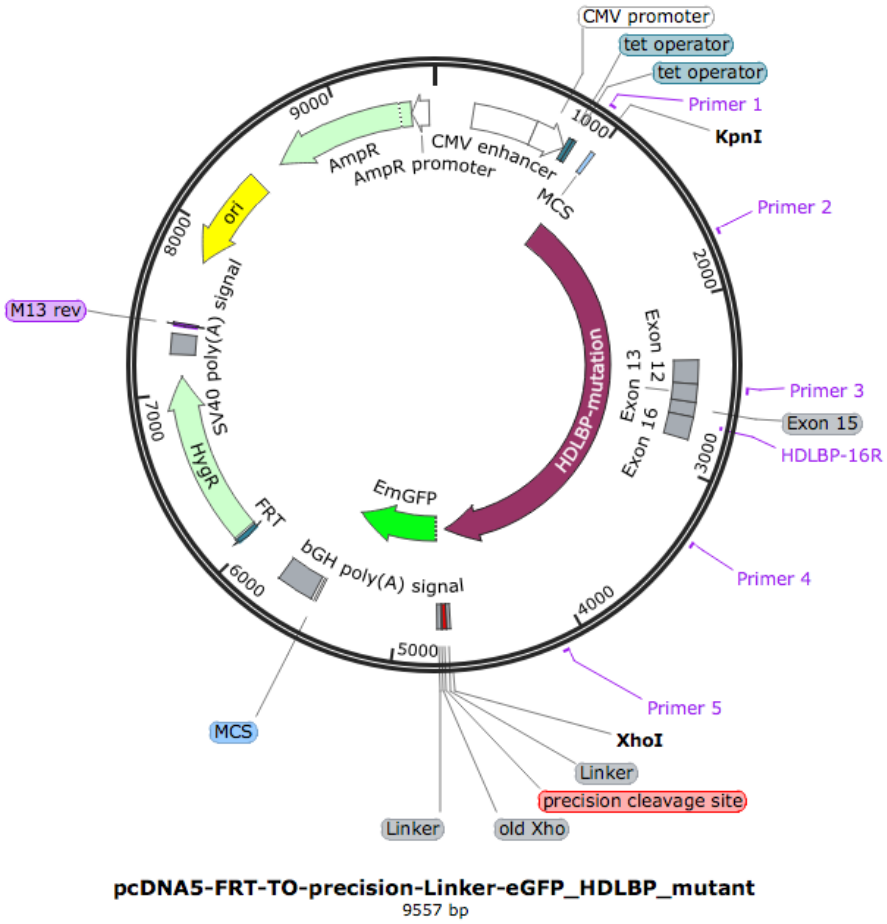


Fig. 6.7 **Plasmid map for *HDLBP* mutant**. This map shows the eGFP-containing plasmid that the *HDLBP* mutant cDNA fragment was cloned into. The restriction enzyme sites for KpnI and XhoI are highlighted, as well as the primers ('HDLBP-12F' and 'HDLBP-16R') used for the PCR amplification of exons twelve to 16 (grey) of *HDLBP*, excluding the skipped exon 14, and primers used for quality control ('Primer 1' through 'Primer 5'). The expected PCR amplicon size is 309 bp.

The cells were cultured and induced by doxycycline to produce wildtype and mutant vigilin-GFP fusion proteins.

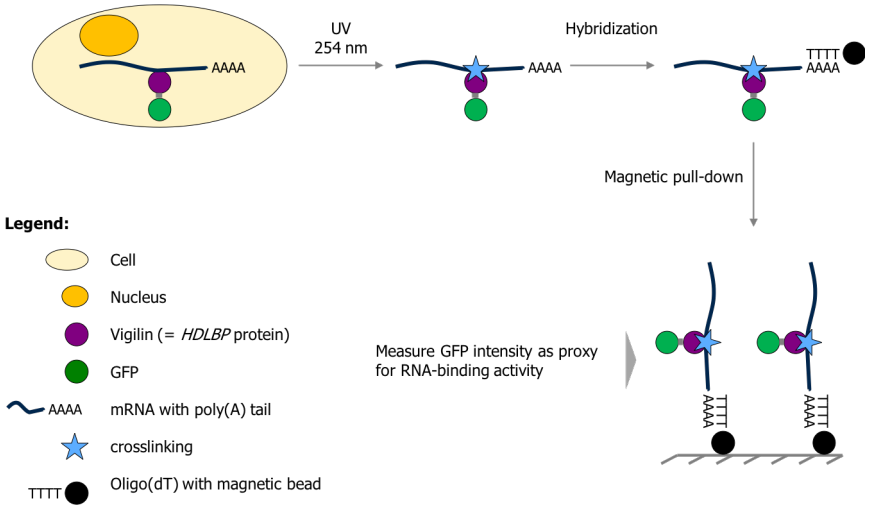
Western Blots of vigilin were performed to exclude protein instability as a cause of the phenotype (see Section 2.3.3 for details). Furthermore, the time-dependent protein decay of wildtype and mutant vigilin was inferred by measuring the GFP fluorescence

intensity of the wildtype and mutant-GFP fusion proteins over 48h in a plate reader post expression induction with doxycycline (see Section 2.3.4 for details).

Second, I tested the effect of the candidate variant on the intra-cellular localisation of vigilin using fluorescence microscopy (described in detail in Section 2.3.5). Wildtype vigilin is expected to be predominantly located in the cytoplasm [149].

Third, to visualise the impact of the splice-site variant on the vigilin mutant's ability to bind RNA, vigilin's tertiary structure was plotted. The tertiary structure simulation and analysis was conducted by Dr Matteo Ferla, a post-doctoral researcher in the Taylor group (see Section 2.3.6 for details).

Fourth, the polyadenylated RNA-binding activity of the vigilin-GFP fusion protein, for both mutant and wildtype, was assessed by hybridisation with an oligo(dT) probe, a method adapted from Strein *et al.* [230] (see Section 2.3.7 for details). The amount of RNA bound by both mutant and wildtype vigilin was estimated by measuring the intensity of the GFP in a plate reader. *HNRNPQ*, a known RNA-binding protein, and GFP, which does not bind RNA, served as positive and negative controls respectively. Figure 6.8 describes the general concept of the RNA-binding assay.



.

Fig. 6.8 **RNA binding activity assay**. The GFP-tagged vigilin is cross-linked with mRNA under UV radiation and cells are lysed. oligo(dT) magnetic beads are hybridised with the poly(A) tails of the cross-linked mRNA and pulled down with a magnet. The GFP intensity of the pulled down complex is measured as a proxy for the protein's RNA binding activity. This figure was adapted from Strein *et al.* [230]

6.3 Results

In this section, results are shown for the assessments focused on quality control (see Section 6.3.1), protein stability (see Section 6.3.2), intra-cellular protein localisation (see Section 6.3.3), simulated tertiary structure (see Section 6.3.4), and RNA-binding activity based on oligo(dT) capture (see Section 6.3.5).

6.3.1 Quality control

PCR of mutant and wildtype *HDLBP* showed a clear difference in size corresponding to the 114 bp long skipped exon 14 (see Figure 6.9), confirming that cloning and transfection were successful.

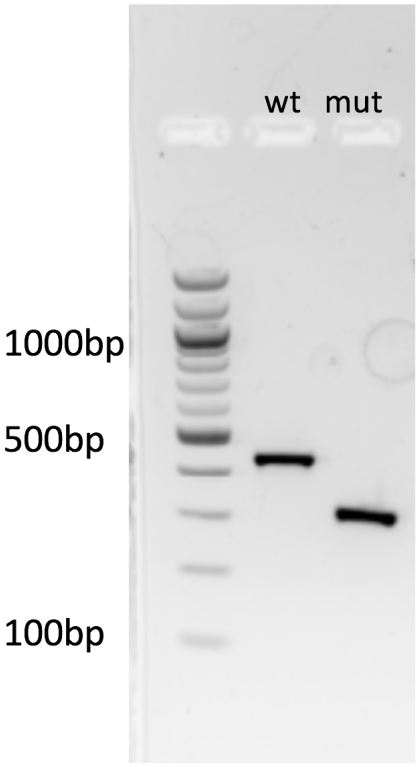


Fig. 6.9 **PCR of mutant and wildtype *HDLBP*.** PCR of mutant and wildtype *HDLBP* was performed with primers that amplify exon 12 - 16. A clear size difference between wildtype and mutant *HDLBP* is visible, corresponding to the 114 bp long skipped exon 14.

Western Blots showed that the two proteins were stable and that the expressed wildtype and mutant vigilin-GFP fusion proteins have the expected sizes in the established HeLa and HEK293 cell lines. The vigilin mutant combined with the eGFP-tag is expected to be approximately 170 kD, as compared to 174 kD for the vigilin wildtype (see Figure 6.10).

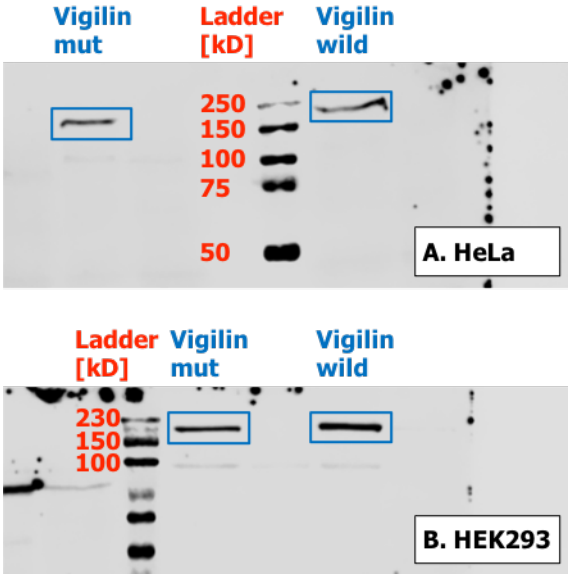


Fig. 6.10 **Western blot of *HDLBP*(vigilin) wildtype/mutant expressed in A. HeLa and B. HEK293 cells.** Both proteins were stable and the relative sizes match the expectations. The eGFP-tag is expected to be 33 kD, the vigilin mutant 137 kD, and the vigilin wildtype 141 kD.

6.3.2 Protein stability

While a marginal difference is visible between the protein decay of wildtype and mutant vigilin over 48h, with the mutant decaying slightly faster than the wildtype, the error bars of the analysis are overlapping and the difference of the concentration change over 48h is within the margin of error, thus not providing evidence for a differentiated protein decay between wildtype and mutant vigilin (see Figure 6.11).

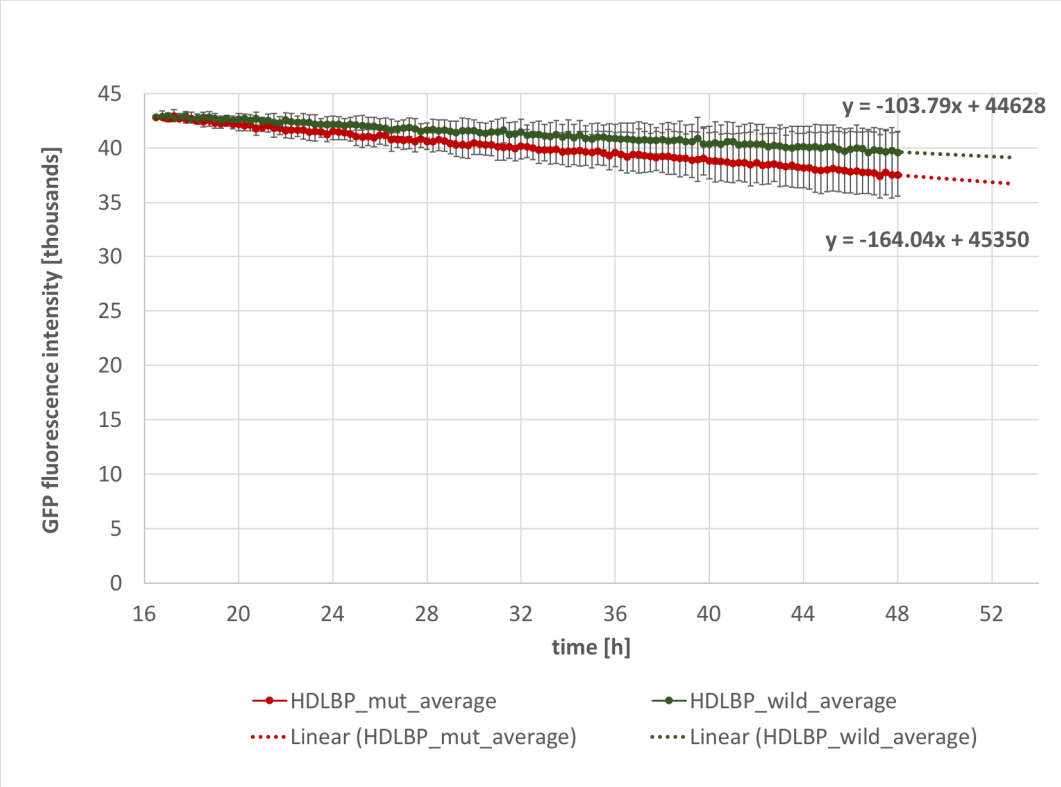


Fig. 6.11 **GFP fluorescence-based protein decay analysis of vigilin wildtype (green, n=6) and mutant (red, n=5) in a plate reader over 48h.** The plot shows the GFP fluorescence intensity as a function of time. The GFP fluorescence intensity of the wild type and mutant vigilin-GFP fusions proteins is used to infer the protein decay of wild type and mutant vigilin. At -164.04, the slope of the vigilin mutant's protein decay curve is slightly steeper than that of the vigilin wildtype with -103.79. However, all data points for both the wildtype and the mutant are well within the margin of error of the experiment, as indicated by the error bars. Thus, no meaningful difference in the decay of both proteins was detected.

6.3.3 Intra-cellular protein localisation

Intra-cellular protein localisation was examined via fluorescence microscopy of the GFP tag fused to the vigilin wildtype and mutant. Wildtype vigilin is expected to be primarily expressed in the cytoplasm [149], as is confirmed in panels A and C of Figure 6.12. The candidate variant did not lead to a difference in intra-cellular localisation for the vigilin mutant, as is confirmed in panels B and D of Figure 6.12.

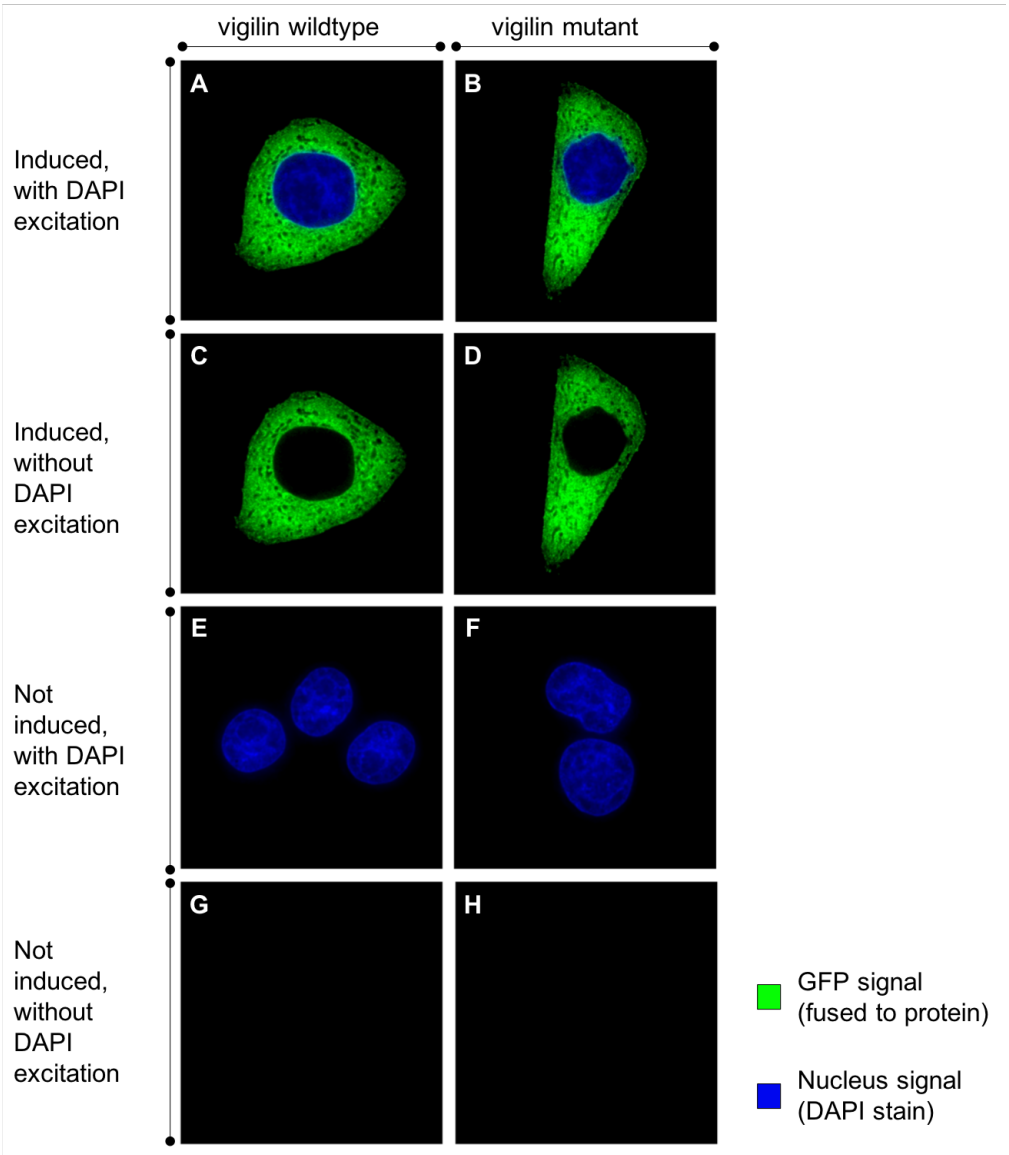


Fig. 6.12 **Fluorescence microscopy of vigilin expressed in HeLa cells.** Panels A to D show HeLa cells in which expression was induced with doxycyclin for vigilin wildtype (A, C) and mutant (B, D) with and without DAPI excitation. Panels E to H show uninduced HeLa cells in which expression was not induced for vigilin wildtype (E, G) and mutant (F, H) with and without DAPI excitation. The green eGFP signal showing vigilin's localisation and the blue DAPI signal indicating the location of the nucleus show no difference in localisation between the vigilin mutant and wildtype. Both mutant and wildtype vigilin are predominantly present in the cytoplasm, as is expected based on Cheng *et al.* [149].

6.3.4 Tertiary structure analysis of the KH6 domain of vigilin

Vigilin's tertiary structure was simulated to examine the effect of the skipping of exon 14 on the RNA-binding KH6 domain. Vigilin's KH6 domain contains the minimal KH motif shared between all KH domains, consisting of two alpha helices enclosed by two beta strands ($\beta\alpha\alpha\beta$), followed by a variable loop, another beta strand and an alpha helix (see green structure in Figure 6.13). Between the two alpha helices of the minimal KH motif sits the GXXG motif that defines a classical KH domain [149]. The β strands form an antiparallel sheet adjacent to the three α helices. The α-GXXG-α element, together with the $\beta 2$ strand and the variable loop, form a hydrophobic groove. This groove can bind to a nucleic acid tetramer, the sequence of which is specified by hydrogen bonding to amino acid side chains [149]. The hydrogen bond interactions are known to favour adenine and cytosine, but not guanine [149]. The tertiary structure simulation in Figure 6.13 shows that exon 14 skipping causes a loss of the $\beta 2$ strand, the variable loop, the $\beta 3$ strand, and part of the $\alpha 3$ helix, resulting in the loss of the KH6 domain's tertiary structure and thereby the hydrophobic groove, which is suspected to lead to a decrease in RNA binding.

6.3.5 oligo(dT) capture to quantify RNA-binding activity of vigilin

An oligo(dT) capture method was used to quantify the RNA-binding activity of the vigilin mutant and the vigilin wildtype. hnRNPQ, a known RNA-binding protein, was used as a positive control. GFP, which does not bind RNA, was used as a negative control. The oligo(dT) capture showed considerably reduced RNA-binding activity of the vigilin mutant-GFP fusion protein compared to the wildtype (see Figure 6.14). The highest and lowest RNA-binding activities were measured for hnRNPQ and GFP respectively.

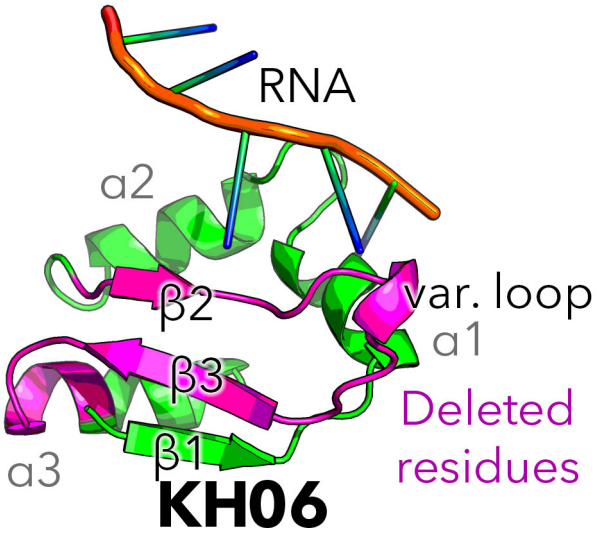


Fig. 6.13 **Tertiary structure of vigilin's KH6 domain and the spliced-out exon 14.** 3D modeling of *HDLBP*'s KH6 domain shows the loss of the $\beta 2$ strand, the variable loop, the $\beta 3$ strand, and part of the $\alpha 3$ helix (all shown in purple), resulting in the loss of the KH6 domain's tertiary structure and thereby the hydrophobic groove, which is suspected to lead to a decrease in RNA binding due to in-frame skipping of exon 14.

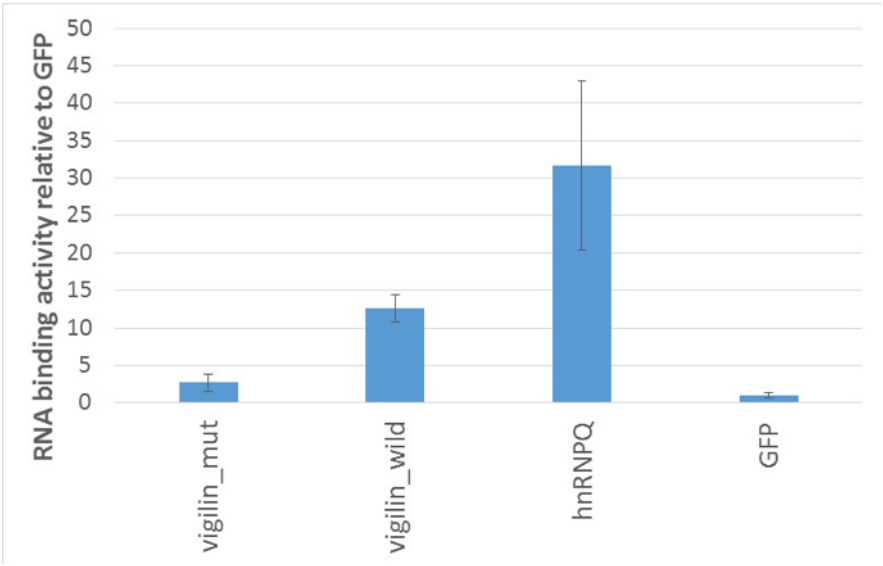


Fig. 6.14 **Polyadenylated RNA-binding activity of vigilin-GFP fusion protein (for mutant and wildtype).** The polyadenylated RNA-binding activity of the vigilin-GFP fusion protein shows a significantly reduced RNA-binding activity for the vigilin mutant as compared to the vigilin wildtype. hnRNPQ, a known RNA-binding protein, serves as a positive control. GFP, which does not bind RNA, serves as a negative control that all signals were normalised to. For each protein, two biological repeats with each three technical repeats, for a total of six repeats, were created.

6.4 Discussion

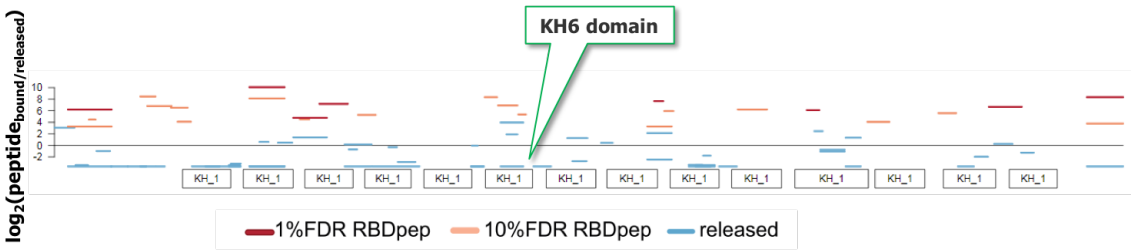
My preliminary data suggest that intra-cellular protein localisation and protein stability can likely be excluded as a potential cause of the phenotype. Fluorescence microscopy showed no differentiation in the intra-cellular protein localisation of wildtype and mutant vigilin. Additionally, vigilin's half-life is expected to be approximately 29h [231] and the protein decay experiment was run for 48h, during which no meaningful difference in the decay of vigilin wildtype and mutant was detected. The oligo(dT) capture demonstrated that the RNA-binding activity of the vigilin mutant is considerably reduced compared to the wildtype, despite the fact that only one of vigilin's 14 KH domains is affected by the spliced out exon 14.

It is however important to note that, due to repeated washing steps in the oligo(dT) capture protocol for each sample, the fluorescence values cannot be used to compare absolute RNA binding activity levels of the vigilin mutant and the wildtype. Thus, the assay does not imply that the vigilin mutant's RNA binding activity is reduced by approximately 80%[1] compared to the wildtype.

Nonetheless, the drop in RNA binding activity from wildtype to mutant suggests that the KH6 domain plays an important role in RNA binding for vigilin. Abolition of RNA binding due to the loss of a subset of KH domains in vigilin has previously been shown for the truncation of the C-terminal most KH13 and KH14 domains [149]. A study conducted by Castello *et al.* [232] provides further evidence for the importance of the KH6 domain's RNA binding. The authors identified RNA-binding domains in proteins in human cells with a mass spectrometry-based approach. In their assay, RNA-binding proteins (RBP) are covalently bound to RNA via UV irradiation, pulled down using polyadenylated oligo(dT) capture, and cleaved into small peptides through protease digestion. The peptides still bound to RNA are pulled down again using

[1] $1 - (RNA\_binding\_activity(vigilin\_mutant)/RNA\_binding\_activity(vigilin\_wildtype)) \approx 80\%$

oligo(dT) capture and mass spectrometry is used to identify the individual peptides. The approach can be used to quantify the amount of each peptide bound to RNA, compared to the amount of peptide released in the process. The ratio of the amount of peptide bound vs released ($peptide_{bound/released}$) thus gives an indication of the RNA binding activity of each peptide sequence, which can be mapped to the original RBP sequence, thereby offering insight into the RNA binding activity of individual protein domains. The result for the vigilin wildtype pre-computed by Castello *et al.* [232] can be readily downloaded from the RBDmap server [233] and is shown in Figure 6.15. The KH6 domain is enclosed by the KH5 and KH7 domains, which, according to Castello *et al.*'s [232] assay, bind less RNA. Thus, it is possible that RNA species that bind close to KH5, KH6, and KH7 are particularly reliant on the KH6 domain. A related effect has previously been shown for the truncation of the C-terminal most KH13 and KH14 domains in Scp160p, vigilin's yeast homolog, which led to the abolition of RNA binding due to the loss of a subset of the protein's KH domains [149]. Hirschmann *et al.* [234] showed that interaction of Scp160p with a specific subset of mRNAs requires the conserved KH13 and KH 14 domains. Similarly, it is possible that other types of RNA require a conserved KH6 domain to interact with vigilin.



.

Fig. 6.15 **Distribution of the peptide fragments of the RNA-bound regions of the RBP and released fragments in vigilin.** The x axis represents the protein sequence from N to C terminus, and the y axis shows the RNA-bound/released peptide intensity ratios. Individual lines represent peptide fragments for false discovery rates of 1% (red) and 10% (blue) respectively. This figure was adapted from Castello *et al.* [232] and can be downloaded from the RBDmap server [233]

Vigilin has over 700 potential mRNA targets and has been observed to influence various stages of RNA metabolism, including mRNA transport, nuclear-cytoplasmic tRNA shuttling, translation, degradation, and formation of stress granules and processing bodies [149]. The oligo(dT) capture method used in this chapter gives an indication of the quantity of RNA bound by the proteins of interest. The method, however, does not provide insight into the types of RNA species that are not bound by the protein of interest anymore due to reduced RNA-binding activity, which is a key limitation of this approach. To further investigate the relationship between the *HDLBP* splice-site variant and the FLS phenotype, it is therefore necessary to conduct a qualitative comparison of the RNA species bound by wildtype and mutant vigilin. This comparison is currently being conducted by Dr Pamela Kaisaki, a post-doctoral researcher in the Taylor Group.

The RNA-binding assays conducted by Dr Kaisaki include:

1. Cross-linking immunoprecipitation (CLIP) with either anti-GFP or anti-*HDLBP* antibodies, followed by RT-PCR of known targets of *HDLBP*, to identify differences between wildtype and mutant,

2. CLIP followed by sequencing (CLIP-seq) to determine the range of RNA species differentially bound, and

3. CLIP followed by mass-spectrometry (iCLIP), which allows identification of the RNA-recognition element.

Simultaneously, CRISPR-Cas9 will be used to introduce the *HDLBP* splice-site variant into the genomes of HEK293 and induced pluripotent stem cell (iPS cell) lines. RNA-binding activity will be measured in the edited HEK293 cells to see whether differences observed previously in target RNA binding in the stably transfected HEK293 cells are confirmed. The data presented in this chapter were generated using HeLa and HEK293 cells, which are derived from cervical cancer and embryonic

kidney cell lines. Transferability of insights drawn from these cell lines for FLS, a developmental disease, needs to be assessed. Therefore, the edited iPS cells will be differentiated into three lineages of endoderm, mesoderm, and ectoderm cells, to detect potential differences between wildtype and mutant *HDLBP* cells, and provide insights into potential links between the developmental symptoms of FLS and defects caused by mutant *HDLBP* in early development.

The work presented in this Chapter suggests that one effect of the patients' splice-site variant in *HDLBP* is a quantitative reduction in RNA-binding activity. The next steps outlined in the previous paragraphs will qualitatively characterise how reduced RNA-binding activity can lead to the FLS phenotype to potentially confirm *HDLBP* as a disease gene for FLS.

Chapter 7

Discussion and Conclusion

This chapter concludes the thesis, beginning with a summary of results (see Section 7.1), a commentary on future work (see Section 7.2), and some concluding remarks (see Section 7.3).

7.1 Summary of results

In this thesis, I examined the performance of existing algorithms for the analysis of WGS data of RGD patients, developed a new method for variant analysis, validated this new method in real RGD patient cases, and worked on the functional validation of *HDLBP* as a novel disease gene candidate for Fine-Lubinsky syndrome.

In Chapter 3, I compared the performance of variant prioritisation algorithms for WGS data based on purely genotypic data with those based on genotypic and phenotypic data for two analysis frameworks, Exomiser's hiPHIVE algorithm and VAAST+Phevor. Furthermore, I analysed the change in performance of Exomiser's hiPHIVE for two different time points and versions of the algorithm, two years apart from each other in 2016 and 2018, and versions 7.2.1 and 11.0.0. For the analysis, I used eleven rare genetic disease patient cases for which the disease-causing variant

had previously been identified in the HICF2 study, which consisted of singletons and trios, different inheritance patterns, including autosomal dominant/*de novo*, autosomal recessive, and X-linked recessive. For eight of the patient cases, the disease-causing variant lies in a gene known for the patient's observed phenotype and those genes were thus annotated in the HPO. The remaining three genes are novel candidates for the patient phenotype, had not been published at the time of analysis and were not annotated in the HPO. All analysed algorithms successfully ranked a significant number of variants at the top of their respective distributions. The GPAs overall outperformed the GAs. Five candidate variants were ranked first and six in the top five by Exomiser's hiPHIVE's genotype-and-phenotype-based combined score, compared to Exomiser's genotype-based variant score, which ranked four variants first and four in the top five. Similarly, the genotype-based VAAST algorithm ranked one variant first and eight variants in the top five, outperformed by the genotype-and-phenotype-based VAAST+Phevor combination, which ranked eight variants first and nine in the top five. For the eleven patient cases analysed in this chapter, VAAST+Phevor outperforms Exomiser's hiPHIVE, but a larger sample size is necessary to draw definitive conclusions about population-scale performance. As for the time series analysis, the newer version of Exomiser ranked more benchmark variants first, in the top 5, and top 10 than the older version of Exomiser, thus making an argument for continuous reanalysis of the sequencing data of RGD patient cohorts. Furthermore, I analysed the percentage of variants that receive a score sufficiently high to warrant further analysis by bioinformaticians. This metric serves as a proxy to determine how much time an analyst has to spend per case. GPAs meaningfully increase the percentage of variants receiving a significant score compared to GAs, from 32.3% to 2.2% and from 25.4% to 11.2% respectively for Exomiser's hiPHIVE and VAAST+Phevor.

Chapter 4 introduces a new algorithm called GPET that is based on not only genotypic and phenotypic, but also tissue-specific expression data for variant ranking.

I used a dataset provided by SWISSProt consisting of disease-causing variants from ClinVar and non-disease-causing variants from the 1,000 Genomes Project as the basis for my analysis. After several data-cleansing steps, the final dataset used for my analysis consisted of 50,514 variants, of which 14,929 were labeled pathogenic and 35,585 benign. Based on the HPO annotations of each gene in the dataset in OMIM, I created an HPO profile for each variant in the dataset. Genes without OMIM-based HPO annotations were randomly allocated HPO annotations from genes with HPO annotations, so that every gene in the dataset was annotated with HPO terms. Based on an HPO-to-tissue mapping, I predicted which tissues were likely affected for each gene's phenotype, creating a matrix of so-called binary tissue labels (BTLs). In addition to the BTLs, I annotated each variant with three expression scores for each of 25 tissues, creating a total of 100 new expression features for each variant. I trained a random forest classifier on the expression features and each variant's Exomiser hiPHIVE scores. Subsequently, I calculated the AUC of GPET on the dataset for different scenarios, ranging from all variants in the testing and training dataset being perfectly scored by Exomiser's hiPHIVE's phenotype score, to no variants receiving an accurate phenotype score. GPET's AUC was consistently higher than Exomiser's hiPHIVE's AUC, reaching 0.95, as opposed to 0.91. The results of this chapter demonstrate two things: first, tissue-specific gene expression data is a powerful tool for RGD WGS variant prioritisation. Second, the HPO carries implicit information indicating which tissues are likely affected by a disease phenotype, information which in turn can be used in conjunction with tissue-specific expression data to improve RGD variant classification.

In Chapter 5, I applied the GPET framework developed in Chapter 4 to the RGD patient cases introduced in Chapter 3. Patient VCFs were first run through the same Exomiser pipeline used in Chapter 3, and variants were subsequently reranked using GPET. GPET uses each case's HPO terms to create a list of tissues relevant for the

patient. Gene expression in each tissue factors into each variant's ranking. In a first analysis, I showed that the expression of candidate disease genes is on average higher in tissues predicted-to-be-affected by the patient's HPO profile than in tissues not implicated by the phenotypic profile. Expression here is measured by three scores, which assess both the gene's relative expression compared to other genes in the same tissue, as well as each gene's expression in one tissue relative to all other tissues. By those metrics, mean gene expression of disease genes is approximately twice as high in tissues relevant for the phenotype as in other tissues. To my knowledge, this is the first time that effect was shown in real patient cases. Thereafter, I compared Exomiser's hiPHIVE's performance with GPET's performance. Exomiser's hiPHIVE outperforms GPET for cases where the candidate gene was annotated with HPO terms at the time of analysis. hiPHIVE ranks four candidate variants first, five in the top five and five in the top 20, compared to zero, zero and one for GPET respectively. In contrast to that, GPET outperforms hiPHIVE for candidate genes that were not annotated with HPO terms, ranking one novel candidate variant first, two in the top five, and three in the top 20, compared to zero, one, and two for hiPHIVE. Notably, GPET also improves the ranking of *HDLBP* as a candidate variant for Fine-Lubinsky syndrome.

In Chapter 6, I presented results to work towards the functional validation of *HDLBP* as a candidate gene for Fine-Lubinsky syndrome. FLS is a rare developmental disorder with a complex phenotype, including plagiocephaly, megalocornea, digital abnormalities, cleft palate, facial dysmorphism, structural brain abnormalities, and sometimes deafness. In this study, WGS was conducted for two of five affected relatives and a variant, c.1731+1G>A, leading to alternative splicing and consequently in-frame skipping of exon 14 of *HDLBP* was found. Exon 14 of *HDLBP* makes up $\approx$51% of the RNA-binding KH6 domain of vigilin, the protein encoded by *HDLBP*. Vigilin is a known RNA-binding protein and the loss of its RNA-binding function was hypothesised to stand in relation to the FLS phenotype. To elucidate the impact of the

HDLBP variant on vigilin's function and potential connection to the FLS phenotype, I cloned cDNA fragments of the *HDLBP* wildtype and our patient's *HDLBP* mutant into an eGFP-containing plasmid and transfected into HeLA and HEK293 cells. The cells were cultured and induced by doxycycline to produce wildtype and mutant vigilin-GFP fusion proteins. Western Blots showed that both constructs were stable. In addition to that, I inferred and analysed protein decay of both wildtype and mutant vigilin over 48h based on the GFP fluorescence intensity of wildtype and mutant vigilin-GFP fusion protein. No large significant difference in decay was detected, thus excluding potentially diminished stability of the mutant protein as a cause of the FLS phenotype. Subsequently, I used fluorescence microscopy of mutant and wildtype vigilin expressed in HeLa cells to assess intra-cellular protein localisation. No difference in intra-cellular localisation between the wildtype and mutant were detected. Hence, protein localisation did likely not cause the phenotype. Finally, I analysed the polyadenylated RNA-binding activity of the wildtype and mutant vigilin-GFP fusion protein, by hybridisation with an oligo(DT) probe. The amount of RNA bound by both mutant and wildtype was then measured as a function of the GFP tag's intensity in a fluorescence-based assay. The assay showed a considerable reduction of RNA-binding activity of the vigilin mutant-GFP fusion protein compared to the wildtype, showing that our patients' splice-site variant affects the RNA-binding activity of vigilin's KH6 domain. Vigilin has over 700 potential mRNA targets and has been observed to influence various stages of RNA metabolism, including mRNA transport, nuclear-cytoplasmic tRNA shuttling, translation, degradation, and formation of stress granules and processing bodies. To further investigate the relationship between the detected *HDLBP* splice-site variant and the FLS phenotype, assays to determine which RNA species are not bound anymore by the vigilin mutant are necessary and currently being conducted by Dr Pamela Kaisaki, a senior post-doctoral researcher in the Taylor Group.

7.2 Future work

The work presented in this thesis suggests several avenues for the improvement of the analysis of WGS data for rare genetic disease patients and the functional validation of *HDLBP* as a disease gene candidate for Fine-Lubinsky syndrome. Here, I split relevant future work into three distinct themes, namely the comparison of algorithmic frameworks (see Section 7.2.1), improvements for GPET (see Section 7.2.2), and the functional validation of *HDLBP* (see Section 7.2.3).

7.2.1 Comparison of algorithmic frameworks

The comparison of GAs, GPAs and GPET presented in Chapter 3 and Chapter 5 on real patient cases should be expanded to larger WGS patient cohorts. The 100,000 Genomes Project and the DDD study would be perfectly suited for this purpose. An expanded analysis would provide several benefits. First, a re-analysis of cases first sequenced several years ago would likely yield results for currently unsolved cases. Second, no independently published large-scale benchmark comparison of Exomiser's hiPHIVE, and its commercial integration into Congenica's platform, as well as Fabric Genomics' VAAST+Phevor algorithm exists. Existing comparisons were either conducted by the creators of Exomiser [58], or the creators of VAAST+Phevor [113] for a smaller number of cases. Cipriani *et al.* [119] published the to my knowledge largest evaluation of Exomiser on 134 WES patient cases. However, the analysis did not include VAAST+Phevor. A large-scale comparison of these established algorithm frameworks could provide best practices for the global RGD scientific community. Third, an analysis covering multiple different disease areas, including neurological, cardiovascular, skeletal, muscular and other conditions, would illustrate in which disease areas lower diagnostic yields are still achieved and thus provide guidance for research resource allocation.

7.2.2 GPET

GPET represents a first attempt at connecting GPAs and the HPO with tissue-specific expression data and many areas for improvement exist.

First, data on HPO terms could be leveraged better. Similar to hiPHIVE and Phevor, GPET only accounts for HPO terms that are present for the patient. Clinicians, however, also provide explicit information on which phenotypic features are not observed in a patient. The absence of those phenotypic features is meaningful, both for the link between HPO terms and genes used in hiPHIVE and Phevor, as well as for the implication of tissues in which disease genes are likely more highly expressed.

Second, the mapping of HPO terms to tissues can be significantly improved. In the current version of the algorithm, only 25 tissues are covered. The most recent version of GTEx (V8) however includes 54 tissue categories, most of which can be mapped to HPO terms [235].

Third, all features used for GPET can likely be constructed in a more sophisticated manner. BTLs are constructed based on whether or not an HPO term is present in the patient and are binary in nature. However, not all phenotypic features are equally strongly expressed in all patients. If information was collected on the severity of an individual patient's phenotype, that information could be used to create continuous tissue labels, instead of binary ones. Information on tissue-specific gene expression can likely be leveraged more efficiently. Instead of relying on the cross-gene and cross-tissue score pre-computed by Feiglin *et al* [132], a machine learning framework like GPET could use the raw expression data contained in GTEx as inputs. Furthermore, Min *et al* [236] and others have shown that expression patterns of certain genes are strongly correlated with each other. If gene A is up-regulated, so is gene B. Information on such co-expression networks could be used to shed light on biological pathways. An approach similar to GADO published by Deelen *et al.* [142] could be investigated.

GADO calculates gene co-regulation based on expression data across a range of tissues. The co-regulation predictions are used to determine which genes likely affect which phenotypic terms. GADO can be used to rank potentially disease-causing genes and was competitive with Exomiser in a benchmark study conducted by Deelen *et al.* [142]. In addition to improving how GPET uses expression data, improvements are possible for the use of genotypic data. GPET should not use hiPHIVE's pre-computed variant score as input, but should instead use the component features used to compute those scores (see Chapter 1 for a detailed description of those features).

7.2.3 Functional validation of *HDLBP*

While the functional experiments conducted to confirm *HDLBP* as a disease gene for FLS showed that the mutant version of vigilin found in our patients has significantly decreased RNA-binding activity, more work is necessary to fully characterise the disease pathway. Dr Kaisaki in the Taylor Group is conducting a range of RNA-binding assays for this purpose, including CLIP-RTPCR and CLIP-seq to highlight differences in which RNA species are bound by wildtype and mutant vigilin, and iCLIP to identify the RNA-recognition element.

At the same time, CRISPR-Cas9 will be used to introduce our patients' splice-site variant into the genomes of HEK293 and induced pluripotent stem cell lines. RNA-binding activity will be measured in the edited HEK293 cells to assess if differences in RNA-binding activity previously identified in the stably transfected HEK293 cells are confirmed. In addition to that, the edited iPS cells will be differentiated into lineages of endoderm, mesoderm and ectoderm cells, to detect potential differences between wildtype and mutant *HDLBP* cells. This could provide further insights into potential links between the developmental symptoms of FLS and defects caused by mutant vigilin in early development.

7.3 Concluding remarks

It is an exciting time for rare genetic disease diagnostics because our understanding of the underlying causes of disease is constantly improving. Advancements in four distinct themes elaborated on in this section are necessary to continue this trend, namely algorithm improvements (see Section 7.3.1), sequencing technology improvements (see Section 7.3.2), increased adoption of WGS in the clinic (see Section 7.3.3), and an increasing public awareness for rare genetic diseases (see Section 7.3.4).

7.3.1 Algorithm improvements

Significant improvements of analysis algorithms are necessary to increase the diagnostic yield of RGD population studies.

Better statistics: Work is underway to not only take advantage of information on which phenotypic features are present in patients, but also on phenotypic features that are explicitly not present in patients [237]. Capturing a patient's phenotype more accurately will improve diagnostic yield.

Furthermore, a small but increasing number of approaches exist to analyse intronic variation, e.g. Genomiser [120].

Improvements of existing databases: Significant improvements in diagnostic yield over time can be attributed to growing databases of various kinds.

Phenotype databases such as the HPO can be improved in a number of ways. First, the more genes are annotated with HPO terms, the more the diagnostic yield of algorithms will improve. Second, the HPO could move from a gene-centric to a variant-centric model, since type and position of the variant dramatically influence the phenotypic presentation of patients. Third, phenotype databases such as ORPHANET [114] are integrating both population frequencies of diseases annotated with HPO

terms, as well as disease-specific frequencies of each term. Finally, an increasing number of tools are being introduced to assist healthcare professionals with appropriately phenotyping patients. These tools include electronic medical record software such as PhenoTips [104], face recognition applications like FDNA's Face2Gene [238], and natural language processing algorithms like Bio-Lark [239].

An increasing number of patient matching databases is emerging that facilitate the diagnosis of rare and ultra rare conditions in geographically distant places. These databases include GeneMatcher [105], PhenomeCentral [103], Matchmaker Exchange [240], and DECIPHER [241]. Population-scale projects like the 100,000 Genomes Project have the potential to solve similar problems if databases are made accessible for external patient matching.

Furthermore, the increasing growth of databases documenting allele frequency, including ClinVar [69] and gnomAD [242], or additional data modalities such as gene expression, including GTEx [121], is improving the performance of VPA.

7.3.2 Sequencing technology improvements

In addition to databases, significant advancements are being made with sequencing technologies. The development of long-read sequencing by companies such as Oxford Nanopore and Pacific Biosciences will likely increase the diagnostic yield for RGD cases by enabling the analysis of repetitive regions, complex structural variation, and haplotype phasing [193]. Beyond long-read sequencing, the introduction of RNA sequencing is opening up new avenues for RGD diagnostics [243].

7.3.3 Adoption of WGS in the clinic

The growing evidence for the clinical utility of WGS is paving the way for large-scale adoption of WGS into clinical care. Institutions such as Rady Children's Hospital

have demonstrated the value of rapid WGS for seriously ill infants in neonatal and paediatric intensive care units [216], while the UK's National Healthcare Service and the UK BioBank announced the whole genome sequencing of five million patients until 2023 [244].

7.3.4 Increasing public awareness of genetics

Finally, the field of RGD is benefiting from increasing public awareness, partially driven by popular science contributions such as the New York Times column 'Diagnosis' by Dr. Lisa Sanders [245], in which the author describes undiagnosed rare disease cases, which are often genetic in nature, to crowd-source help from scientists, physicians, patients, allies and others around the globe to get patients answers. Dr. Sanders' column also provided the basis for a Netflix documentary series called 'Diagnosis'.

With continuous improvements in the field, one can be hopeful that - one day - all 400 million rare genetic disease patients globally will know the cause of their medical struggles and the diagnostic odyssey will be reduced from its current seven year average in the UK to mere hours. The results presented in this thesis are a further step in that direction by demonstrating the usefulness of phenotypic data and tissue-specific expression data for rare genetic disease diagnosis based on WGS data and provide evidence to support *HDLBP* as a disease gene for FLS.

References

[1] Warren Kaplan, Veronika J Wirtz, Aukje Mantel-Teeuwisse, Pieter Stolk, Beatrice Duthey, and Richard Laing. Priority Medicines for Europe and the World 2013 Update. *World Health Organization*, 2013.

[2] Flemming Ørnskov, Alastair Kent, and Nicole Boice. Rare Disease Impact Report: Insights from patients and the medical community. *Shire Report*, 2013.

[3] RARE Diseases: Facts and Statistics. https://globalgenes.org/rare-diseases-facts-statistics/, accessed 2018-08-20.

[4] ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing. https://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/, accessed 2018-08-21.

[5] Jessica X. Chong, Kati J. Buckingham, Shalini N. Jhangiani, Corinne Boehm, Nara Sobreira, Joshua D. Smith, Tanya M. Harrell, Margaret J. McMillin, Wojciech Wiszniewski, Tomasz Gambin, Zeynep H. Coban Akdemir, Kimberly Doheny, Alan F. Scott, Dimitri Avramopoulos, Aravinda Chakravarti, Julie Hoover-Fong, Debra Mathews, P. Dane Witmer, Hua Ling, Kurt Hetrick, Lee Watkins, Karynne E. Patterson, Frederic Reinier, Elizabeth Blue, Donna Muzny, Martin Kircher, Kaya Bilguvar, Francesc López-Giráldez, V. Reid Sutton, Holly K. Tabor, Suzanne M. Leal, Murat Gunel, Shrikant Mane, Richard A. Gibbs, Eric Boerwinkle, Ada Hamosh, Jay Shendure, James R. Lupski, Richard P. Lifton, David Valle, Deborah A. Nickerson, and Michael J. Bamshad. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *American Journal of Human Genetics*, 97(2):199–215, 2015.

[6] Johan T. den Dunnen, Raymond Dalgleish, Donna R. Maglott, Reece K. Hart, Marc S. Greenblatt, Jean Mcgowan-Jordan, Anne Francoise Roux, Timothy Smith, Stylianos E. Antonarakis, and Peter E.M. Taschner. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human Mutation*, 37(6):564–569, 2016.

[7] Frederick Sanger and Alan Coulson. A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase. *Journal of Molecular Biology*, 94:441–448, 1975.

[8] Frederick Sanger, Steve Nicklen, and Alan Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7, 1977.

[9] Stefan M Pulst. Genetic linkage analysis. *Archives of Neurology*, 56(6):667–672, 1999.

[10] Microarray-based Comparative Genomic Hybridization (aCGH) | Learn Science at Scitable. https://www.nature.com/scitable/topicpage/microarray-based-comparative-genomic-hybridization-acgh-45432/, accessed 2020-08-23.

[11] Lisenka E L M Vissers, Joris A Veltman, Ad Geurts Van Kessel, and Han G Brunner. Identification of disease genes by wholegenome CGH arrays. *Human Molecular Genetics*, 14(2):215–223, 2005.

[12] Illumina Inc. An introduction to Next-Generation Sequencing Technology. www.illumina.com/documents/products/illumina_sequencing_introduction.pdf, accessed 2020-10-14, 2017.

[13] Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F Twigg, Andrew O M Wilkie, Gil McVean, and Gerton Lunter. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 46(8):912–918, 2014.

[14] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[15] Gerton Lunter and Martin Goodson. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6):936–939, 2011.

[16] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–8, 2011.

[17] Marten Jäger, Kai Wang, Sebastian Bauer, Damian Smedley, Peter Krawitz, and Peter N. Robinson. Jannovar: A Java Library for Exome Annotation. *Human Mutation*, 35(5):548–555, 2014.

[18] K Wang, M Li, and H Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, 2010.

[19] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham R S Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(122), 2016.

[20] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, 2014.

[21] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–1081, 2009.

[22] Recommended Uniform Screening Panel | Official web site of the U.S. Health Resources & Services Administration. https://www.hrsa.gov/advisory-committees/heritable-disorders/rusp/index.html, accessed 2020-10-14.

[23] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H Shah, Atul J Butte, Michael D Howell, Claire Cui, Greg S Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):1–10, 2018.

[24] Saadet Mercimek-Mahmutoglu, Jaina Patel, Dawn Cordeiro, Stacy Hewson, David Callen, Elizabeth J Donner, Cecil D Hahn, Peter Kannu, Jeff Kobayashi, Berge A Minassian, Mahendranath Moharir, Komudi Siriwardena, Shelly K Weiss, Rosanna Weksberg, and O Carter Snead. Diagnostic yield of genetic testing in epileptic encephalopathy in childhood. *Epilepsia*, 56(5):707–716, 2015.

[25] G Bradley Schaefer and Richard E Lutz. Diagnostic yield in the clinical genetic evaluation of autism spectrum disorders. *Genetics in Medicine*, 8(9):549–556, 2006.

[26] UK BioBank. http://www.ukbiobank.ac.uk/, accessed 2018-08-21.

[27] UK Biobank. Regeneron announces major collaboration to exome sequence UK Biobank genetic data more quickly | UK Biobank. http://www.ukbiobank.ac.uk/2018/01/regeneron-announces-major-collaboration-to-exome-sequence-uk-biobank-genetic-data-more-quickly/, accessed 2018-08-21.

[28] Caroline F Wright, Jeremy F McRae, Stephen Clayton, Giuseppe Gallone, Stuart Aitken, Tomas W FitzGerald, Philip Jones, Elena Prigmore, Diana Rajan, Jenny Lord, Alejandro Sifrim, Rosemary Kelsell, Michael J Parker, Jeffrey C Barrett, Matthew E Hurles, David R FitzPatrick, and Helen V Firth. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genetics in Medicine*, 20(10):1216–1223, 2018.

[29] Leanne M Dibbens, Boukje De Vries, Simona Donatello, Sarah E Heron, Bree L Hodgson, Satyan Chintawar, Douglas E Crompton, James N Hughes, Susannah T Bellows, Karl Martin Klein, Petra M C Callenbach, Mark A Corbett, Alison E Gardner, Sara Kivity, Xenia Iona, Brigid M Regan, Claudia M Weller, Denis Crimmins, Terence J O'Brien, Rosa Guerrero-López, John C Mulley, Francois Dubeau, Laura Licchetta, Francesca Bisulli, Patrick Cossette, Paul Q Thomas, Jozef Gecz, Jose Serratosa, Oebele F Brouwer, Frederick Andermann, Eva Andermann, Arn M J M Van Den Maagdenberg, Massimo Pandolfo,

Samuel F Berkovic, and Ingrid E Scheffer. Mutations in DEPDC5 cause familial focal epilepsy with variable foci. *Nature Genetics*, 45(5):546–551, 2013.

[30] Michael J Bamshad, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755, 2011.

[31] Lucas D. Ward and Manolis Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology*, 30(11):1095–106, 2012.

[32] Vijay Kumar Pounraja, Gopal Jayakar, Matthew Jensen, Neil Kelkar, and Santhosh Girirajan. A machine-learning approach for accurate detection of copy number variants from exome sequencing. *Genome Research*, 29(7):1134–1143, 2019.

[33] Rick M. Tankard, Mark F. Bennett, Peter Degorski, Martin B. Delatycki, Paul J. Lockhart, and Melanie Bahlo. Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. *American Journal of Human Genetics*, 103(6):858–873, 2018.

[34] A B P Van Kuilenburg, M Tarailo-Graovac, P A Richmond, B I Drögemöller, M A Pouladi, R Leen, K Brand-Arzamendi, D Dobritzsch, E Dolzhenko, M A Eberle, B Hayward, M J Jones, F Karbassi, M S Kobor, J Koster, D Kumari, M Li, J MacIsaac, C McDonald, J Meijer, C Nguyen, I S Rajan-Babu, S W Scherer, B Sim, B Trost, L A Tseng, M Turkenburg, J J F A Van Vugt, J H Veldink, J S Walia, Y Wang, M Van Weeghel, G E B Wright, X Xu, R K C Yuen, J Zhang, C J Ross, W W Wasserman, M T Geraghty, S Santra, R J A Wanders, X Y Wen, H R Waterham, K Usdin, and C D M Van Karnebeek. Glutaminase deficiency caused by short tandem repeat expansion in GLS. *New England Journal of Medicine*, 380(15):1433–1441, 2019.

[35] Jenny C Taylor, Hilary C Martin, Stefano Lise, John Broxholme, Jean-baptiste Cazier, Andy Rimmer, Alexander Kanapin, Gerton Lunter, Simon Fiddy, Chris Allan, a Radu Aricescu, Moustafa Attar, Christian Babbs, Jennifer Becq, David Beeson, Celeste Bento, Patricia Bignell, Edward Blair, Veronica J Buckle, Katherine Bull, Ondrej Cais, Holger Cario, Helen Chapel, Richard R Copley, Richard Cornall, Jude Craft, Karin Dahan, Emma E Davenport, Calliope Dendrou, Olivier Devuyst, Aimée L Fenwick, Jonathan Flint, Lars Fugger, Rodney D Gilbert, Anne Goriely, Angie Green, Ingo H Greger, Russell Grocock, Anja V Gruszczyk, Robert Hastings, Edouard Hatton, Doug Higgs, Adrian Hill, Chris Holmes, Malcolm Howard, Linda Hughes, Peter Humburg, David Johnson, Fredrik Karpe, Zoya Kingsbury, Usha Kini, Julian C Knight, Jonathan Krohn, Sarah Lamble, Craig Langman, Lorne Lonie, Joshua Luck, Davis McCarthy, Simon J McGowan, Mary Frances McMullin, Kerry a Miller, Lisa Murray, Andrea H Németh, M Andrew Nesbit, David Nutt, Elizabeth Ormondroyd, Annette Bang Oturai, Alistair Pagnamenta, Smita Y Patel, Melanie Percy, Nayia Petousi, Paolo Piazza, Sian E Piret, Guadalupe Polanco-Echeverry, Niko Popitsch, Fiona Powrie, Chris Pugh, Lynn Quek, Peter A Robbins, Kathryn Robson, Alexandra Russo, Natasha Sahgal, Pauline A van Schouwenburg, Anna Schuh, Earl Silverman, Alison Simmons, Per Soelberg Sørensen, Elizabeth

Sweeney, John Taylor, Rajesh V Thakker, Ian Tomlinson, Amy Trebes, Stephen R F Twigg, Holm H Uhlig, Paresh Vyas, Tim Vyse, Steven a Wall, Hugh Watkins, Michael P Whyte, Lorna Witty, Ben Wright, Chris Yau, David Buck, Sean Humphray, Peter J Ratcliffe, John I Bell, Andrew O M Wilkie, David Bentley, Peter Donnelly, and Gilean McVean. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*, 47(7):717–726, 2015.

[36] Mark Stevenson, Alistair T Pagnamenta, Silvia Reichart, Charlotte Philpott, Kate E Lines, Caroline M Gorvin, Karl Lhotta, Jenny C Taylor, and Rajesh V Thakker. Whole genome sequence analysis identifies a PAX2 mutation to establish a correct diagnosis for a syndromic form of hyperuricemia. *American Journal of Medical Genetics, Part A*, 2020.

[37] André B P van Kuilenburg, Maja Tarailo-Graovac, Judith Meijer, Britt Drogemoller, Jerry Vockley, Dirk Maurer, Doreen Dobritzsch, Colin J Ross, Wyeth Wasserman, Rutger Meinsma, Lida Zoetekouw, and Clara D M van Karnebeek. Genome sequencing reveals a novel genetic mechanism underlying dihydropyrimidine dehydrogenase deficiency: A novel missense variant c.1700G>A and a large intragenic inversion in DPYD spanning intron 8 to intron 12. *Human Mutation*, 39(7):947–953, jul 2018.

[38] Anath C Lionel, Gregory Costain, Nasim Monfared, Susan Walker, Miriam S Reuter, S Mohsen Hosseini, Bhooma Thiruvahindrapuram, Daniele Merico, Rebekah Jobling, Thomas Nalpathamkalam, Giovanna Pellecchia, Wilson W L Sung, Zhuozhi Wang, Peter Bikangaga, Cyrus Boelman, Melissa T Carter, Dawn Cordeiro, Cheryl Cytrynbaum, Sharon D Dell, Priya Dhir, James J Dowling, Elise Heon, Stacy Hewson, Linda Hiraki, Michal Inbar-Feigenberg, Regan Klatt, Jonathan Kronick, Ronald M Laxer, Christoph Licht, Heather MacDonald, Saadet Mercimek-Andrews, Roberto Mendoza-Londono, Tino Piscione, Rayfel Schneider, Andreas Schulze, Earl Silverman, Komudi Siriwardena, O Carter Snead, Neal Sondheimer, Joanne Sutherland, Ajoy Vincent, Jonathan D Wasserman, Rosanna Weksberg, Cheryl Shuman, Chris Carew, Michael J Szego, Robin Z Hayeems, Raveen Basran, Dimitri J Stavropoulos, Peter N Ray, Sarah Bowdin, M Stephen Meyn, Ronald D Cohn, Stephen W Scherer, and Christian R Marshall. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genetics in Medicine*, 20(4):435–443, 2018.

[39] Andrew M Gross, Subramanian S Ajay, Vani Rajan, Carolyn Brown, Krista Bluske, Nicole J Burns, Aditi Chawla, Alison J Coffey, Alka Malhotra, Alicia Scocchia, Erin Thorpe, Natasa Dzidic, Karine Hovanes, Trilochan Sahoo, Egor Dolzhenko, Bryan Lajoie, Amirah Khouzam, Shimul Chowdhury, John Belmont, Eric Roller, Sergii Ivakhno, Stephen Tanner, Julia McEachern, Tina Hambuch, Michael Eberle, R Tanner Hagelstrom, David R Bentley, Denise L Perry, and Ryan J Taft. Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. *Genetics in Medicine*, 21(5):1121–1130, 2019.

[40] The 100,000 Genomes Project | Genomics England. https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/, accessed 2018-12-30.

[41] Suthesh Sivapalaratnam and NIHR Bioresource. The Rare Diseases Pilot for the 100,000 Genomes Project: Findings in Known and New Genes by Analysis of 3,549 Whole Genome Sequenced Samples from Patients and Relatives with Haematological, Haemostasis and Immune Disorders. *Blood*, 132(Supplement 1), 2018.

[42] Daniel F Gudbjartsson, Hannes Helgason, Sigurjon A Gudjonsson, Florian Zink, Asmundur Oddson, Arnaldur Gylfason, Soren Besenbacher, Gisli Magnusson, Bjarni V Halldorsson, Eirikur Hjartarson, Gunnar Th Sigurdsson, Simon N Stacey, Michael L Frigge, Hilma Holm, Saemundsdottir Jona, Hafdis Th Helgadottir, Hrefna Johannsdottir, Gunnlaugur Sigfusson, Gudmundur Thorgeirsson, Jon Th Sverrisson, Solveig Gretarsdottir, G Bragi Walters, Thorunn Rafnar, Bjarni Thjodleifsson, Einar S Bjornsson, Sigurdur Olafsson, Hildur Thorarinsdottir, Thora Steingrimsdottir, Thora S Gudmundsdottir, Asgeir Theodors, Jon G Jonasson, Asgeir Sigurdsson, Gyda Bjornsdottir, Jon J Jonsson, Olafur Thorarensen, Petur Ludvigsson, Hakon Gudbjartsson, Gudmundur I Eyjolfsson, Olof Sigurdardottir, Isleifur Olafsson, David O Arnar, Olafur Th Magnusson, Augustine Kong, Gisli Masson, Unnur Thorsteinsdottir, Agnar Helgason, Patrick Sulem, and Kari Stefansson. Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, 47(5):435–444, 2015.

[43] Human Longevity, Inc. https://www.humanlongevity.com/, accessed 2018-11-05.

[44] The Cost of Sequencing a Human Genome - National Human Genome Research Institute (NHGRI). https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/, accessed 2018-08-21.

[45] Erika Check Hayden. Genome researchers raise alarm over big data. http://www.nature.com/doifinder/10.1038/nature.2015.17912, accessed 2018-08-21, 2015.

[46] Ingenuity Analysis for Sequencing Data - Variant Analysis, IPA and iReport. http://www.ingenuity.com/data-analysis/sequencing-data, accessed 2016-10-30.

[47] Andrew R Carson, Erin N Smith, Hiroko Matsui, Sigrid K Brækkan, Kristen Jepsen, John-Bjarne Hansen, and Kelly A Frazer. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics*, 15:125, 2014.

[48] Leslie G Biesecker. Exome sequencing makes medical genomics a reality. *Nature Genetics*, 42(1):13–14, 2010.

[49] Adam Auton, Gonçalo R Abecasis, David M Altshuler, Richard M Durbin, Gonçalo R Abecasis, David R Bentley, Aravinda Chakravarti, Andrew G Clark, Peter Donnelly, Evan E Eichler, Paul Flicek, Stacey B Gabriel, Richard A Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Jun Wang, Yuqi Chang, Qiang

Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Eric S. Lander, David M. Altshuler, Stacey B. Gabriel, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Paul Flicek, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, David R. Bentley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Hans Lehrach, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie-Laure Yaspo, Elaine R. Mardis, Richard K. Wilson, Lucinda Fulton, Robert Fulton, Stephen T. Sherry, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Gil A. McVean, Richard M. Durbin, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Jeanette P. Schmidt, Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping Zhan, Adam Auton, Christopher L. Campbell, Yu Kong, Anthony Marcketta, Richard A. Gibbs, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Donna Muzny, Aniko Sabo, Zhuoyi Huang, Jun Wang, Lachlan J. M. Coin, Lin Fang, Xiaosen Guo, Xin Jin, Guoqing Li, Qibin Li, Yingrui Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly, Mark A. DePristo, Robert E. Handsaker, David M. Altshuler, Eric Banks, Gaurav Bhatia, Guillermo del Angel, Stacey B. Gabriel, Giulio Genovese, Namrata Gupta, Heng Li, Seva Kashin, Eric S. Lander, Steven A. McCarroll, James C. Nemesh, Ryan E. Poplin, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Andrew G. Clark, Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Jan O. Korbel, Tobias Rausch, Markus H. Fritz, Adrian M. Stütz, Paul Flicek, Kathryn Beal, Laura Clarke, Avik Datta, Javier Herrero, William M. McLaren, Graham R. S. Ritchie, Richard E. Smith, Daniel Zerbino, Xiangqun Zheng-Bradley, Pardis C. Sabeti, Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper, Edward V. Ball, Peter D. Stenson, David R. Bentley, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Sean Humphray, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E. Kenny, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Ralf Sudbrak, Vyacheslav S. Amstislavskiy, Ralf Herwig, Elaine R. Mardis, Li Ding, Daniel C. Koboldt, David Larson, Kai Ye, Simon Gravel, Anand Swaroop, Emily Chew, Tuuli Lappalainen, Yaniv Erlich, Melissa Gymrek, Thomas Frederick Willems, Jared T. Simpson, Mark D. Shriver, Jeffrey A. Rosenfeld, Carlos D. Bustamante, Stephen B. Montgomery, Francisco M. De La Vega, Jake K. Byrnes, Andrew W.

Carroll, Marianne K. DeGorter, Phil Lacroute, Brian K. Maples, Alicia R. Martin, Andres Moreno-Estrada, Suyash S. Shringarpure, Fouad Zakharia, Eran Halperin, Yael Baran, Charles Lee, Eliza Cerveira, Jaeho Hwang, Ankit Malhotra, Dariusz Plewczynski, Kamen Radew, Mallory Romanovitch, Chengsheng Zhang, Fiona C. L. Hyland, David W. Craig, Alexis Christoforides, Nils Homer, Tyler Izatt, Ahmet A. Kurdoglu, Shripad A. Sinari, Kevin Squire, Stephen T. Sherry, Chunlin Xiao, Jonathan Sebat, Danny Antaki, Madhusudan Gujral, Amina Noor, Kenny Ye, Esteban G. Burchard, Ryan D. Hernandez, Christopher R. Gignoux, David Haussler, Sol J. Katzman, W. James Kent, Bryan Howie, Andres Ruiz-Linares, Emmanouil T. Dermitzakis, Scott E. Devine, Gonçalo R. Abecasis, Hyun Min Kang, Jeffrey M. Kidd, Tom Blackwell, Sean Caron, Wei Chen, Sarah Emery, Lars Fritsche, Christian Fuchsberger, Goo Jun, Bingshan Li, Robert Lyons, Chris Scheller, Carlo Sidore, Shiya Song, Elzbieta Sliwerska, Daniel Taliun, Adrian Tan, Ryan Welch, Mary Kate Wing, Xiaowei Zhan, Philip Awadalla, Alan Hodgkinson, Yun Li, Xinghua Shi, Andrew Quitadamo, Gerton Lunter, Gil A. McVean, Jonathan L. Marchini, Simon Myers, Claire Churchhouse, Olivier Delaneau, Anjali Gupta-Hinch, Warren Kretzschmar, Zamin Iqbal, Iain Mathieson, Androniki Menelaou, Andy Rimmer, Dionysia K. Xifara, Taras K. Oleksyk, Yunxin Fu, Xiaoming Liu, Momiao Xiong, Lynn Jorde, David Witherspoon, Jinchuan Xing, Evan E. Eichler, Brian L. Browning, Sharon R. Browning, Fereydoun Hormozdiari, Peter H. Sudmant, Ekta Khurana, Richard M. Durbin, Matthew E. Hurles, Chris Tyler-Smith, Cornelis A. Albers, Qasim Ayub, Senduran Balasubramaniam, Yuan Chen, Vincenza Colonna, Petr Danecek, Luke Jostins, Thomas M. Keane, Shane McCarthy, Klaudia Walter, Yali Xue, Mark B. Gerstein, Alexej Abyzov, Suganthi Balasubramanian, Jieming Chen, Declan Clarke, Yao Fu, Arif O. Harmanci, Mike Jin, Donghoon Lee, Jeremy Liu, Xinmeng Jasmine Mu, Jing Zhang, Yan Zhang, Yingrui Li, Ruibang Luo, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Wan-Ping Lee, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Steven A. McCarroll, Robert E. Handsaker, David M. Altshuler, Eric Banks, Guillermo del Angel, Giulio Genovese, Chris Hartl, Heng Li, Seva Kashin, James C. Nemesh, Khalid Shakir, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Jeremiah Degenhardt, Jan O. Korbel, Markus H. Fritz, Sascha Meiers, Benjamin Raeder, Tobias Rausch, Adrian M. Stütz, Paul Flicek, Francesco Paolo Casale, Laura Clarke, Richard E. Smith, Oliver Stegle, Xiangqun Zheng-Bradley, David R. Bentley, Bret Barnes, R. Keira Cheetham, Michael Eberle, Sean Humphray, Scott Kahn, Lisa Murray, Richard Shaw, Eric-Wubbo Lameijer, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Li Ding, Ira Hall, Kai Ye, Phil Lacroute, Charles Lee, Eliza Cerveira, Ankit Malhotra, Jaeho Hwang, Dariusz Plewczynski, Kamen Radew, Mallory Romanovitch, Chengsheng Zhang, David W. Craig, Nils Homer, Deanna Church, Chunlin Xiao, Jonathan Sebat, Danny Antaki, Vineet Bafna, Jacob Michaelson, Kenny Ye, Scott E. Devine, Eugene J. Gardner, Gonçalo R. Abecasis, Jeffrey M. Kidd, Ryan E. Mills, Gargi Dayama, Sarah Emery, Goo Jun, Xinghua Shi, Andrew Quitadamo, Gerton Lunter, Gil A. McVean, Ken Chen, Xian Fan, Zechen Chong, Tenghui Chen, David Witherspoon, Jinchuan Xing, Evan E. Eichler, Mark J. Chaisson, Fereydoun Hormozdiari, John Huddleston, Maika Malig, Bradley J. Nelson, Peter H. Sudmant, Nicholas F. Parrish, Ekta Khurana, Matthew E. Hurles, Ben Blackburne, Sarah J. Lindsay, Zemin Ning, Klaudia Walter, Yujun Zhang, Mark B. Gerstein, Alexej Abyzov, Jieming Chen,

Declan Clarke, Hugo Lam, Xinmeng Jasmine Mu, Cristina Sisu, Jing Zhang, Yan Zhang, Richard A. Gibbs, Fuli Yu, Matthew Bainbridge, Danny Challis, Uday S. Evani, Christie Kovar, James Lu, Donna Muzny, Uma Nagaswamy, Jeffrey G. Reid, Aniko Sabo, Jin Yu, Xiaosen Guo, Wangshen Li, Yingrui Li, Renhua Wu, Gabor T. Marth, Erik P. Garrison, Wen Fung Leong, Alistair N. Ward, Guillermo del Angel, Mark A. DePristo, Stacey B. Gabriel, Namrata Gupta, Chris Hartl, Ryan E. Poplin, Andrew G. Clark, Juan L. Rodriguez-Flores, Paul Flicek, Laura Clarke, Richard E. Smith, Xiangqun Zheng-Bradley, Daniel G. MacArthur, Elaine R. Mardis, Robert Fulton, Daniel C. Koboldt, Simon Gravel, Carlos D. Bustamante, David W. Craig, Alexis Christoforides, Nils Homer, Tyler Izatt, Stephen T. Sherry, Chunlin Xiao, Emmanouil T. Dermitzakis, Gonçalo R. Abecasis, Hyun Min Kang, Gil A. McVean, Mark B. Gerstein, Suganthi Balasubramanian, Lukas Habegger, Haiyuan Yu, Paul Flicek, Laura Clarke, Fiona Cunningham, Ian Dunham, Daniel Zerbino, Xiangqun Zheng-Bradley, Kasper Lage, Jakob Berg Jespersen, Heiko Horn, Stephen B. Montgomery, Marianne K. DeGorter, Ekta Khurana, Chris Tyler-Smith, Yuan Chen, Vincenza Colonna, Yali Xue, Mark B. Gerstein, Suganthi Balasubramanian, Yao Fu, Donghoon Kim, Adam Auton, Anthony Marcketta, Rob Desalle, Apurva Narechania, Melissa A. Wilson Sayres, Erik P. Garrison, Robert E. Handsaker, Seva Kashin, Steven A. McCarroll, Juan L. Rodriguez-Flores, Paul Flicek, Laura Clarke, Xiangqun Zheng-Bradley, Yaniv Erlich, Melissa Gymrek, Thomas Frederick Willems, Carlos D. Bustamante, Fernando L. Mendez, G. David Poznik, Peter A. Underhill, Charles Lee, Eliza Cerveira, Ankit Malhotra, Mallory Romanovitch, Chengsheng Zhang, Gonçalo R. Abecasis, Lachlan Coin, Haojing Shao, David Mittelman, Chris Tyler-Smith, Qasim Ayub, Ruby Banerjee, Maria Cerezo, Yuan Chen, Thomas W. Fitzgerald, Sandra Louzada, Andrea Massaia, Shane McCarthy, Graham R. Ritchie, Yali Xue, Fengtang Yang, Richard A. Gibbs, Christie Kovar, Divya Kalra, Walker Hale, Donna Muzny, Jeffrey G. Reid, Jun Wang, Xu Dan, Xiaosen Guo, Guoqing Li, Yingrui Li, Chen Ye, Xiaole Zheng, David M. Altshuler, Paul Flicek, Laura Clarke, Xiangqun Zheng-Bradley, David R. Bentley, Anthony Cox, Sean Humphray, Scott Kahn, Ralf Sudbrak, Marcus W. Albrecht, Matthias Lienhard, David Larson, David W. Craig, Tyler Izatt, Ahmet A. Kurdoglu, Stephen T. Sherry, Chunlin Xiao, David Haussler, Gonçalo R. Abecasis, Gil A. McVean, Richard M. Durbin, Senduran Balasubramaniam, Thomas M. Keane, Shane McCarthy, James Stalker, Aravinda Chakravarti, Bartha M. Knoppers, Gonçalo R. Abecasis, Kathleen C. Barnes, Christine Beiswanger, Esteban G. Burchard, Carlos D. Bustamante, Hongyu Cai, Hongzhi Cao, Richard M. Durbin, Norman P. Gerry, Neda Gharani, Richard A. Gibbs, Christopher R. Gignoux, Simon Gravel, Brenna Henn, Danielle Jones, Lynn Jorde, Jane S. Kaye, Alon Keinan, Alastair Kent, Angeliki Kerasidou, Yingrui Li, Rasika Mathias, Gil A. McVean, Andres Moreno-Estrada, Pilar N. Ossorio, Michael Parker, Alissa M. Resch, Charles N. Rotimi, Charmaine D. Royal, Karla Sandoval, Yeyang Su, Ralf Sudbrak, Zhongming Tian, Sarah Tishkoff, Lorraine H. Toji, Chris Tyler-Smith, Marc Via, Yuhong Wang, Huanming Yang, Ling Yang, Jiayong Zhu, Walter Bodmer, Gabriel Bedoya, Andres Ruiz-Linares, Zhiming Cai, Yang Gao, Jiayou Chu, Leena Peltonen, Andres Garcia-Montero, Alberto Orfao, Julie Dutil, Juan C. Martinez-Cruzado, Taras K. Oleksyk, Kathleen C. Barnes, Rasika A. Mathias, Anselm Hennis, Harold Watson, Colin McKenzie, Firdausi Qadri, Regina LaRocque, Pardis C. Sabeti, Jiayong Zhu, Xiaoyan Deng, Pardis C. Sabeti, Danny Asogun, Onikepe

Folarin, Christian Happi, Omonwunmi Omoniwa, Matt Stremlau, Ridhi Tariyal, Muminatou Jallow, Fatoumatta Sisay Joof, Tumani Corrah, Kirk Rockett, Dominic Kwiatkowski, Jaspal Kooner, Trâ'n Tnh Hiê'n, Sarah J. Dunstan, Nguyen Thuy Hang, Richard Fonnie, Robert Garry, Lansana Kanneh, Lina Moses, Pardis C. Sabeti, John Schieffelin, Donald S. Grant, Carla Gallo, Giovanni Poletti, Danish Saleheen, Asif Rasheed, Lisa D. Brooks, Adam L. Felsenfeld, Jean E. McEwen, Yekaterina Vaydylevich, Eric D. Green, Audrey Duncanson, Michael Dunn, Jeffery A. Schloss, Jun Wang, Huanming Yang, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, Jonathan L. Marchini, Shane McCarthy, Gil A. McVean, and Gonçalo R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[50] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, Taru Tukiainen, Daniel P. Birnbaum, Jack A. Kosmicki, Laramie E. Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N. Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I. Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M. Peloso, Ryan Poplin, Manuel A. Rivas, Valentin Ruano-Rubio, Samuel A. Rose, Douglas M. Ruderfer, Khalid Shakir, Peter D. Stenson, Christine Stevens, Brett P. Thomas, Grace Tiao, Maria T. Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M. Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C. Florez, Stacey B. Gabriel, Gad Getz, Stephen J. Glatt, Christina M. Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M. Neale, Aarno Palotie, Shaun M. Purcell, Danish Saleheen, Jeremiah M. Scharf, Pamela Sklar, Patrick F. Sullivan, Jaakko Tuomilehto, Ming T. Tsuang, Hugh C. Watkins, James G. Wilson, Mark J. Daly, and Daniel G. MacArthur. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016.

[51] NHLBI Grand Opportunity Exome Sequencing Project (ESP). https://esp.gs.washington.edu/drupal/, accessed 2019-01-21.

[52] dbSNP database. https://www.ncbi.nlm.nih.gov/snp, accessed 2019-04-11.

[53] Regis A James, Ian M Campbell, Edward S Chen, Philip M Boone, Mitchell A Rao, Matthew N Bainbridge, James R Lupski, Yaping Yang, Christine M Eng, Jennifer E Posey, and Chad A Shaw. A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. *Genome Medicine*, 8(1):13, 2016.

[54] Anthony Fejes, Björn Stade, Stephan Ritter, Anna Lewis, Edward Kiruluta, and Martin Reese. Fabric Genomics' Opal Clinical Variant Interpretation Platform Enables Rapid Whole Genome Analysis Turnaround in Under an Hour. In *ACMG 2017 Poster #291*, 2017.

[55] Hao Hu, Chad D. Huff, Barry Moore, Steven Flygare, Martin G. Reese, and Mark Yandell. VAAST 2.0: Improved variant classification and disease-gene

identification using a conservation-controlled amino acid substitution matrix. *Genetic Epidemiology*, 37(6):622–634, 2013.

[56] I A Adzhubei, S Schmidt, L Peshkin, V E Ramensky, A Gerasimova, P Bork, A S Kondrashov, and S R Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, 2010.

[57] Jana Marie Schwarz, Christian Rödelsperger, Markus Schuelke, and Dominik Seelow. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, 7(8):575–576, 2010.

[58] Damian Smedley and Peter N. Robinson. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Medicine*, 7(1):81, 2015.

[59] Fang Shi, Yao Yao, Yannan Bin, Chun Hou Zheng, and Junfeng Xia. Computational identification of deleterious synonymous variants in human genomes using a feature-based approach. *BMC Medical Genomics*, 12(Suppl 1), 2019.

[60] Fran Supek, Belén Miñana, Juan Valcárcel, Toni Gabaldón, and Ben Lehner. Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, 156(6):1324–1335, 2014.

[61] Ryan C. Hunt, Vijaya L. Simhadri, Matthew Iandoli, Zuben E. Sauna, and Chava Kimchi-Sarfaty. Exposing synonymous mutations. *Trends in Genetics*, 30(7):308–321, 2014.

[62] Paige S. Spencer, Efraín Siller, John F. Anderson, and José M. Barral. Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *Journal of Molecular Biology*, 422(3):328–335, 2012.

[63] Luca Cartegni, Shern L. Chew, and Adrian R. Krainer. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. *Nature Reviews Genetics*, 3(4):285–298, 2002.

[64] Pengbo Wen, Peng Xiao, and Junfeng Xia. dbDSM: a manually curated database for deleterious synonymous mutations. *Bioinformatics*, 32(12):1914–1916, jun 2016.

[65] Peter N Robinson, Sebastian Köhler, Anika Oellrich, Sanger Mouse Genetics, Kai Wang, Christopher J Mungall, Suzanna E. Lewis, Nicole Washington, Sebastian Bauer, Dominik Seelow, Peter Krawitz, Christian Gilissen, Melissa Haendel, and Damian Smedley. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, 24(2):340–348, 2014.

[66] Sebastian Köhler, Marcel H Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N Robinson. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *American Journal of Human Genetics*, 85(4):457–464, 2009.

[67] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L Rehm. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–423, 2015.

[68] Arindam Bhattacharjee, Tanya Sokolsky, Stacia K Wyman, Martin G Reese, Erik Puffenberger, Kevin Strauss, Holmes Morton, Richard B Parad, and Edwin W Naylor. Development of DNA Confirmatory and High-Risk Diagnostic Testing for Newborns Using Targeted Next-Generation DNA Sequencing. *Genetics in Medicine*, 17(5):337–47, 2014.

[69] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1):980–985, 2014.

[70] HGMD home page. http://www.hgmd.cf.ac.uk/ac/index.php, accessed 2018-08-21.

[71] SwissVar - Portal to Swiss-Prot diseases and variants. https://swissvar.expasy.org/, accessed 2018-08-21.

[72] gnomAD browser. http://gnomad.broadinstitute.org/, accessed 2018-08-23.

[73] ExAC Browser. http://exac.broadinstitute.org/, 2018-08-21.

[74] Laura M. Amendola, Michael O. Dorschner, Peggy D. Robertson, Joseph S. Salama, Ragan Hart, Brian H. Shirts, Mitzi L. Murray, Mari J. Tokita, Carlos J. Gallego, Daniel Seung Kim, James T. Bennett, David R. Crosslin, Jane Ranchalis, Kelly L. Jones, Elisabeth A. Rosenthal, Ella R. Jarvik, Andy Itsara, Emily H. Turner, Daniel S. Herman, Jennifer Schleit, Amber Burt, Seema M. Jamal, Jenica L. Abrudan, Andrew D. Johnson, Laura K. Conlin, Matthew C. Dulik, Avni Santani, Danielle R. Metterville, Melissa Kelly, Ann Katherine M. Foreman, Kristy Lee, Kent D. Taylor, Xiuqing Guo, Kristy Crooks, Lesli A. Kiedrowski, Leslie J. Raffel, Ora Gordon, Kalotina Machini, Robert J. Desnick, Leslie G. Biesecker, Steven A. Lubitz, Surabhi Mulchandani, Greg M. Cooper, Steven Joffe, C. Sue Richards, Yaoping Yang, Jerome I. Rotter, Stephen S. Rich, Christopher J. O'Donnell, Jonathan S. Berg, Nancy B. Spinner, James P. Evans, Stephanie M. Fullerton, Kathleen A. Leppig, Robin L. Bennett, Thomas Bird, Virginia P. Sybert, William M. Grady, Holly K. Tabor, Jerry H. Kim, Michael J. Bamshad, Benjamin Wilfond, Arno G. Motulsky, C. Ronald Scott, Colin C. Pritchard, Tom D. Walsh, Wylie Burke, Wendy H. Raskind, Peter Byers, Fuki M. Hisama, Heidi Rehm, Debbie A. Nickerson, and Gail P. Jarvik. Actionable exomic incidental findings in 6503 participants: Challenges of variant classification. *Genome Research*, 25(3):305–315, 2015.

[75] Laura M. Amendola, Gail P. Jarvik, Michael C. Leo, Heather M. McLaughlin, Yassmine Akkari, Michelle D. Amaral, Jonathan S. Berg, Sawona Biswas, Kevin M. Bowling, Laura K. Conlin, Greg M. Cooper, Michael O. Dorschner, Matthew C. Dulik, Arezou A. Ghazani, Rajarshi Ghosh, Robert C. Green, Ragan Hart, Carrie Horton, Jennifer J. Johnston, Matthew S. Lebo, Aleksandar

Milosavljevic, Jeffrey Ou, Christine M. Pak, Ronak Y. Patel, Sumit Punj, Carolyn Sue Richards, Joseph Salama, Natasha T. Strande, Yaping Yang, Sharon E. Plon, Leslie G. Biesecker, and Heidi L. Rehm. Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *American Journal of Human Genetics*, 99(1):247, 2016.

[76] Shan Yang, Stephen E. Lincoln, Yuya Kobayashi, Keith Nykamp, Robert L. Nussbaum, and Scott Topper. Sources of discordance among germ-line variant classifications in ClinVar. *Genetics in Medicine*, 19(10):1118–1126, 2017.

[77] Naisha Shah, Ying Chen Claire Hou, Hung Chun Yu, Rachana Sainger, C. Thomas Caskey, J. Craig Venter, and Amalio Telenti. Identification of Misclassified ClinVar Variants via Disease Population Prevalence. *American Journal of Human Genetics*, 102(4):609–619, 2018.

[78] Rodrigo A. Toledo. Inflated pathogenic variant profiles in the ClinVar database. *Nature Reviews Endocrinology*, 14:387–389, 2018.

[79] Heidi L Rehm, Jonathan S Berg, Lisa D Brooks, Carlos D Bustamante, James P Evans, Melissa J Landrum, David H Ledbetter, Donna R Maglott, Christa Lese Martin, Robert L Nussbaum, Sharon E Plon, Erin M Ramos, Stephen T Sherry, and Michael S Watson. ClinGen — The Clinical Genome Resource. *New England Journal of Medicine*, 372(23):2235–2242, 2015.

[80] Review Guidelines ClinVar NCBI. ncbi.nlm.nih.gov/clinvar/docs/review_guidelines/, accessed 2018-08-23.

[81] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard A Gibbs, W James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15:1034–1050, 2005.

[82] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics*, 76(1), 2014.

[83] P C Ng, J G Henikoff, and S Henikoff. PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics*, 17(3), 2001.

[84] Helen M. Berman, Gerard J. Kleywegt, Haruki Nakamura, and John L. Markley. The future of the protein data bank. *Biopolymers*, 99(3):218–222, mar 2013.

[85] Karen Eilbeck, Aaron Quinlan, and Mark Yandell. Settling the score: variant prioritization and Mendelian disease. *Nature Reviews Genetics*, 18(10):599–612, 2017.

[86] Jana Marie Schwarz, David N. Cooper, Markus Schuelke, and Dominik Seelow. Mutationtaster2: Mutation prediction for the deep-sequencing age. *Nature Methods*, 11(4):361–362, 2014.

[87] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

[88] Nilah M. Ioannidis, Joseph H. Rothstein, Vikas Pejaver, Sumit Middha, Shannon K. McDonnell, Saurabh Baheti, Anthony Musolf, Qing Li, Emily Holzinger, Danielle Karyadi, Lisa A. Cannon-Albright, Craig C. Teerlink, Janet L. Stanford, William B. Isaacs, Jianfeng Xu, Kathleen A. Cooney, Ethan M. Lange, Johanna Schleutker, John D. Carpten, Isaac J. Powell, Olivier Cussenot, Geraldine Cancel-Tassin, Graham G. Giles, Robert J. MacInnis, Christiane Maier, Chih Lin Hsieh, Fredrik Wiklund, William J. Catalona, William D. Foulkes, Diptasri Mandal, Rosalind A. Eeles, Zsofia Kote-Jarai, Carlos D. Bustamante, Daniel J. Schaid, Trevor Hastie, Elaine A. Ostrander, Joan E. Bailey-Wilson, Predrag Radivojac, Stephen N. Thibodeau, Alice S. Whittemore, and Weiva Sieh. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American Journal of Human Genetics*, 99(4):877–885, 2016.

[89] Damian Smedley, Julius O B Jacobsen, Marten Jäger, Sebastian Köhler, Manuel Holtgrewe, Max Schubach, Enrico Siragusa, Tomasz Zemojtel, Orion J Buske, Nicole L Washington, William P Bone, Melissa a Haendel, and Peter N Robinson. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols*, 10(12):2004–2015, 2015.

[90] The Exomiser: A Tool to Annotate and Prioritize Exome Variants.

[91] Exome Variant Server. http://evs.gs.washington.edu/EVS/, accessed 2018-08-24.

[92] Precision Medicine Activities | NHLBI, NIH - TOPMed database.

[93] UK10K - Home page.

[94] Mark Yandell, Chad Huff, Hao Hu, Marc Singleton, Barry Moore, Jinchuan Xing, Lynn B. Jorde, and Martin G. Reese. A probabilistic disease-gene finder for personal genomes. *Genome Research*, 21(9):1529–1542, 2011.

[95] Dg MacArthur, Suganthi Balasubramanian, and Adam Frankish. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science*, 335(6070):1–14, 2012.

[96] Peter N. Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *American Journal of Human Genetics*, 83(5):610–615, 2008.

[97] Peter N. Robinson and Stefan Mundlos. The Human Phenotype Ontology. *Clinical Genetics*, 77(6):525–534, 2010.

[98] Tudor Groza, Sebastian Köhler, Dawid Moldenhauer, Nicole Vasilevsky, Gareth Baynam, Tomasz Zemojtel, Lynn Marie Schriml, Warren Alden Kibbe, Paul N. Schofield, Tim Beck, Drashtti Vasant, Anthony J. Brookes, Andreas Zankl, Nicole L. Washington, Christopher J. Mungall, Suzanna E. Lewis, Melissa A. Haendel, Helen Parkinson, and Peter N. Robinson. The Human Phenotype

Ontology: Semantic Unification of Common and Rare Disease. *American Journal of Human Genetics*, 97(1):111–124, 2015.

[99] Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, Nomi L Harris, Nicolas Matentzoglu, Julie A McMurry, David Osumi-Sutherland, Valentina Cipriani, James P Balhoff, Tom Conlin, Hannah Blau, Gareth Baynam, Richard Palmer, Dylan Gratian, Hugh Dawkins, Michael Segal, Anna C Jansen, Ahmed Muaz, Willie H Chang, Jenna Bergerson, Stanley J F Laulederkind, Zafer Yüksel, Sergi Beltran, Alexandra F Freeman, Panagiotis I Sergouniotis, Daniel Durkin, Andrea L Storm, Marc Hanauer, Michael Brudno, Susan M Bello, Murat Sincan, Kayli Rageth, Matthew T Wheeler, Renske Oegema, Halima Lourghi, Maria G Della Rocca, Rachel Thompson, Francisco Castellanos, James Priest, Charlotte Cunningham-Rundles, Ayushi Hegde, Ruth C Lovering, Catherine Hajek, Annie Olry, Luigi Notarangelo, Morgan Similuk, Xingmin A Zhang, David Gómez-Andrés, Hanns Lochmüller, Hélène Dollfus, Sergio Rosenzweig, Shruti Marwaha, Ana Rath, Kathleen Sullivan, Cynthia Smith, Joshua D Milner, Dorothée Leroux, Cornelius F Boerkoel, Amy Klion, Melody C Carter, Tudor Groza, Damian Smedley, Melissa A Haendel, Chris Mungall, and Peter N Robinson. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, pages 1–10, 2018.

[100] Peter N. Robinson. Human Phenotype Ontology Website. https://hpo.jax.org/app/, accessed 2018-08-02.

[101] Quentin Ferry, Julia Steinberg, Caleb Webber, David R FitzPatrick, Chris P Ponting, Andrew Zisserman, and Christoffer Nellåker. Diagnostically relevant facial gestalt information from ordinary photos. *eLife*, 3:e02020, 2014.

[102] Face2Gene Homepage. https://www.face2gene.com/, accessed 2018-08-24.

[103] Orion J. Buske, Marta Girdea, Sergiu Dumitriu, Bailey Gallinger, Taila Hartley, Heather Trang, Andriy Misyura, Tal Friedman, Chandree Beaulieu, William P. Bone, Amanda E. Links, Nicole L. Washington, Melissa A. Haendel, Peter N. Robinson, Cornelius F. Boerkoel, David Adams, William A. Gahl, Kym M. Boycott, and Michael Brudno. PhenomeCentral: A Portal for Phenotypic and Genotypic Matchmaking of Patients with Rare Genetic Diseases. *Human Mutation*, 36(10):931–940, 2015.

[104] Marta Girdea, Sergiu Dumitriu, Marc Fiume, Sarah Bowdin, Kym M. Boycott, Sébastien Chénier, David Chitayat, Hanna Faghfoury, M. Stephen Meyn, Peter N. Ray, Joyce So, Dimitri J. Stavropoulos, and Michael Brudno. PhenoTips: Patient phenotyping software for clinical and research use. *Human Mutation*, 34(8):1057–1065, 2013.

[105] Nara Sobreira, François Schiettecatte, David Valle, and Ada Hamosh. GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene Nara. *Human Mutation*, 36(10):928–930, 2015.

[106] HPO annotations for "congenital bilateral hip dislocation". http://compbio.charite.de/hpoweb/showterm?id=HP:0008780, accessed 2016-11-02.

[107] Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/, accessed 2018-08-12.

[108] WHO | International Classification of Diseases, 11th Revision (ICD-11). http://www.who.int/classifications/icd/en/, accessed 2018-08-24.

[109] SNOMED International. https://www.snomed.org/snomed-ct, accessed 2018-08-24.

[110] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the Interactome for Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics*, 82:949–958, 2008.

[111] Cynthia L Smith, Carrollann W Goldsmith, and Janan T Eppig. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(R7), 2004.

[112] Divya Sardana, Suresh Vasa, Nishanth Vepachedu, Jing Chen, Ranga Chandra Gudivada, Bruce J. Aronow, and Anil G. Jegga. PhenoHM: Human-mouse comparative phenome-genome server. *Nucleic Acids Research*, 38:165–174, 2010.

[113] Marc V. Singleton, Stephen L. Guthery, Karl V. Voelkerding, Karin Chen, Brett Kennedy, Rebecca L. Margraf, Jacob Durtschi, Karen Eilbeck, Martin G. Reese, Lynn B. Jorde, Chad D. Huff, and Mark Yandell. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *American Journal of Human Genetics*, 94(4):599–610, 2014.

[114] Orphanet. https://www.orpha.net/, accessed 2018-08-25.

[115] Damian Smedley, Sebastian Köhler, Johanna Christina Czeschik, Joanna Amberger, Carol Bocchini, Ada Hamosh, Julian Veldboer, Tomasz Zemojtel, and Peter N. Robinson. Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*, 30(22):3215–3222, 2014.

[116] Damian Smedley, Anika Oellrich, Sebastian Köhler, Barbara Ruef, Monte Westerfield, Peter Robinson, Suzanna Lewis, and Christopher Mungall. PhenoDigm: Analyzing curated annotations to associate animal models with human diseases. *Database*, 2013:1–11, 2013.

[117] Nicole L Washington, Melissa A Haendel, Christopher J Mungall, Michael Ashburner, Monte Westerfield, and Suzanna E Lewis. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biology*, 7(11), 2009.

[118] William P Bone, Nicole L Washington, Orion J Buske, David R Adams, Joie Davis, David Draper, Elise D. Flynn, Marta Girdea, Rena Godfrey, Gretchen Golas, Catherine Groden, Julius Jacobsen, Sebastian Köhler, Elizabeth M. J. Lee, Amanda E. Links, Thomas C. Markello, Christopher J. Mungall, Michele Nehrebecky, Peter N. Robinson, Murat Sincan, Ariane G. Soldatos, Cynthia J. Tifft, Camilo Toro, Heather Trang, Elise Valkanas, Nicole Vasilevsky, Colleen

Wahl, Lynne A. Wolfe, Cornelius F. Boerkoel, Michael Brudno, Melissa A. Haendel, William A. Gahl, and Damian Smedley. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genetics in Medicine*, 18(6):608–617, 2016.

[119] Valentina Cipriani, Nikolas Pontikos, Gavin Arno, Panagiotis I. Sergouniotis, Eva Lenassi, Penpitcha Thawong, Daniel Danis, Michel Michaelides, Andrew R. Webster, Anthony T. Moore, Peter N. Robinson, Julius O.B. Jacobsen, and Damian Smedley. An Improved Phenotype-Driven Tool for Rare Mendelian Variant Prioritization: Benchmarking Exomiser on Real Patient Whole-Exome Data. *Genes*, 11(4):460, 2020.

[120] Damian Smedley, Max Schubach, Julius O B Jacobsen, Sebastian Ko, Tomasz Zemojtel, Malte Spielmann, Harry Hochheiser, Nicole L Washington, Julie A Mcmurry, Melissa A Haendel, Christopher J Mungall, Suzanna E Lewis, and Tudor Groza. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *The American Journal of Human Genetics*, 99:595–606, 2016.

[121] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J. Cox, Dan L. Nicolae, Eric R. Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T. Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M. Anderson, Elizabeth L. Wilder, Leslie K. Derr, Eric D. Green, Jeffery P. Struewing, Gary Temple, Simona Volpi, Joy T. Boyer, Elizabeth J. Thomson, Mark S. Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R. Insel, Susan E. Koester, A Roger Little, Patrick K. Bender, Thomas Lehner, Yin Yao, Carolyn C. Compton, Jimmie B. Vaught, Sherilyn Sawyer, Nicole C. Lockhart, Joanne Demchok, and Helen F. Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

[122] GTEx Portal. https://gtexportal.org/home/tissueSummaryPage, accessed 2018-08-26.

[123] Downloadable data - The Human Protein Atlas. https://www.proteinatlas.org/about/download, 2018-08-26.

[124] A. I Su, T Wiltshire, S Batalov, H Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences*, 101(16):6062–6067, 2004.

[125] Robert Petryszak, Tony Burdett, Benedetto Fiorelli, Nuno A. Fonseca, Mar Gonzalez-Porta, Emma Hastings, Wolfgang Huber, Simon Jupp, Maria Keays, Nataliya Kryvych, Julie McMurry, John C. Marioni, James Malone, Karine Megy, Gabriella Rustici, Amy Y. Tang, Jan Taubert, Eleanor Williams, Oliver Mannion, Helen E. Parkinson, and Alvis Brazma. Expression Atlas update - A database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Research*, 42:926–932, 2014.

[126] Markus Krupp, Jens U. Marquardt, Ugur Sahin, Peter R. Galle, John Castle, and Andreas Teufel. RNA-Seq Atlas-a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics*, 28(8):1184–1185, 2012.

[127] Philippe Rocca-Serra, Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Sergio Contrino, Jaak Vilo, Niran Abeygunawardena, Gaurab Mukherjee, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, Ahmet Oezcimen, and Susanna Assunta Sansone. ArrayExpress: A public database of gene expression data at EBI. *Comptes Rendus - Biologies*, 326(10-11):1075–1078, 2003.

[128] Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F. Kim, Alexandra Soboleva, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Rolf N. Muertter, and Ron Edgar. NCBI GEO: Archive for high-throughput functional genomic data. *Nucleic Acids Research*, 37:885–890, 2009.

[129] Anika Oellrich and Damian Smedley. Linking tissues to phenotypes using gene expression profiles. *Database*, 2014, 2014.

[130] Ruth Barshir, Omer Shwartz, Ilan Y. Smoly, and Esti Yeger-Lotem. Comparative Analysis of Human Tissue Interactomes Reveals Factors Leading to Tissue-Specific Manifestation of Hereditary Diseases. *PLoS Computational Biology*, 10(6), 2014.

[131] Samir Zaidi, Murim Choi, Hiroko Wakimoto, Lijiang Ma, Jianming Jiang, John D. Overton, Angela Romano-Adesman, Robert D. Bjornson, Roger E. Breitbart, Kerry K. Brown, Nicholas J. Carriero, Yee Him Cheung, John Deanfield, Steve Depalma, Khalid A. Fakhro, Joseph Glessner, Hakon Hakonarson, Michael J. Italia, Jonathan R. Kaltman, Juan Kaski, Richard Kim, Jennie K. Kline, Teresa Lee, Jeremy Leipzig, Alexander Lopez, Shrikant M. Mane, Laura E. Mitchell, Jane W. Newburger, Michael Parfenov, Itsik Pe'Er, George Porter, Amy E. Roberts, Ravi Sachidanandam, Stephan J. Sanders, Howard S. Seiden, Mathew W. State, Sailakshmi Subramanian, Irina R. Tikhonova, Wei

Wang, Dorothy Warburton, Peter S. White, Ismee A. Williams, Hongyu Zhao, Jonathan G. Seidman, Martina Brueckner, Wendy K. Chung, Bruce D. Gelb, Elizabeth Goldmuntz, Christine E. Seidman, and Richard P. Lifton. De novo mutations in histone-modifying genes in congenital heart disease. *Nature*, 498(7453):220–223, 2013.

[132] Ariel Feiglin, Bryce K Allen, Isaac S Kohane, and Sek Won Kong. Comprehensive Analysis of Tissue-wide Gene Expression and Phenotype Data Reveals Tissues Affected in Rare Genetic Disorders. *Cell Systems*, 5:140–148, 2017.

[133] Diane Fatkin, Christine E Seidman, and Jonathan G Seidman. Genetics and disease of ventricular muscle. *Cold Spring Harbor Perspectives in Medicine*, 4, 2014.

[134] Jonathan Mosley, Mark Benson, J Gustav Smith, Olle Melander, Debby Ngo, Christian M Shaffer, Jane F Ferguson, Herzig Matthew, Catherine A McCarty, Christopher Chute, Gail P Jarvik, and Adam Gordon. Probing the virtual proteome to identify novel disease biomarkers. *Circulation*, 138:22, 2018.

[135] Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, Daniel I Chasman, Garret A Fitzgerald, Kara Dolinski, Tilo Grosser, and Olga G Troyanskaya. Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6):569–576, 2015.

[136] Marta Melé, Pedro G Ferreira, Ferran Reverter, David S DeLuca, Jean Monlong, and Michael Sammeth. The human transcriptome across tissues and individuals. *Human Genomics*, 348(6235), 2015.

[137] Jeanne Hansen, Chelsi Snow, Emily Tuttle, Dalia H. Ghoneim, Chun Song Yang, Adam Spencer, Sonya A. Gunter, Christopher D. Smyser, Christina A. Gurnett, Marwan Shinawi, William B. Dobyns, James Wheless, Marc W. Halterman, Laura A. Jansen, Bryce M. Paschal, and Alex R. Paciorkowski. De Novo Mutations in SIK1 Cause a Spectrum of Developmental Epilepsies. *American Journal of Human Genetics*, 96(6):1009, 2015.

[138] Kasper Lage, Niclas Tue Hansen, E Olof Karlberg, Aron C Eklund, Francisco S Roque, Patricia K Donahoe, Zoltan Szallasi, Thomas Skøt Jensen, and Søren Brunak. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 105(52):20870–5, 2008.

[139] Janna E Hutz, Aldi T Kraja, Howard L McLeod, and Michael A Province. CANDID: A flexible method for prioritizing candidate genes for complex human traits. *Genetic Epidemiology*, 32(8):779–790, 2008.

[140] Agne Antanaviciute, Catherine Daly, Laura A Crinnion, Alexander F Markham, Christopher M Watson, David T Bonthron, and Ian M Carr. GeneTIER: Prioritization of candidate disease genes using tissue-specific gene expression profiles. *Bioinformatics*, 31(16):2728–2735, 2015.

[141] Léon-Charles Tranchevent, Amin Ardeshirdavani, Sarah ElShal, Daniel Alcaide, Jan Aerts, Didier Auboeuf, and Yves Moreau. Candidate gene prioritization with Endeavour. *Nucleic Acids Research*, 44:W117–W121, 2016.

[142] Patrick Deelen, Sipko van Dam, Johanna C Herkert, Juha M Karjalainen, Harm Brugge, Kristin M Abbott, Cleo C van Diemen, Paul A van der Zwaag, Erica H Gerkes, Evelien Zonneveld-Huijssoon, Jelkje J Boer-Bergsma, Pytrik Folkertsma, Tessa Gillett, K. Joeri van der Velde, Roan Kanninga, Peter C van den Akker, Sabrina Z. Jan, Edgar T. Hoorntje, Wouter P. te Rijdt, Yvonne J. Vos, Jan D.H. Jongbloed, Conny M.A. van Ravenswaaij-Arts, Richard Sinke, Birgit Sikkema-Raddatz, Wilhelmina S. Kerstjens-Frederikse, Morris A. Swertz, and Lude Franke. Improving the diagnostic yield of exome- sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. *Nature Communications*, 10, 2019.

[143] Rasko Leinonen, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, Richard Gibson, Gemma Hoad, Mikyung Jang, Nima Pakseresht, Sheila Plaister, Rajesh Radhakrishnan, Kethi Reddy, Siamak Sobhany, Petra Ten Hoopen, Robert Vaughan, Vadim Zalunin, and Guy Cochrane. The European Nucleotide Archive. *Nucleic Acids Research*, 39, 2010.

[144] Loris Bertoldi, Claudio Forcato, Nicola Vitulo, Giovanni Birolo, Fabio De Pascale, Erika Feltrin, Riccardo Schiavon, Franca Anglani, Susanna Negrisolo, Alessandra Zanetti, Francesca D'Avanzo, Rosella Tomanin, Georgine Faulkner, Alessandro Vezzi, and Giorgio Valle. QueryOR: a comprehensive web platform for genetic variant analysis and prioritization. *BMC Bioinformatics*, 18(225), 2017.

[145] Daniel C. Koboldt, David E. Larson, Lori S. Sullivan, Sara J. Bowne, Karyn M. Steinberg, Jennifer D. Churchill, Aimee C. Buhr, Nathan Nutter, Eric A. Pierce, Susan H. Blanton, George M. Weinstock, Richard K. Wilson, and Stephen P. Daiger. Exome-based mapping and variant prioritization for inherited mendelian disorders. *The American Journal of Human Genetics*, 94:373–384, 2014.

[146] Rare Diseases Models & Mechanisms – Europe (RDMM-Europe) – Solve-RD. http://solve-rd.eu/rdmm-europe/, accessed 2020-05-17.

[147] Fine-Lubinsky syndrome | Genetic and Rare Diseases Information Center (GARD) – an NCATS Program. https://rarediseases.info.nih.gov/diseases/958/fine-lubinsky-syndrome, accessed 2018-08-13.

[148] J Roman Corona-Rivera, Eloy Lopez-Marure, Diana Garcia-Cruz, Carmen O Romo-Huerta, Alejandro Rea-Rosas, L Gustavo Orozco-Alatorre, and J Manuel Ramirez-Valdivia. Further clinical delineation of Fine-Lubinsky syndrome. *The American Journal of Medical Genetics*, Part A:1070–1075, 2009.

[149] Matthew H.K. Cheng and Ralf Peter Jansen. A jack of all trades: the RNA-binding protein vigilin. *WIREs RNA*, 8:1–15, 2017.

[150] Alfredo Castello, Bernd Fischer, Matthias W. Hentze, and Thomas Preiss. RNA-binding proteins in Mendelian disease. *Trends in Genetics*, 29(5):318–327, 2013.

[151] Alger M Fredericks, Kamil J Cygan, Brian a Brown, William G Fairbrother, and Cell Biology. RNA-Binding Proteins: Splicing Factors and Disease. *Biomolecules*, 5:893–909, 2015.

[152] HuaLin Zhou, Marie Mangelsdorf, JiangHong Liu, Li Zhu, and Jane Y Wu. RNA-binding proteins in neurological diseases. *Science China Life Sciences*, 57(4):432–444, 2014.

[153] Kiven E Lukong, Kai-wei Chang, Edouard W Khandjian, and Stéphane Richard. RNA-binding proteins in human genetic disease. http://dx.doi.org/10.1016/j.tig.2008.05.004, access 2019-08-04, 2008.

[154] OMIM entry on 2q370-deletion syndrome.

[155] Camille Leroy, Emilie Landais, Sylvain Briault, Albert David, Olivier Tassy, Nicolas Gruchy, Bruno Delobel, Marie Jose Grégoire, Bruno Leheup, Laurence Taine, Didier Lacombe, Marie Ange Delrue, Annick Toutain, Agathe Paubel, Francine Mugneret, Christel Thauvin-Robinet, Stephanie Arpin, Cedric Le Caignec, Philippe Jonveaux, Mylene Beri, Nathalie Leporrier, Jacques Motte, Caroline Fiquet, Olivier Brichet, Monique Mozelle-Nivoix, Pascal Sabouraud, Nathalie Golovkine, Nathalie Bednarek, Dominique Gaillard, and Martine Doco-Fenzy. The 2q37-deletion syndrome: An update of the clinical spectrum including overweight, brachydactyly and behavioural features in 14 new patients. *European Journal of Human Genetics*, 21:602–612, 2013.

[156] Bärbel Felder, Bernhard Radlwimmer, Axel Benner, Antoaneta Mincheva, Grischa Tödt, Kim S Beyer, Claudia Schuster, Sven Bölte, Gabriele Schmötzer, Sabine M Klauck, Fritz Poustka, Peter Lichter, and Annemarie Poustka. FARP2, HDLBP and PASK are downregulated in a patient with autism and 2q37.3 deletion syndrome. *American Journal of Medical Genetics, Part A*, 149:952–959, 2009.

[157] Frank Rauch and Francis H Glorieux. Chromosome 2q37 Deletion: Clinical and Molecular Aspects. *American Journal of Medical Genetics Part C, Seminars in medical genetics*, 145C:357–371, 2007.

[158] Xiaoqin Yu, Qiuying Liu, Jinyang He, Yuan Huang, Lei Jiang, Xiaoyan Xie, Ji Liu, Lihong Chen, Ling Wei, and Yang Qin. Vigilin interacts with CTCF and is involved in the maintenance of imprinting of IGF2 through a novel RNA–mediated mechanism. *International Journal of Biological Macromolecules*, 108:515–522, 2018.

[159] Picard Tools - By Broad Institute. http://broadinstitute.github.io/picard/, accessed 2019-07-08.

[160] Sean Davis. GitHub - seandavi/ngCGH: Tools for producing pseudo-cgh of next-generation sequencing data. https://github.com/seandavi/ngCGH, accessed 2020-04-05.

[161] Katayoon Darvishi. Application of nexus copy number software for CNV detection and analysis. *Current Protocols in Human Genetics*, 4(14):1–28, 2010.

[162] Niko Popitsch. CODOC: Efficient access, analysis and compression of depth of coverage signals. *Bioinformatics*, 30(18):2676–2677, 2014.

[163] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature Bio*, 29(1):24–26, 2011.

[164] Fabric Genomics. Omicia Opal 4.24.0. https://app.omicia.com/, accessed 2016-11-07.

[165] GeneArt Gene Synthesis and Services | Thermo Fisher Scientific - US. https://www.thermofisher.com/us/en/home/life-science/cloning/gene-synthesis.html, accessed 2020-04-09.

[166] PCR Primer Design Tool.

[167] Q00341 | SWISS-MODEL Repository.

[168] RCSB PDB - 4B8T: RNA BINDING PROTEIN Solution structure of the third KH domain of KSRP in complex with the G-rich target sequence.

[169] Relax application.

[170] PyMOL | pymol.org.

[171] Vanessa Sawyer. Fabric Genomics Partners to Improve Pediatric Care | Business Wire. https://www.businesswire.com/news/home/20171019005274/en/Fabric-Genomics-Partners-Improve-Pediatric-Care, accessed 2018-12-04.

[172] Melina Claussnitzer, Judy H Cho, Rory Collins, Nancy J Cox, Emmanouil T Dermitzakis, Matthew E Hurles, Sekar Kathiresan, Eimear E Kenny, Cecilia M Lindgren, Daniel G Macarthur, Kathryn N North, Sharon E Plon, Heidi L Rehm, Neil Risch, Charles N Rotimi, Jay Shendure, Nicole Soranzo, and Mark I McCarthy. A brief history of human disease genetics. *Nature*, 577:179–189, 2020.

[173] Reuben J Pengelly, Thahmina Alom, Zijian Zhang, David Hunt, Sarah Ennis, and Andrew Collins. Evaluating phenotype-driven approaches for genetic diagnoses from exomes in a clinical setting. *Scientific Reports*, 7(13509), 2017.

[174] Agne Antanaviciute, Christopher M. Watson, Sally M. Harrison, Carolina Lascelles, Laura Crinnion, Alexander F. Markham, David T. Bonthron, and Ian M. Carr. OVA: Integrating molecular and physical phenotype data from multiple biomedical domain ontologies with variant filtering for enhanced variant prioritization. *Bioinformatics*, 31(23):3822–3829, 2015.

[175] Teresa Requena, Alvaro Gallego-Martinez, and Jose A. Lopez-Escamez. A pipeline combining multiple strategies for prioritizing heterozygous variants for the identification of candidate genes in exome datasets. *Human Genomics*, 11(1):1–11, 2017.

[176] Asif Javed, Saloni Agrawal, and Pauline C Ng. Phen-Gen : combining phenotype and genotype to analyze rare disorders. *Nature Methods*, 11(9), 2014.

[177] UCSC Table Browser. https://genome.ucsc.edu/cgi-bin/hgTables, accessed 2016-04-29.

[178] Aaron R Quinlan and Ira M Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[179] Human Phenotype Ontology Statistics. https://hpo.jax.org/app/, accessed 2019-05-27.

[180] Vida Čulić, Noriko Miyake, Sunčana Janković, Davor Petrović, Marko Šimunović, Tomislav Đapić, Masaaki Shiina, Kazuhiro Ogata, and Naomichi Matsumoto. Distal arthrogryposis with variable clinical expression caused by TNNI2 mutation. *Human Genome Variation*, 3(September):16035, 2016.

[181] Arjune Sen, Patricia Dugan, Piero Perucca, Daniel Costello, Hyunmi Choi, Carl Bazil, Rod Radtke, Danielle Andrade, Chantal Depondt, Sinead Heavin, Jane Adcock, W. Owen Pickrell, Ronan McGinty, Fábio Nascimento, Philip Smith, Mark I. Rees, Patrick Kwan, Terence J. O'Brien, David Goldstein, and Norman Delanty. The phenotype of bilateral hippocampal sclerosis and its management in "real life" clinical settings. *Epilepsia*, 59(7):1410–1420, 2018.

[182] CACNA1E calcium voltage-gated channel subunit alpha1 E [Homo sapiens (human)]. http://www.ncbi.nlm.nih.gov/pubmed/777, accessed 2019-05-27, 2019.

[183] Katherine L Helbig, Robert J Lauerer, Jacqueline C Bahr, Ivana A Souza, Candace T Myers, Betül Uysal, Niklas Schwarz, Maria A. Gandini, Sun Huang, Boris Keren, Cyril Mignot, Alexandra Afenjar, Thierry Billette de Villemeur, Delphine Héron, Caroline Nava, Stéphanie Valence, Julien Buratti, Christina R. Fagerberg, Kristina P. Soerensen, Maria Kibaek, Erik Jan Kamsteeg, David A. Koolen, Boudewijn Gunning, H. Jurgen Schelhaas, Michael C. Kruer, Jordana Fox, Somayeh Bakhtiari, Randa Jarrar, Sergio Padilla-Lopez, Kristin Lindstrom, Sheng Chih Jin, Xue Zeng, Kaya Bilguvar, Antigone Papavasileiou, Qinghe Xin, Changlian Zhu, Katja Boysen, Filippo Vairo, Brendan C. Lanpher, Eric W. Klee, Jan Mendelt Tillema, Eric T. Payne, Margot A. Cousin, Teresa M. Kruisselbrink, Myra J. Wick, Joshua Baker, Eric Haan, Nicholas Smith, Mark A. Corbett, Alastair H. MacLennan, Jozef Gecz, Saskia Biskup, Eva Goldmann, Lance H. Rodan, Elizabeth Kichula, Eric Segal, Kelly E. Jackson, Alexander Asamoah, David Dimmock, Julie McCarrier, Lorenzo D. Botto, Francis Filloux, Tatiana Tvrdik, Gregory D. Cascino, Sherry Klingerman, Catherine Neumann, Raymond Wang, Jessie C. Jacobsen, Melinda A. Nolan, Russell G. Snell, Klaus Lehnert, Lynette G. Sadleir, Britt Marie Anderlid, Malin Kvarnung, Renzo Guerrini, Michael J. Friez, Michael J. Lyons, Jennifer Leonhard, Gabriel Kringlen, Kari Casas, Christelle M. El Achkar, Lacey A. Smith, Alexander Rotenberg, Annapurna Poduri, Alba Sanchis-Juan, Keren J. Carss, Julia Rankin, Adam Zeman, F. Lucy Raymond, Moira Blyth, Bronwyn Kerr, Karla Ruiz, Jill Urquhart, Imelda Hughes, Siddharth Banka, Ulrike B.S. Hedrich, Ingrid E. Scheffer, Ingo Helbig, Gerald W. Zamponi, Holger Lerche, and Heather C. Mefford. De Novo Pathogenic Variants in CACNA1E Cause Developmental and

Epileptic Encephalopathy with Contractures, Macrocephaly, and Dyskinesias. *The American Journal of Human Genetics*, 103:666–678, 2018.

[184] Juliette Piard, Lara Hawkes, Mathieu Milh, Laurent Villard, Renato Borgatti, Romina Romaniello, Melanie Fradin, Yline Capri, Delphine Héron, Marie-Christine Nougues, Caroline Nava, Oana Tarta Arsene, Debbie Shears, Yoshimi Sogawa, Diana Johnson, Helen Firth, Pradeep Vasudevan, Gabriela Jones, Marie-Ange Nguyen-Morel, Tiffany Busa, Agathe Roubertie, Myrthe van den Born, Elise Brischoux-Boucher, Michel Koenig, Cyril Mignot, Usha Kini, and Christophe Philippe. The phenotypic spectrum of WWOX-related disorders: 20 additional cases of WOREE syndrome and review of the literature. *Genetics in Medicine*, 0(0):1–11, 2018.

[185] Salma Ben-Salem, Aisha M Al-Shamsi, Anne John, Bassam R Ali, and Lihadh Al-Gazali. A Novel Whole Exon Deletion in WWOX Gene Causes Early Epilepsy, Intellectual Disability and Optic Atrophy. *Journal of Molecular Neuroscience*, 56:17–23, 2015.

[186] Cyril Mignot, Laetitia Lambert, Laurent Pasquier, Thierry Bienvenu, Andrée Delahaye-Duriez, Boris Keren, Jérémie Lefranc, Aline Saunier, Lila Allou, Virginie Roth, Mylène Valduga, Aissa Moustaïne, Stéphane Auvin, Catherine Barrey, Sandra Chantot-Bastaraud, Nicolas Lebrun, Marie-Laure Moutard, Marie-Christine Nougues, Anne-Isabelle Vermersch, Bénédicte Héron, Eva Pipiras, Delphine Héron, Laurence Olivier-Faivre, Jean-Louis Guéant, Philippe Jonveaux, and Christophe Philippe. WWOX-related encephalopathies: delineation of the phenotypical spectrum and emerging genotype-phenotype correlation. *Journal of Medical Genetics*, 52(1):61–70, jan 2015.

[187] Mylène Valduga, Christophe Philippe, Laetitia Lambert, Pascale Bach-Segura, Emmanuelle Schmitt, Jean Pierre Masutti, Bénédicte François, Patrick Pinaud, Mireille Vibert, and Philippe Jonveaux. WWOX and severe autosomal recessive epileptic encephalopathy: first case in the prenatal period. *Journal of Human Genetics*, 60:267–271, 2015.

[188] Kae Won Cho, Jongsung Lee, and Youngjo Kim. Genetic Variations Leading to Familial Dilated Cardiomyopathy. *Molecules and Cells*, 2016.

[189] Timothy M Olson, Virginia V Michels, Stephen N Thibodeau, Yin-Shan Tai, and Mark T Keating. Actin mutations in dilated cardiomyopathy, a heritable form of heart failure. *Science*, 280:750–752, 1998.

[190] Zakiya S Al-Mosawi, Khulood K Al-Saad, Roya Ijadi-Maghsoodi, Hatem I El-Shanti, and Polly J Ferguson. A Splice Site Mutation Confirms the Role of LPIN2 in Majeed Syndrome. *Arthritis*, 56(3):960–964, 2007.

[191] Elisabeth J Smith, Florence Allantaz, Lynda Bennett, Dongping Zhang, Xiaochong Gao, Geryl Wood, Daniel L Kastner, Marilynn Punaro, Ivona Aksentijevich, Virginia Pascual, and Carol A Wise. Clinical, Molecular, and Genetic Characteristics of PAPA Syndrome: A Review. *Current Genomics*, 11:519–527, 2010.

[192] Hatem El-Shanti and Polly Ferguson. Majeed Syndrome. *GeneReviews*, 2008.

[193] Rory Bowden, Robert W Davies, Andreas Heger, Alistair T Pagnamenta, Mariateresa de Cesare, Laura E Oikkonen, Duncan Parkes, Colin Freeman, Fatima Dhalla, Smita Y Patel, Niko Popitsch, Camilla L C Ip, Hannah E Roberts, Silvia Salatino, Helen Lockstone, Gerton Lunter, Jenny C Taylor, David Buck, Michael A Simpson, and Peter Donnelly. Sequencing of human genomes with nanopore technology. *Nature Communications*, 10(1869), 2019.

[194] Dong-Hui Chen, Jennifer E Below, Akiko Shimamura, Sioban B Keel, Mark Matsushita, John Wolff, Youngmee Sul, Emily Bonkowski, Maria Castella, Toshiyasu Taniguchi, Deborah Nickerson, Thalia Papayannopoulou, Thomas D Bird, and Wendy H Raskind. Ataxia-Pancytopenia Syndrome Is Caused by Missense Mutations in SAMD9L. *The American Journal of Human Genetics*, 98:1146–1158, 2016.

[195] Bianca Tesi, Josef Davidsson, Matthias Voss, Elisa Rahikkala, Tim D Holmes, Samuel C.C. Chiang, Jonna Komulainen-Ebrahim, Sorina Gorcenco, Alexandra Rundberg Nilsson, Tim Ripperger, Hannaleena Kokkonen, David Bryder, Thoas Fioretos, Jan Inge Henter, Merja Möttönen, Riitta Niinimäki, Lars Nilsson, Cornelis Jan Pronk, Andreas Puschmann, Hong Qian, Johanna Uusimaa, Jukka Moilanen, Ulf Tedgård, Jörg Cammenga, and Yenan T. Bryceson. Gain-of-function SAMD9L mutations cause a syndrome of cytopenia, immunodeficiency, MDS, and neurological symptoms. *Blood*, 129(16):2266–2279, 2017.

[196] Jason R Schwartz, Jing Ma, Tamara Lamprecht, Michael Walsh, and Shuoguo Wang. The genomic landscape of pediatric myelodysplastic syndromes. *Nature Communications*, 8(1557), 2017.

[197] Olivier Bluteau, Marie Sebert, Thierry Leblanc, Régis Peffault De Latour, Samuel Quentin, Elodie Lainey, Lucie Hernandez, Jean Hugues Dalle, Flore Sicre De Fontbrune, Etienne Lengline, Raphael Itzykson, Emmanuelle Clappier, Nicolas Boissel, Nadia Vasquez, Mélanie Da Costa, Julien Masliah-Planchon, Wendy Cuccuini, Anna Raimbault, Louis De Jaegere, Lionel Adès, Pierre Fenaux, Sébastien Maury, Claudine Schmitt, Marc Muller, Carine Domenech, Nicolas Blin, Bénédicte Bruno, Isabelle Pellier, Mathilde Hunault, Stéphane Blanche, Arnaud Petit, Guy Leverger, Gérard Michel, Yves Bertrand, André Baruchel, Gérard Socié, and Jean Soulier. A landscape of germ line mutations in a cohort of inherited bone marrow failure patients. *Blood*, 131(7):717–732, 2018.

[198] Jasmine C Wong, Kevin Shannon, and Jeffery M Klco. Germline SAMD9 and SAMD9L mutations are associated with extensive genetic evolution and diverse hematologic outcomes. *JCI Insight*, 3(14), 2018.

[199] Orphanet: Antley Bixler syndrome. http://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=EN&Expert=83, accessed 2016-11-02.

[200] Ashley M Holder, Brett H Graham, Brendan Lee, and Daryl A Scott. Fine–Lubinsky syndrome: Sibling pair suggests possible autosomal recessive inheritance. *American Journal of Medical Genetics Part A*, 143A:2576–2580, nov 2007.

[201] Federico Andrea Santoni, Periklis Makrythanasis, and Stylianos E Antonarakis. CATCHing putative causative variants in consanguineous families. *BMC Bioinformatics*, 16(1):310, 2015.

[202] M F McMullin. Congenital erythrocytosis. *International Journal of Laboratory Hematology*, 38:59–65, 2016.

[203] Marialuisa Quadri, Antonio Federico, Tianna Zhao, Guido J Breedveld, Carla Battisti, Cathérine Delnooz, Lies Anne Severijnen, Lara Di Toro Mammarella, Andrea Mignarri, Lucia Monti, Antioco Sanna, Peng Lu, Francesca Punzo, Giovanni Cossu, Rob Willemsen, Fabrizio Rasi, Ben A. Oostra, Bart P. Van De Warrenburg, and Vincenzo Bonifati. Mutations in SLC30A10 cause Parkinsonism and Dystonia with Hypermanganesemia, Polycythemia, and Chronic Liver Disease. *American Journal of Human Genetics*, 90:467–477, 2012.

[204] Karin Tuschl, Peter T. Clayton, Sidney M. Gospe, Shamshad Gulab, Shahnaz Ibrahim, Pratibha Singhi, Roosy Aulakh, Reinaldo T. Ribeiro, Orlando G. Barsottini, Maha S. Zaki, Maria Luz Del Rosario, Sarah Dyack, Victoria Price, Andrea Rideout, Kevin Gordon, Ron A. Wevers, W. K. Kling Chong, and Philippa B. Mills. Syndrome of hepatic cirrhosis, dystonia, polycythemia, and hypermanganesemia caused by mutations in SLC30A10, a manganese transporter in man. *American Journal of Human Genetics*, 90:457–466, 2012.

[205] Benjamin L Ebert and H Franklin Bunn. Regulation of the Erythropoietin Gene. *Blood*, 94(6):1864–1877, 1999.

[206] M Michael Cohen. Klippel-Trenaunay Syndrome. *American Journal of Human Genetics*, 93:171–175, 2000.

[207] Ann P Quick, Qiongling Wang, Leonne E Philippen, Giselle Barreto-Torres, David Y Chiang, David L Beavers, Guoliang Wang, Maha Khalid, Julia O Reynolds, Hannah M Campbell, Jordan Showell, Mark D McCauley, Arjen Scholten, and Xander H Wehrens. Striated Muscle Preferentially Expressed Protein Kinase (SPEG) Is Essential for Cardiac Function by Regulating Junctional Membrane Complex Activity. *Circulation Research*, 2016.

[208] Thomas Gridley. Notch signaling in the vasculature. *Current Topics in Developmental Biology*, 92:277–309, 2010.

[209] Helen C Su. DOCK8 (Dedicator of cytokinesis 8) deficiency. *Current Opinion in Allergy and Clinical Immunology*, 10(6):515–520, 2010.

[210] Fang Yang, Sherman Silber, N Adrian Leu, Robert D Oates, Janet D Marszalek, Helen Skaletsky, Laura G Brown, Steve Rozen, David C Page, and P Jeremy Wang. TEX11 is mutated in infertile men with azoospermia and regulates genome-wide recombination rates in mouse. *EMBO Molecular Medicine*, 7(9):1198–210, 2015.

[211] Kenneth I Aston, Csilla Krausz, Ilaria Laface, E Ruiz-Castané, and Douglas T Carrell. Evaluation of 172 candidate polymorphisms for association with oligozoospermia or azoospermia in a large cohort of men of European descent. *Human Reproduction*, 25(6):1383–1397, 2010.

[212] Alexander N Yatsenko, Andrew P Georgiadis, Albrecht Röpke, Andrea J Berman, Thomas Jaffe, Marta Olszewska, Birgit Westernströer, Joseph Sanfilippo, Maciej Kurpisz, Aleksandar Rajkovic, Svetlana A Yatsenko, Sabine Kliesch, Stefan Schlatt, and Frank Tüttelmann. X-Linked TEX11 Mutations, Meiotic Arrest, and Azoospermia in Infertile Men. *The New England Journal of Medicine*, 372(22):2097–107, 2015.

[213] Anna C Need, Vandana Shashi, Kelly Schoch, Slavé Petrovski, and David B Goldstein. The importance of dynamic re-analysis in diagnostic whole exome sequencing. *Journal of Medical Genetics*, 0(0), 2016.

[214] Aaron M Wenger, Harendra Guturu, Jonathan A Bernstein, and Gill Bejerano. Systematic reanalysis of clinical exome data yields additional diagnoses: Implications for providers. *Genetics in Medicine*, 19(2):209–214, 2017.

[215] Jamie M. Ellingford, Stephanie Barton, Sanjeev Bhaskar, Simon G Williams, Panagiotis I Sergouniotis, James O'Sullivan, Janine A. Lamb, Rahat Perveen, Georgina Hall, William G Newman, Paul N Bishop, Stephen A Roberts, Rick Leach, Rick Tearle, Stuart Bayliss, Simon C Ramsden, Andrea H Nemeth, and Graeme C M Black. Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current Diagnostic Testing for Inherited Retinal Disease. *Ophthalmology*, 123:1143–1150, 2016.

[216] Michelle M Clark, Amber Hildreth, Sergey Batalov, Yan Ding, Shimul Chowdhury, Kelly Watkins, Katarzyna Ellsworth, Brandon Camp, Cyrielle I Kint, Calum Yacoubian, Lauge Farnaes, Matthew N Bainbridge, Curtis Beebe, Joshua J A Braun, Margaret Bray, Jeanne Carroll, Julie A Cakici, Sara A Caylor, Christina Clarke, Mitchell P Creed, Jennifer Friedman, Alison Frith, Richard Gain, Mary Gaughran, Shauna George, Sheldon Gilmer, Joseph Gleeson, Jeremy Gore, Haiying Grunenwald, Raymond L Hovey, Marie L Janes, Kejia Lin, Paul D Mcdonagh, Kyle Mcbride, Patrick Mulrooney, Shareef Nahas, Daeheon Oh, Albert Oriol, Laura Puckett, Zia Rady, Martin G Reese, Julie Ryu, Lisa Salz, Erica Sanford, Lawrence Stewart, Nathaly Sweeney, Mari Tokita, Luca Van Der Kraan, Sarah White, Kristen Wigby, Brett Williams, Terence Wong, Meredith S Wright, Catherine Yamada, Peter Schols, John Reynders, Kevin Hall, David Dimmock, Narayanan Veeraraghavan, Thomas Defay, and Stephen F Kingsmore. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Science Translational Medicine*, 11, 2019.

[217] Max Schubach, Matteo Re, Peter N Robinson, and Giorgio Valentini. Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. *Scientific Reports*, 7(2959), 2017.

[218] FATHMM data downloads. http://fathmm.biocompute.org.uk/downloads.html, accessed 2019-11-24.

[219] Hashem A Shihab, Julian Gough, Matthew Mort, David N Cooper, Ian N M Day, and Tom R Gaunt. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Human Genomics*, 8(11), 2014.

[220] Mark F Rogers, Hashem A Shihab, Matthew Mort, David N Cooper, Tom R Gaunt, and Colin Campbell. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, pages 1–3, 2017.

[221] scikit-learn 0.21.2 documentation. https://scikit-learn.org/stable/, accessed 2019-07-25.

[222] scikit-learn 0.21.2 documentation - Random Forests. https://scikit-learn.org/stable/modules/ensemble.html#forest, 2019-07-26.

[223] scikit-learn 0.21.2 documentation - StratifiedKFold. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html, accessed 2019-07-25.

[224] ExAC HDLBP entry. http://exac.broadinstitute.org/gene/ENSG00000115677, accessed 2018-09-04.

[225] Jana Marie Schwarz, David N. Cooper, Markus Schuelke, and Dominik Seelow. Mutationtaster2: Mutation prediction for the deep-sequencing age. *Nature Methods*, 11(4):361–362, 2014.

[226] Stephen R Williams, Micheala A Aldred, Vazken M Der Kaloustian, Fahed Halal, Gordon Gowans, D. Ross McLeod, Sara Zondag, Helga V Toriello, R. Ellen Magenis, and Sarah H Elsea. Haploinsufficiency of HDAC4 causes brachydactyly mental retardation syndrome, with brachydactyly type E, developmental delays, and behavioral problems. *The American Journal of Human Genetics*, 87:219–228, 2010.

[227] Pablo Villavicencio-Lorini, Eva Klopocki, Marc Trimborn, Randi Koll, Stefan Mundlos, and Denise Horn. Phenotypic variant of Brachydactyly-mental retardation syndrome in a family with an inherited interstitial 2q37.3 microdeletion including HDAC4. *European Journal of Human Genetics*, 21:743–748, 2013.

[228] Histone Deacetylase 4 (HDAC4) OMIM entry. https://www.omim.org/entry/605314?search=brachydactyly mental retardation syndrome&highlight=mental brachydactyly syndromic retardation brachydactylous brachydactylic brachydactylia syndrome, accessed 2018-09-03.

[229] Patricia G Wheeler, Dongli Huang, and Zunyan Dai. Haploinsufficiency of HDAC4 does not cause intellectual disability in all affected individuals. *American Journal of Medical Genetics Part A*, 164:1826–1829, jul 2014.

[230] Claudia Strein, Anne-Marie Alleaume, Ulrich Rothbauer, Matthias W Hentze, and Alfredo Castello. A versatile assay for RNA-binding proteins in living cells. *RNA*, 20:721–731, 2014.

[231] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473:337–342, 2011.

[232] Alfredo Castello, Bernd Fischer, Christian K. Frese, Rastislav Horos, Anne Marie Alleaume, Sophia Foehr, Tomaz Curk, Jeroen Krijgsveld, and Matthias W. Hentze. Comprehensive Identification of RNA-Binding Domains in Human Cells. *Molecular Cell*, 63:696–710, 2016.

[233] Alfredo Castello, Bernd Fischer, Christian K. Frese, Rastislav Horos, Anne Marie Alleaume, Sophia Foehr, Tomaz Curk, Jeroen Krijgsveld, and Matthias W. Hentze. RBDmap server.

[234] Wolf D. Hirschmann, Heidrun Westendorf, Andreas Mayer, Gina Cannarozzi, Patrick Cramer, and Ralf Peter Jansen. Scp160p is required for translational efficiency of codon-optimized mRNAs in yeast. *Nucleic Acids Research*, 42(6):4043–4055, 2014.

[235] GTEx Portal v8. https://www.gtexportal.org/home/tissueSummaryPage, accessed 2020-03-29.

[236] Josine L Min, George Nicholson, Ingileif Halgrimsdottir, Kristian Almstrup, Andreas Petri, Amy Barrett, Mary Travers, Nigel W Rayner, Reedik Mägi, Fredrik H Pettersson, John Broxholme, Matt J Neville, Quin F Wills, Jane Cheeseman, Maxine Allen, Chris C Holmes, Tim D. Spector, Jan Fleckner, Mark I. McCarthy, Fredrik Karpe, Cecilia M. Lindgren, and Krina T. Zondervan. Coexpression network analysis in abdominal and gluteal adipose tissue reveals regulatory genetic loci for metabolic syndrome and related phenotypes. *PLoS Genetics*, 8(2), 2012.

[237] Tomasz Konopka Id and Damian Smedley. Incremental data integration for tracking genotype-disease associations. *PLoS Computational Biology*, 16(1), 2020.

[238] Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M. Krawitz, Susanne B Kamphausen, Martin Zenker, Lynne M Bird, and Karen W Gripp. Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, 25:60–64, 2019.

[239] Tudor Groza, Sebastian Köhler, Sandra Doelken, Nigel Collier, Anika Oellrich, Damian Smedley, Francisco M. Couto, Gareth Baynam, Andreas Zankl, and Peter N. Robinson. Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database*, pages 1–13, 2015.

[240] Anthony A Philippakis, Danielle R Azzariti, Sergi Beltran, Anthony J Brookes, Catherine A Brownstein, Michael Brudno, Han G Brunner, Orion J. Buske, Knox Carey, Cassie Doll, Sergiu Dumitriu, Stephanie O M Dyke, Johan T. den Dunnen, Helen V. Firth, Richard A. Gibbs, Marta Girdea, Michael Gonzalez, Melissa A. Haendel, Ada Hamosh, Ingrid A. Holm, Lijia Huang, Matthew E. Hurles, Ben Hutton, Joel B. Krier, Andriy Misyura, Christopher J. Mungall, Justin Paschall, Benedict Paten, Peter N. Robinson, François Schiettecatte, Nara L. Sobreira, Ganesh J. Swaminathan, Peter E. Taschner, Sharon F. Terry, Nicole L. Washington, Stephan Züchner, Kym M. Boycott, and Heidi L. Rehm. The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. *Human Mutation*, 36(10):915–921, 2015.

[241] DECIPHER. https://decipher.sanger.ac.uk/index, accessed 2016-10-30.

[242] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, Laura D Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M England, Eleanor G Seaby, Jack A Kosmicki, Raymond K Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X Chong, Kaitlin E Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S Ware, Christopher Vittal, Irina M Armean, Louis Bergelson, Kristian Cibulskis, Kristen M Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E Talkowski, Carlos A. Aguilar Salinas, Tariq Ahmad, Christine M. Albert, Diego Ardissino, Gil Atzmon, John Barnard, Laurent Beaugerie, Emelia J. Benjamin, Michael Boehnke, Lori L. Bonnycastle, Erwin P. Bottinger, Donald W. Bowden, Matthew J. Bown, John C. Chambers, Juliana C. Chan, Daniel Chasman, Judy Cho, Mina K. Chung, Bruce Cohen, Adolfo Correa, Dana Dabelea, Mark J. Daly, Dawood Darbar, Ravindranath Duggirala, Josée Dupuis, Patrick T. Ellinor, Roberto Elosua, Jeanette Erdmann, Tõnu Esko, Martti Färkkilä, Jose Florez, Andre Franke, Gad Getz, Benjamin Glaser, Stephen J. Glatt, David Goldstein, Clicerio Gonzalez, Leif Groop, Christopher Haiman, Craig Hanis, Matthew Harms, Mikko Hiltunen, Matti M. Holi, Christina M. Hultman, Mikko Kallela, Jaakko Kaprio, Sekar Kathiresan, Bong Jo Kim, Young Jin Kim, George Kirov, Jaspal Kooner, Seppo Koskinen, Harlan M. Krumholz, Subra Kugathasan, Soo Heon Kwak, Markku Laakso, Terho Lehtimäki, Ruth J.F. Loos, Steven A. Lubitz, Ronald C.W. Ma, Daniel G. MacArthur, Jaume Marrugat, Kari M. Mattila, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, James B. Meigs, Olle Melander, Andres Metspalu, Benjamin M. Neale, Peter M. Nilsson, Michael C. O'Donovan, Dost Ongur, Lorena Orozco, Michael J. Owen, Colin N.A. Palmer, Aarno Palotie, Kyong Soo Park, Carlos Pato, Ann E. Pulver, Nazneen Rahman, Anne M. Remes, John D. Rioux, Samuli Ripatti, Dan M. Roden, Danish Saleheen, Veikko Salomaa, Nilesh J. Samani, Jeremiah Scharf, Heribert Schunkert, Moore B. Shoemaker, Pamela Sklar, Hilkka Soininen, Harry Sokol, Tim Spector, Patrick F. Sullivan, Jaana Suvisaari, E. Shyong Tai, Yik Ying Teo, Tuomi Tiinamaija, Ming Tsuang, Dan Turner, Teresa Tusie-Luna, Erkki Vartiainen, James S. Ware, Hugh Watkins, Rinse K. Weersma, Maija Wessman, James G. Wilson, Ramnik J. Xavier, Benjamin M. Neale, Mark J. Daly, and Daniel G. MacArthur. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581:434–443, 2020.

[243] Laure Frésard, Craig Smail, Nicole M Ferraro, Nicole A Teran, Xin Li, Kevin S Smith, Devon Bonner, Kristin D. Kernohan, Shruti Marwaha, Zachary Zappala, Brunilda Balliu, Joe R. Davis, Boxiang Liu, Cameron J. Prybol, Jennefer N. Kohler, Diane B. Zastrow, Chloe M. Reuter, Dianna G. Fisk, Megan E. Grove, Jean M. Davidson, Taila Hartley, Ruchi Joshi, Benjamin J. Strober, Sowmithri Utiramerur, David R. Adams, Aaron Aday, Mercedes E. Alejandro, Patrick Allard, Euan A. Ashley, Mahshid S. Azamian, Carlos A. Bacino, Eva Baker, Ashok Balasubramanyam, Hayk Barseghyan, Gabriel F. Batzli, Alan H. Beggs,

Babak Behnam, Hugo J. Bellen, Jonathan A. Bernstein, Gerard T. Berry, Anna Bican, David P. Bick, Camille L. Birch, Devon Bonner, Braden E. Boone, Bret L. Bostwick, Lauren C. Briere, Elly Brokamp, Donna M. Brown, Matthew Brush, Elizabeth A. Burke, Lindsay C. Burrage, Manish J. Butte, Shan Chen, Gary D. Clark, Terra R. Coakley, Joy D. Cogan, Heather A. Colley, Cynthia M. Cooper, Heidi Cope, William J. Craigen, Precilla D'Souza, Mariska Davids, Jean M. Davidson, Jyoti G. Dayal, Esteban C. Dell'Angelica, Shweta U. Dhar, Katrina M. Dipple, Laurel A. Donnell-Fink, Naghmeh Dorrani, Daniel C. Dorset, Emilie D. Douine, David D. Draper, Annika M. Dries, Laura Duncan, David J. Eckstein, Lisa T. Emrick, Christine M. Eng, Gregory M. Enns, Ascia Eskin, Cecilia Esteves, Tyra Estwick, Liliana Fernandez, Carlos Ferreira, Elizabeth L. Fieg, Paul G. Fisher, Brent L. Fogel, Noah D. Friedman, William A. Gahl, Emily Glanton, Rena A. Godfrey, Alica M. Goldman, David B. Goldstein, Sarah E. Gould, Jean Philippe F. Gourdine, Catherine A. Groden, Andrea L. Gropman, Melissa Haendel, Rizwan Hamid, Neil A. Hanchard, Frances High, Ingrid A. Holm, Jason Hom, Ellen M. Howerton, Yong Huang, Fariha Jamal, Yong hui Jiang, Jean M. Johnston, Angela L. Jones, Lefkothea Karaviti, David M. Koeller, Isaac S. Kohane, Jennefer N. Kohler, Donna M. Krasnewich, Susan Korrick, Mary Koziura, Joel B. Krier, Jennifer E. Kyle, Seema R. Lalani, C. Christopher Lau, Jozef Lazar, Kimberly LeBlanc, Brendan H. Lee, Hane Lee, Shawn E. Levy, Richard A. Lewis, Sharyn A. Lincoln, Sandra K. Loo, Joseph Loscalzo, Richard L. Maas, Ellen F. Macnamara, Calum A. MacRae, Valerie V. Maduro, Marta M. Majcherska, May Christine V. Malicdan, Laura A. Mamounas, Teri A. Manolio, Thomas C. Markello, Ronit Marom, Martin G. Martin, Julian A. Martínez-Agosto, Shruti Marwaha, Thomas May, Allyn McConkie-Rosell, Colleen E. McCormack, Alexa T. McCray, Jason D. Merker, Thomas O. Metz, Matthew Might, Paolo M. Moretti, Marie Morimoto, John J. Mulvihill, David R. Murdock, Jennifer L. Murphy, Donna M. Muzny, Michele E. Nehrebecky, Stan F. Nelson, J. Scott Newberry, John H. Newman, Sarah K. Nicholas, Donna Novacic, Jordan S. Orange, James P. Orengo, J. Carl Pallais, Christina Gs Palmer, Jeanette C. Papp, Neil H. Parker, Loren Dm Pena, John A. Phillips, Jennifer E. Posey, John H. Postlethwait, Lorraine Potocki, Barbara N. Pusey, Genecee Renteria, Chloe M. Reuter, Lynette Rives, Amy K. Robertson, Lance H. Rodan, Jill A. Rosenfeld, Jacinda B. Sampson, Susan L. Samson, Kelly Schoch, Daryl A. Scott, Lisa Shakachite, Prashant Sharma, Vandana Shashi, Rebecca Signer, Edwin K. Silverman, Janet S. Sinsheimer, Kevin S. Smith, Rebecca C. Spillmann, Joan M. Stoler, Nicholas Stong, Jennifer A. Sullivan, David A. Sweetser, Queenie K.G. Tan, Cynthia J. Tifft, Camilo Toro, Alyssa A. Tran, Tiina K. Urv, Eric Vilain, Tiphanie P. Vogel, Daryl M. Waggott, Colleen E. Wahl, Nicole M. Walley, Chris A. Walsh, Melissa Walker, Jijun Wan, Michael F. Wangler, Patricia A. Ward, Katrina M. Waters, Bobbie Jo M. Webb-Robertson, Monte Westerfield, Matthew T. Wheeler, Anastasia L. Wise, Lynne A. Wolfe, Elizabeth A. Worthey, Shinya Yamamoto, John Yang, Yaping Yang, Amanda J. Yoon, Guoyun Yu, Diane B. Zastrow, Chunli Zhao, Allison Zheng, Kym Boycott, Alex MacKenzie, Jacek Majewski, Michael Brudno, Dennis Bulman, David Dyment, Lars Lind, Erik Ingelsson, Alexis Battle, Gill Bejerano, Jonathan A. Bernstein, Euan A. Ashley, Kym M. Boycott, Jason D. Merker, Matthew T. Wheeler, and Stephen B. Montgomery. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nature Medicine*, 25:911–919, 2019.

[244] Secretary of State for Health and Social Care announces ambition to sequence 5 million genomes within five years | Genomics England. https://www.genomicsengland.co.uk/matt-hancock-announces-5-million-genomes-within-five-years/, accessed 2020-03-29.

[245] Diagnosis - The New York Times. https://www.nytimes.com/column/diagnosis, accessed 2020-03-29.

Appendix A

Appendices for chapter 2

A.1 cDNA fragment sequences and plasmid maps

Legend for sequences:

- Restriction enzyme site
- Bases added to increase restriction enzyme accuracy
- C'C: restriction enzyme cuts at the apostrophe
- Kozak sequence (RNNATGG)
- Coding sequence

A.1.1 *HDLBP* wildtype

C G G G G T A C ' C A T G A G T T C C G T T G C A G T T T T G A C C C A A
G A G A G T T T T G C T G A A C A C C G A A G T G G G C T G G T T C C G C
A A C A A A T C A A A G T T G C C A C T C T A A A T T C A G A A G A G G A
G A G C G A C C C T C C A A C C T A C A A G G A T G C C T T C C C T C C A
C T T C C T G A G A A A G C T G C T T G C C T G G A A A G T G C C C A G G
A A C C C G C T G G A G C C T G G G G G A A C A A G A T C C G A C C C A T
C A A G G C T T C T G T C A T C A C T C A G G T G T T C C A T G T A C C C C
T G G A G G A G A G A A A A T A C A A G G A T A T G A A C C A G T T T G G
A G A A G G T G A A C A A G C A A A A A A T C T G C C T T G A G A T C A T G
C A G A G A A C T G G T G C T C A C T T G G A G C T G T C T T T G G C C A
A A G A C C A A G G C C T C T C C A T C A T G G T G T C A G G A A A G C T
G G A T G C T G T C A T G A A A G C T C G G A A G G A C A T T G T T G C T
A G A C T G C A G A C T C A G G C C T C A G C A A C T G T T G C C A T T C
C C A A A G A A C A C C A T C G C T T T G T T A T T G G C A A A A A T G G
A G A G A A A C T G C A A G A C T T G G A G C T A A A A A C T G C A A C C
A A A A T C C A G A T C C C A C G C C C A G A T G A C C C C A G C A A T C
A G A T C A A G A T C A C T G G C A C C A A A G A G G G C A T C G A G A A
A G C T C G C C A T G A A G T C T T A C T C A T C T C T G C C G A G C A G
G A C A A A C G T G C T G T G G A G A G G C T A G A A G T A G A A A A G
G C A T T C C A C C C C T T C A T C G C T G G G C C G T A T A A T A G A C
T G G T T G G C G A G A T C A T G C A G G A G A C A G G C A C G C G C A T
C A A C A T C C C C C C A C C C A G C G T G A A C C G G A C A G A G A T T
G T C T T C A C T G G A G A G A A G G A A C A G T T G G C T C A G G C T G
T G G C T C G C A T C A A G A A G A T T T A T G A G G A G A A G A A A A A
G A A G A C T A C A A C C A T T G C A G T G G A A G T G A A G A A A T C C
C A A C A C A A G T A T G T C A T T G G G C C C A A G G G C A A T T C A T

TGCAGGAGATCCTTGAGAGAACTGGAGTTTCCGTTGA
GATCCCACCCTCAGACAGCATCTCTGAGACTGTAATA
CTTCGAGGCGAACCTGAAAAGTTAGGTCAGGCGTTGA
CTGAAGTCTATGCCAAGGCCAATAGCTTCACCGTCTC
CTCTGTCGCCGCCCCTTCCTGGCTTCACCGTTTCATCA
TTGGCAAGAAAGGGCAGAACCTGGCCAAAATCACTCA
GCAGATGCCAAAGGTTCACATCGAGTTCACAGAGGGC
GAAGACAAGATCACCCTGGAGGGCCCTACAGAGGATG
TCAATGTGGCCCAGGAACAGATAGAAGGCATGGTCAA
AGATTTGATTAACCGGATGGACTATGTGGAGATCAAC
ATCGACCACAAGTTCCACAGGCACCTCATTGGGAAGA
GCGGTGCCAACATAAACAGAATCAAAGACCAGTACAA
GGTGTCCGTGCGCATCCCTCCTGACAGTGAGAAGAGC
AATTTGATCCGCATCGAGGGGGACCCACAGGGCGTGC
AGCAGGCCAAGCGAGAGCTGCTGGAGCTTGCATCTCG
CATGGAAAATGAGCGTACCAAGGATCTAATCATTGAG
CAAAGATTTCATCGCACAATCATTGGGCAGAAGGGTG
AACGGATCCGTGAAATTCGTGACAAATTCCCAGAGGT
CATCATTAACTTTCCAGACCCAGCACAAAAAAGTGAC
ATTGTCCAGCTCAGAGGACCTAAGAATGAGGTGGAAA
AATGCACAAAATACATGCAGAAGATGGTGGCAGATCT
GGTGGAAAATAGCTATTCAATTTCTGTTCCGATCTTCA
AACAGTTTCACAAGAATATCATTGGGAAAGGAGGCGC
AAACATTAAAAAGATTCGTGAAGAAAGCAACACCAAA
ATCGACCTTCCAGCAGAGAATAGCAATTCAGAGACCA
TTATCATCACAGGCAAGCGAGCCAACTGCGAAGCTGC
CCGGAGCAGGATTCTGTCTATTCAGAAAGACCTGGCC
AACATAGCCGAGGTAGAGGTCTCCATCCCTGCCAAGC

TGCACAACTCCCTCATTGGCACCAAGGGCCGTCTGAT
CCGCTCCATCATGGAGGAGTGCGGCGGGGTCCACATT
CACTTTCCCGTGGAAGGTTCAGGAAGCGACACCGTTG
TTATCAGGGGCCCTTCCTCGGATGTGGAGAAGGCCAA
GAAGCAGCTCCTGCATCTGGCGGAGGAGAAGCAAACC
AAGAGTTTCACTGTTGACATCCGCGCCAAGCCAGAAT
ACCACAAATTCCTCATCGGCAAGGGGGGCGGCAAAAT
TCGCAAGGTGCGCGACAGCACTGGAGCACGTGTCATC
TTCCCTGCGGCTGAGGACAAGGACCAGGACCTGATCA
CCATCATTGGAAAGGAGGACGCCGTCCGAGAGGCACA
GAAGGAGCTGGAGGCCTTGATCCAAAACCTGGATAAT
GTGGTGGAAGACTCCATGCTGGTGGACCCCAAGCACC
ACCGCCACTTCGTCATCCGCAGAGGCCAGGTCTTGCG
GGAGATTGCTGAAGAGTATGGCGGGGTGATGGTCAGC
TTCCCACGCTCTGGCACACAGAGCGACAAAGTCACCC
TCAAGGGCGCCAAGGACTGTGTGGAGGCAGCCAAGAA
ACGCATTCAGGAGATCATTGAGGACCTGGAAGCTCAG
GTGACATTAGAATGTGCTATACCCCAGAAATTCCATC
GATCTGTCATGGGCCCCAAAGGTTCCAGAATCCAGCA
GATTACTCGGGATTTCAGTGTTCAAATTAAATTCCCA
GACAGAGAGGAGAACGCAGTTCACAGTACAGAGCCA
GTTGTCCAGGAGAATGGGGACGAAGCTGGGGAGGGG
AGAGAGGCTAAAGATTGTGACCCCGGCTCTCCAAGGA
GGTGTGACATCATCATCATCTCTGGCCGGAAAGAAAA
GTGTGAGGCTGCCAAGGAAGCTCTGGAGGCATTGGTT
CCTGTCACCATTGAAGTAGAGGTGCCCTTTGACCTTC
ACCGTTACGTTATTGGGCAGAAAGGAAGTGGGATCCG
CAAGATGATGGATGAGTTTGAGGTGAACATACATGTC

CCGGCACCTGAGCTGCAGTCTGACATCATCGCCATCA

CGGGCCTCGCTGCAAATTTGGACCGGGCCAAGGCTGG

ACTGCTGGAGCGTGTGAAGGAGCTACAGGCCGAGCAG

GAGGACCGGGCTTTAAGGAGTTTTAAGCTGAGTGTCA

CTGTAGACCCCAAATACCATCCCAAGATTATCGGGAG

AAAGGGGGCAGTAATTACCCAAATCCGGTTGGAGCAT

GACGTGAACATCCAGTTTCCTGATAAGGACGATGGGA

ACCAGCCCCAGGACCAAATTACCATCACAGGGTACGA

AAAGAACACAGAAGCTGCCAGGGATGCTATACTGAGA

ATTGTGGGTGAACTTGAGCAGATGGTTTCTGAGGACG

TCCCGCTGGACCACCGCGTTCACGCCCGCATCATTGG

TGCCCGCGGCAAAGCCATTCGCAAAATCATGGACGAA

TTCAAGGTGGACATTCGCTTCCCACAGAGCGGAGCCC

CAGACCCCAACTGCGTCACTGTGACGGGGCTCCCAGA

GAATGTGGAGGAAGCCATCGACCACATCCTCAATCTG

GAGGAGGAATACCTAGCTGACGTGGTGGACAGTGAGG

CGCTGCAGGTATACATGAAACCCCCAGCACACGAAGA

GGCCAAGGCACCTTCCAGAGGCTTTGTGGTGCGGGAC

GCACCCTGGACCGCCAGCAGCAGTGAGAAGGCTCCTG

ACATGAGCAGCTCTGAGGAATTTCCCAGCTTTGGGGC

TCAGGTGGCTCCCAAGACCCTCCCTTGGGGCCCCAAA

CGAC'TCGAGCGG

A.1.2 *HDLBP* mutant

C G G G G T A C ' C A T G A G T T C C G T T G C A G T T T T G A C C C A A
G A G A G T T T T G C T G A A C A C C G A A G T G G G C T G G T T C C G C
A A C A A A T C A A A G T T G C C A C T C T A A A T T C A G A A G A G G A
G A G C G A C C C T C C A A C C T A C A A G G A T G C C T T C C C T C C A
C T T C C T G A G A A A G C T G C T T G C C T G G A A A G T G C C C A G G
A A C C C G C T G G A G C C T G G G G G A A C A A G A T C C G A C C C A T
C A A G G C T T C T G T C A T C A C T C A G G T G T T C C A T G T A C C C C
T G G A G G A G A G A A A A T A C A A G G A T A T G A A C C A G T T T G G
A G A A G G T G A A C A A G C A A A A A A T C T G C C T T G A G A T C A T G
C A G A G A A C T G G T G C T C A C T T G G A G C T G T C T T T G G C C A
A A G A C C A A G G C C T C T C C A T C A T G G T G T C A G G A A A G C T
G G A T G C T G T C A T G A A A G C T C G G A A G G A C A T T G T T G C T
A G A C T G C A G A C T C A G G C C T C A G C A A C T G T T G C C A T T C
C C A A A G A A C A C C A T C G C T T T G T T A T T G G C A A A A A T G G
A G A G A A A C T G C A A G A C T T G G A G C T A A A A A C T G C A A C C
A A A A T C C A G A T C C C A C G C C C A G A T G A C C C C A G C A A T C
A G A T C A A G A T C A C T G G C A C C A A A G A G G G C A T C G A G A A
A G C T C G C C A T G A A G T C T T A C T C A T C T C T G C C G A G C A G
G A C A A A C G T G C T G T G G A G A G G C T A G A A G T A G A A A A G
G C A T T C C A C C C C T T C A T C G C T G G G C C G T A T A A T A G A C
T G G T T G G C G A G A T C A T G C A G G A G A C A G G C A C G C G C A T
C A A C A T C C C C C C A C C C A G C G T G A A C C G G A C A G A G A T T
G T C T T C A C T G G A G A G A A G G A A C A G T T G G C T C A G G C T G
T G G C T C G C A T C A A G A A G A T T T A T G A G G A G A A G A A A A A
G A A G A C T A C A A C C A T T G C A G T G G A A G T G A A G A A A T C C
C A A C A C A A G T A T G T C A T T G G G C C C A A G G G C A A T T C A T

TGCAGGAGATCCTTGAGAGAACTGGAGTTTCCGTTGA
GATCCCACCCTCAGACAGCATCTCTGAGACTGTAATA
CTTCGAGGCGAACCTGAAAAGTTAGGTCAGGCGTTGA
CTGAAGTCTATGCCAAGGCCAATAGCTTCACCGTCTC
CTCTGTCGCCGCCCCTTCCTGGCTTCACCGTTTCATCA
TTGGCAAGAAAGGGCAGAACCTGGCCAAAATCACTCA
GCAGATGCCAAAGGTTCACATCGAGTTCACAGAGGGC
GAAGACAAGATCACCCTGGAGGGCCCTACAGAGGATG
TCAATGTGGCCCAGGAACAGATAGAAGGCATGGTCAA
AGATTTGATTAACCGGATGGACTATGTGGAGATCAAC
ATCGACCACAAGTTCCACAGGCACCTCATTGGGAAGA
GCGGTGCCAACATAAACAGAATCAAAGACCAGTACAA
GGTGTCCGTGCGCATCCCTCCTGACAGTGAGAAGAGC
AATTTGATCCGCATCGAGGGGGACCCACAGGGCGTGC
AGCAGGCCAAGCGAGAGCTGCTGGAGCTTGCATCTCG
CATGGAAAATGAGCGTACCAAGGATCTAATCATTGAG
CAAAGATTTCATCGCACAATCATTGGGCAGAAGGGTG
AACGGATCCGTGAAATTCGTGACAAATTCCCAGAGGT
GGAAAATAGCTATTCAATTTCTGTTCCGATCTTCAAA
CAGTTTCACAAGAATATCATTGGGAAAGGAGGCGCAA
ACATTAAAAAGATTCGTGAAGAAAGCAACACCAAAAT
CGACCTTCCAGCAGAGAATAGCAATTCAGAGACCATT
ATCATCACAGGCAAGCGAGCCAACTGCGAAGCTGCCC
GGAGCAGGATTCTGTCTATTCAGAAAGACCTGGCCAA
CATAGCCGAGGTAGAGGTCTCCATCCCTGCCAAGCTG
CACAACTCCCTCATTGGCACCAAGGGCCGTCTGATCC
GCTCCATCATGGAGGAGTGCGGCGGGGTCCACATTCA
CTTTCCCGTGGAAGGTTCAGGAAGCGACACCGTTGTT

A T C A G G G G C C C T T C C T C G G A T G T G G A G A A G G C C A A G A
A G C A G C T C C T G C A T C T G G C G G A G G A G A A G C A A A C C A A
G A G T T T C A C T G T T G A C A T C C G C G C C A A G C C A G A A T A C
C A C A A A T T C C T C A T C G G C A A G G G G G G C G G C A A A A T T C
G C A A G G T G C G C G A C A G C A C T G G A G C A C G T G T C A T C T T
C C C T G C G G C T G A G G A C A A G G A C C A G G A C C T G A T C A C C
A T C A T T G G A A A G G A G G A C G C C G T C C G A G A G G C A C A G A
A G G A G C T G G A G G C C T T G A T C C A A A A C C T G G A T A A T G T
G G T G G A A G A C T C C A T G C T G G T G G A C C C C A A G C A C C A C
C G C C A C T T C G T C A T C C G C A G A G G C C A G G T C T T G C G G G
A G A T T G C T G A A G A G T A T G G C G G G G T G A T G G T C A G C T T
C C C A C G C T C T G G C A C A C A G A G C G A C A A A G T C A C C C T C
A A G G G C G C C A A G G A C T G T G T G G A G G C A G C C A A G A A A
C G C A T T C A G G A G A T C A T T G A G G A C C T G G A A G C T C A G G
T G A C A T T A G A A T G T G C T A T A C C C C A G A A A T T C C A T C G
A T C T G T C A T G G G C C C C A A A G G T T C C A G A A T C C A G C A G
A T T A C T C G G G A T T T C A G T G T T C A A A T T A A A T T C C C A G A
C A G A G A G G A G A A C G C A G T T C A C A G T A C A G A G C C A G T T
G T C C A G G A G A A T G G G G A C G A A G C T G G G G A G G G G A G A
G A G G C T A A A G A T T G T G A C C C C G G C T C T C C A A G G A G G T
G T G A C A T C A T C A T C A T C T C T G G C C G G A A A G A A A A G T G
T G A G G C T G C C A A G G A A G C T C T G G A G G C A T T G G T T C C T
G T C A C C A T T G A A G T A G A G G T G C C C T T T G A C C T T C A C C
G T T A C G T T A T T G G G C A G A A A G G A A G T G G G A T C C G C A A
G A T G A T G G A T G A G T T T G A G G T G A A C A T A C A T G T C C C G
G C A C C T G A G C T G C A G T C T G A C A T C A T C G C C A T C A C G G
G C C T C G C T G C A A A T T T G G A C C G G G C C A A G G C T G G A C T
G C T G G A G C G T G T G A A G G A G C T A C A G G C C G A G C A G G A

G G A C C G G G C T T T A A G G A G T T T T A A G C T G A G T G T C A C T

G T A G A C C C C A A A T A C C A T C C C A A G A T T A T C G G G A G A A

A G G G G G C A G T A A T T A C C C A A A T C C G G T T G G A G C A T G A

C G T G A A C A T C C A G T T T C C T G A T A A G G A C G A T G G G A A C

C A G C C C C A G G A C C A A A T T A C C A T C A C A G G G T A C G A A A

A G A A C A C A G A A G C T G C C A G G G A T G C T A T A C T G A G A A T

T G T G G G T G A A C T T G A G C A G A T G G T T T C T G A G G A C G T C

C C G C T G G A C C A C C G C G T T C A C G C C C G C A T C A T T G G T G

C C C G C G G C A A A G C C A T T C G C A A A A T C A T G G A C G A A T T

C A A G G T G G A C A T T C G C T T C C C A C A G A G C G G A G C C C C A

G A C C C C A A C T G C G T C A C T G T G A C G G G G C T C C C A G A G A

A T G T G G A G G A A G C C A T C G A C C A C A T C C T C A A T C T G G A

G G A G G A A T A C C T A G C T G A C G T G G T G G A C A G T G A G G C G

C T G C A G G T A T A C A T G A A A C C C C C A G C A C A C G A A G A G G

C C A A G G C A C C T T C C A G A G G C T T T G T G G T G C G G G A C G C

A C C C T G G A C C G C C A G C A G C A G T G A G A A G G C T C C T G A C

A T G A G C A G C T C T G A G G A A T T T C C C A G C T T T G G G G C T C

A G G T G G C T C C C A A G A C C C T C C C T T G G G G C C C C A A A C G

A C ’ T C G A G C G G

Appendix B

Appendices for chapter 3

All data showed for VAAST+Phevor was created with Omicia Opal version 4.24.0, the same version of the software used for the analyses in Chapter 1, since I was not able to access an older version of the platform using the same databases as Exomiser v7.2.1. Therefore, data showed in Figure B.1 for Exomiser in the following plots should not be interpreted in comparison to the VAAST+Phevor data.

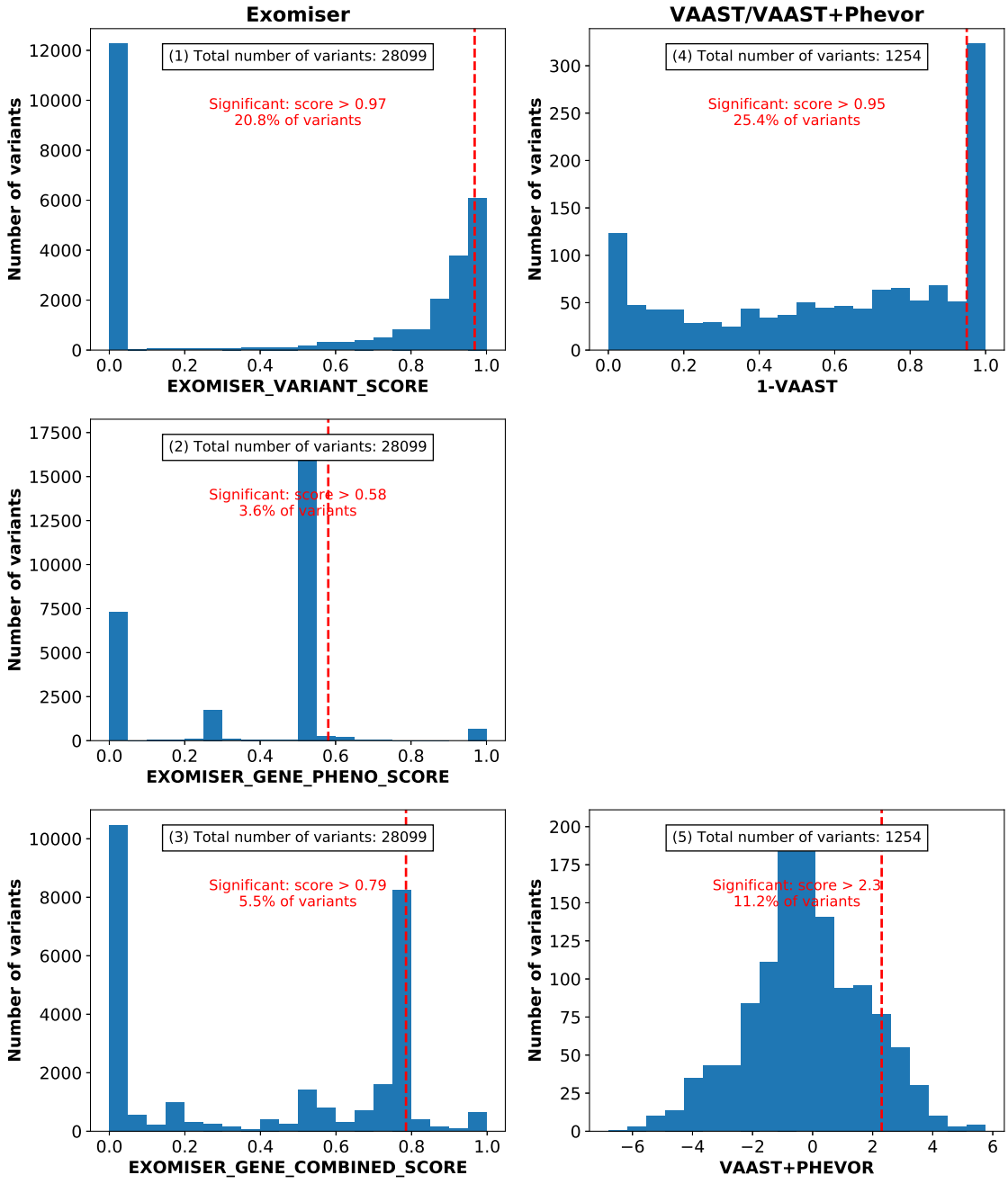


Fig. B.1 Histogram of prioritisation scores, combining all variants for all benchmark cases for each algorithm. The red dotted light shows the significance cut-off for each algorithm: Exomiser variant score $\geq$ 1.00, Exomiser phenotype score $\geq$ 0.58, Exomiser combined score $\geq$ 0.79, 1-VAAST(p-value) $\geq$ 0.95 and Phevor score $\geq$ 2.3. Based on Exomiser v7.2.1