TUM School of Life Sciences

Discovery of genetic causes of rare diseases through analysis of clinical sequencing data

Tim Jeske

2022

TUM School of Life Sciences

TECHNISCHE UNIVERSITÄT MÜNCHEN

Discovery of genetic causes of rare diseases through analysis of clinical sequencing data

Tim Jeske

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Dr. h.c. mult. Martin Hrabě de Angelis
Prüfer der Dissertation:	1. Prof. Dr. Hans-Werner Mewes 2. Prof. Dr. Dmitrij Frischmann 3. Prof. Dr. Inke König

Die Dissertation wurde am 23.06.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 14.03.2022 angenommen.

Danksagung

Zuallererst möchte ich mich bei meinem Doktorvater Werner Mewes bedanken. Seine langjährige Betreuung von meinem Bachelorabschluss, über meinen Masterabschluss bis hin zur Dissertation hat meine bisherige wissenschaftlichen Laufbahn maßgeblich geprägt und erst ermöglicht. Auch an Christoph Klein möchte ich meinen besonderen Dank richten, einerseits für die Bereitstellungen der dieser Dissertation zugrundeliegenden Daten, andererseits für das stets in mich entgegengebrachte Vertrauen in meine Fähigkeiten.

Desweiteren gilt mein Dank Allen, die meine Betreuung zu verschiedenen Zeiten des Projekts übernahmen. Dazu zählen Gabi Kastenmüller, die mich durchgehend als Teil ihrer Gruppe am Helmholtz Zentrum München betrachtete und nach Kräften unterstützte, Rober Küffner, der mir half die Kontakte zum Dr. von Haunerschen Kinderklinikum zu knüpfen und auch später unabhängig von seinem Aufenthaltsort Ideen und Ergebnisse mit mir diskutierte, sowie Matthias Heinig, dessen Anregungen als externer Experte meines Thesis Committees meine Arbeit mit formten. Schließlich haben die Diskussionen meiner Ergebnisse mit Meino Rohlfs und Daniel Kotlarz entscheidend die Entwicklung der beschriebenen Tools und Pipelines vorangetrieben und mich stets angespornt deren Qualität zu optimieren.

Auch möchte ich mich bei "meinen" Studenten bedanken, die zu einem nicht unerheblichen Teil zu dieser Thesis beigetragen haben. Das umfassende biologische Hintergrundwissen von Kaarin Ahomaa gab einen entscheidenden Hinweis zur Identifikation der kausalen genetischen Variante einer Patientin und der darauf folgenden Entdeckung einer neuartigen Pathogenese von Neutropenie. Die Projekte von Verena Burger und Susanne Artmeier waren die Ideengeber für die hier beschriebenen Schritte der Qualitätskontrolle und Auswertung der Daten. Und schließlich möchte ich mich bei Paul Hager bedanken, ohne dessen unermüdlichen Einsatzes die Implementierung und Publikation von Smart-Phase unmöglich gewesen wäre.

Von allen Kollegen, die mich in mich während meiner Promotion begleiteten, möchte ich mich vor allem bei Maximilian Hastreiter bedanken. Als Bürokollegen schon seit meiner Masterarbeit sind wir als Team zusammengewachsen und konnten diverse Projekte nicht nur erfolgreich, sondern auch mit Freude daran abschließen. Darüber hinaus trugen auch Maria Wörheide und Patrick Weinisch als Zimmerkollegen am Helmholtz Zentrum und Sebastian Hollizeck und Maximilian Witzel als Zimmerkollegen am Klinikum zu einer Arbeitsatmosphäre bei, die jeden Tag meiner Zeit als Doktorand lohnenswert machten.

Abstract

Millions of people worldwide suffer from rare diseases. Genetic diagnosis is crucial to identify the molecular cause of disease, adapt the treatment, and decipher the pathogenesis. Whole-exome sequencing (WES) became established as a diagnostic tool in clinical practice, but the majority of rare disease patients remains undiagnosed. As each individual has several thousands of genetic variants in his or her genome, the identification of disease-causing variants is challenging. The aim of my thesis was to improve the analysis of clinical sequencing data, identify genetic causes of rare diseases and gain new insights into the pathomechanisms underlying rare diseases.

In the first part of my work, I developed the tools KNIME4NGS and SmartPhase as well as an algorithm for population stratification and a Candidate Identification Pipeline (CIP). KNIME4NGS simplifies the processing of sequencing data by allowing the user to assemble modular workflows in a graphical user interface. SmartPhase, a tool for accurate and fast phasing of heterozygous variant pairs, improves the detection of compound heterozygosity in clinical pipelines. The implemented algorithm that stratifies a cohort by ethnic origin enables the detection and filtration of population-specific variants, which appear rare overall, but are common in underrepresented populations. Finally, the CIP is a flexible workflow to prioritize genetic variants and to select candidate genes for disease associations based on a multitude of annotations and derived filter criteria.

In the second part, I applied the developed tools and methods to clinical sequencing data from a cohort of pediatric patients suffering from various inborn errors of immunity (IEI). The data set comprises WES data of $1,746$ patients and 705 healthy relatives as well as whole-genome sequencing data of one family. After applying the stratification algorithm to the cohort, filtering by population-specific allele frequencies increased the number of frequency-filtered variants by 23.68%. The use of SmartPhase contributed to a higher efficiency of variant filtering by reducing the number of potentially compound heterozygous variant pairs by 59.16%. The final variant prioritization and candidate gene selection using the CIP resulted in the identification of 205 genes potentially causing IEI.

Further, I thoroughly characterized genetic etiologies of IEI in the analyzed cohort based on 26 selected confirmed pathogenic variants in ten different genes affecting 33 patients. Of the latter, 20 patients suffer from inflammatory bowel disease (IBD). In seven of these, IBD is the manifestation of underlying IEI, while the other 13 patients have defects in the genes *EPCAM* and *SLC5A1*, both not linked to IEI. This finding shows that IBD can be an indication of IEI, but also variants in genes not related to IEI can cause IBD-like symptoms. Additionally, the symptoms observed in the seven patients with underlying IEI enabled an extension of the phenotypic description of defects in the genes *CARMIL2*, *FOXP3*, *G6PC3*, *SRP54* and *RTEL1*. Of two intronic variants among the 26 pathogenic variants, one has proven to be a branch point mutation resulting in a splicing defect of *DKC1* and causing dyskeratosis congenita, which serves as an example of non-coding variation underlying rare diseases. While eight of the ten discussed genes have already been associated with monogenic diseases, I describe the discovery of two novel defects in the genes *SRPRA* and *SRP19* causing severe congenital neutropenia.

In conclusion, the developed tools and methods improved the prioritization of genetic variants, led to the identification of pathogenic variants in coding as well as non-coding regions, and resulted in the discovery of novel genetic defects. As none of the analyses presented is methodologically restricted to a specific set of genes, they can all be applied to any clinical cohort with suspected monogenic causalities. Besides having enabled a definitive diagnosis in a substantial number of children in the cohort analyzed, the developed tools and methods as well as the discovered gene defects will help to increase the diagnostic rate in future rare disease studies.

Zusammenfassung

Millionen Menschen weltweit leiden an seltenen Krankheiten. Die genetische Diagnostik ist entscheidend, um die molekulare Ursache der jeweiligen Erkrankung zu identifizieren, die Behandlung anzupassen und die Pathogenese aufzuklären. Hierzu hat sich Exomsequenzierung in der klinischen Praxis als diagnostisches Werkzeug etabliert, aber die Mehrheit der Patienten mit seltenen Krankheiten kann trotzdem noch nicht mit Gewissheit diagnostiziert werden. Da jeder Mensch mehrere tausend genetische Varianten in seinem Genom hat, ist die Identifizierung von krankheitsverursachenden Varianten eine Herausforderung. Das Ziel meiner Dissertation war es, die Analyse von klinischen Sequenzdaten zu verbessern, genetische Ursachen seltener Krankheiten zu identifizieren und neue Erkenntnisse zu den Pathomechanismen zu gewinnen, die seltenen Krankheiten zugrunde liegen.

Im ersten Teil meiner Arbeit habe ich die Programme KNIME4NGS und SmartPhase, sowie einen Algorithmus zur Populationsstratifikation und eine Kandidaten-Identifikations-Pipeline entwickelt. KNIME4NGS vereinfacht die Prozessierung von Sequenzdaten indem es dem Anwender die Modellierung von Analyseschritten in einer grafischen Oberfläche ermöglicht. SmartPhase, ein Programm für genaues und schnelles Phasen von heterozygoten Variantenpaaren, verbessert die Erkennung von kombinierter Heterozygosität bei der Analyse klinischer Daten. Der implementierte Algorithmus, der eine Kohorte nach ethnischer Herkunft stratifiziert, erlaubt die Identifikation und Filterung von populationsspezifischen Varianten, die insgesamt selten erscheinen, aber in unterrepräsentierten Populationen häufig auftreten. Schließlich stellt die Kandidaten-Identifikations-Pipeline einen flexibel anpassbaren Ablauf von Analyseschritten dar, um auf Basis einer Vielzahl von Annotationen und abgeleiteten Filterkriterien genetische Varianten zu priorisieren und Kandidatengene für Krankheitsassoziationen zu selektieren.

Im zweiten Teil wendete ich die entwickelten Programme und Methoden auf eine klinische Kohorte pädiatrischer Patienten an, die an verschiedenen angeborenen Immunstörungen leiden. Der Datensatz umfasst Exomsequenzdaten von 1.746 Patienten und 705 gesunden Verwandten sowie Genomsequenzdaten einer Familie. Nach Anwendung des Stratifikationsalgorithmus auf die Kohorte erhöhte das Filtern nach populationsspezifischen Allelhäufigkeiten die Menge der nach Häufigkeit gefilterten Varianten um $23,68\%$. Die Verwendung von SmartPhase trug zu einer höheren Effizienz der Variantenfilterung bei, indem es die Menge der potentiell kombiniert heterozygoten Variantenpaare um $59,16\%$

reduzierte. Die abschließende Priorisierung der Varianten und Selektion von Kandidatengenen resultierte in der Identifikation von 205 Genen, die potentiell angeborene Immunstörungen auslösen.

Weiterhin habe ich genetische Ätiologien angeborener Immunstörungen in der analysierten Kohorte umfassend charakterisiert, basierend auf 26 ausgewählten nachgewiesen pathogenen Varianten in zehn verschiedenen Genen, die 33 Patienten betreffen. Von Letzteren leiden 20 Patienten an chronisch-entzündlichen Darmerkrankungen. Bei sieben von diesen ist die Darmerkrankung eine Manifestation zugrundeliegender angeborener Immunstörungen, während die anderen 13 Patienten Defekte in den Genen *EPCAM* und *SLC5A1* haben, die beide nicht mit angeborenen Immunstörungen in Verbindung stehen. Diese Erkenntnis zeigt, dass chronisch-entzündliche Darmerkrankungen zwar ein Hinweis auf angeborene Immunstörungen sein können, aber auch Varianten in Genen ohne Bezug zu angeborenen Immunstörungen Symptome chronisch-entzündlicher Darmerkrankungen verursachen können. Zusätzlich ermöglichten die Symptome, die bei den sieben Patienten beobachtet wurden, die Erweiterung der klinischen Beschreibung von Defekten in den Genen *CARMIL2*, *FOXP3*, *G6PC3*, *SRP54* und *RTEL1*. Von zwei intronischen Varianten unter den 26 pathogenen Varianten hat sich eine als eine sogenannte "branch point" Mutation erwiesen, die zu defektem Spleißen von *DKC1* führt und Dyskeratosis congenita verursacht, was ein Beispiel für nicht-kodierende Variation als Ursache seltener Erkrankungen darstellt. Während acht der zehn diskutierten Gene bereits mit monogenetischen Erkrankungen in Verbindung gebracht wurden, beschreibe ich die Entdeckung von zwei neuartigen Defekten in den Genen *SRPRA* und *SRP19*, die schwere kongenitale Neutropenie verursachen.

Zusammengefasst verbesserten die entwickelten Programme und Methoden die Priorisierung genetischer Varianten, führten zur Identifizierung pathogener Varianten sowohl in kodierenden als auch in nicht-kodierenden Regionen und resultierten in der Entdeckung neuartiger Gendefekte. Da sich keine der vorgestellten Analysen methodisch auf eine bestimmte Gruppe von Genen beschränkt, können sie alle auf jede klinische Kohorte mit vermuteten monogenen Kausalitäten angewandt werden. Neben der Ermöglichung einer definitiven Diagnose bei einer wesentlichen Anzahl von Kindern in der hier analysierten Kohorte, werden sowohl die entwickelten Programme und Methoden als auch die entdeckten Gendefekte dazu beitragen, die Diagnoserate in zukünftigen Studien zu seltenen Krankheiten zu erhöhen.

Scientific contributions

The following list includes my scientific contributions in the fields of next-generation sequencing data analysis, genetics of rare disease, and immunodeficiencies, as peer-reviewed publications or posters at conferences sorted in descending order by publication date.

- L. A. Schuch, M. Forstner, C. K. Rapp, Y. Li, D. E. C. Smith, M. I. Mendes, F. Delhommel, M. Sattler, N. Emiralioğlu, E. Z. Taskiran, D. Orhan, N. Kiper, M. Rohlfs, T. Jeske, M. Hastreiter, M. Gerstlauer, A. Torrent-Vernetta, A. Moreno-Galdó, B. Kammer, F. Brasch, S. Reu-Hofer, M. Griese, **"FARS1-related disorders caused by biallelic mutations in cytosolic phenylalanyl-tRNA synthetase genes: Look beyond the lungs!,"**, *Clin. Genet.*, vol. 99, pp. 789-801, June 2021.

- F. Rahmani, E. Rayzan, M. R. Rahmani, S. Shahkarami, S. Zoghi, A. Rezaei, Z. Aryan, M. Najafi, M. Rohlfs, T. Jeske, M. Aflatoonian, Z. Chavoshzadeh, F. Farahmand, F. Motamed, P. Rohani, H. Alimadadi, A. Mahdaviani, M. Mansouri, M. Tavakol, M. Vanderberg, D. Kotlarz, C. Klein, and N. Rezaei, **"Clinical and Mutation Description of the First Iranian Cohort of Infantile Inflammatory Bowel Disease: The Iranian Primary Immunodeficiency Registry (IPIDR),"** *Immunol. Invest.*, vol. 50, pp. 445–459, May 2021.

- D. T. Duztas, L. Al-Shadfan, H. Ozturk, H. Yazan, E. Cakir, N. Unver, O. Ekinci, B. Dalgic, M. Rohlfs, T. Jeske, C. Klein, D. Kotlarz, and O. E. Gurkan, **"New Findings of Immunodysregulation, Polyendocrinopathy, and Enteropathy X-linked Syndrome (IPEX); Granulomas in Lung and Duodenum,"**, *Pediatr. Dev. Pathol.*, Online ahead of print, Mar. 2021.

- S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower, T. J. Callahan, C. G. Chute, J. L. Est, P. D. Galer, S. Ganesan, M. Griese, M. Haimel, J. Pazmandi, M. Hanauer, N. L. Harris, M. J. Hartnett, M. Hastreiter, F. Hauck, Y. He, T. Jeske, H. Kearney, G. Kindle, C. Klein, K. Knoflach, R. Krause, D. Lagorce, J. A. McMurry, J. A. Miller, M. C. Munoz-Torres, R. L. Peters, C. K. Rapp, A. M. Rath, S. A. Rind, A. Z. Rosenberg, M. M. Segal, M. G. Seidel, D. Smedley, T. Talmy, Y. Thomas, S. A. Wiafe, J. Xian, Z. Yüksel, I. Helbig, C. J. Mungall, M. A. Haendel, and P. N.

Robinson, **“The Human Phenotype Ontology in 2021,”** *Nucleic Acids Res.*, vol. 49, pp. D1207-D1217, Jan. 2021.

- A. Ziv, L. Werner, L. Konnikova, A. Awad, T. Jeske, M. Hastreiter, V. Mitsialis, T. Stauber, S. Wall, D. Kotlarz, C. Klein, S. B. Snapper, Y. Tzfati, B. Weiss, R. Somech, and D. S. Shouval, **“An RTEL1 Mutation Links to Infantile-Onset Ulcerative Colitis and Severe Immunodeficiency,”** *J. Clin. Immunol.*, vol. 40, pp. 1010–1019, Oct. 2020.

- T. Jeske, M. Hastreiter, S. Hollizeck, M. Rohlfs, D. Kotlarz, and C. Klein, **“Deciphering genetic causes of inborn errors of immunity in a clinical setting,”** Poster session at the virtual conference *Genomics of Rare Disease* in Hinxton, United Kingdom, Mar. 2020.

- P. Hager, H.-W. Mewes, M. Rohlfs, C. Klein, and T. Jeske, **“SmartPhase: Accurate and fast phasing of heterozygous variant pairs for genetic diagnosis of rare diseases,”** *PLoS Comput. Biol.*, vol. 16, p. e1007613, Feb. 2020.

- Y. Mizoguchi, S. Hesse, M. Linder, N. Ziętara, M. Łyszkiewicz, Y. Liu, M. Tatematsu, P. Grabowski, K. Ahomaa, T. Jeske, S. Hollizeck, E. Rusha, M. K. Saito, M. Kobayashi, Z. Alizadeh, Z. Pourpak, S. Iurian, N. Rezaei, E. Unal, M. Drukker, B. Walzog, F. Hauck, J. Rappsilber, and C. Klein, **“Defects in Signal Recognition Particle (SRP) Components Reveal an Essential and Non-Redundant Role for Granule Biogenesis and Differentiation of Neutrophil Granulocytes,”** *Blood*, vol. 134 (Supplement_1), p. 216, Nov. 2019.

- T. Jeske, P. Huypens, L. Stirm, S. Höckele, C. M. Wurmser, A. Böhm, C. Weigert, H. Staiger, C. Klein, J. Beckers, and M. Hastreiter, **“DEUS: an R package for accurate small RNA profiling based on differential expression of unique sequences,”** *Bioinformatics*, vol. 35, pp. 4834–4836, Nov. 2019.

- T. Magg, A. Shcherbina, D. Arslan, M. M. Desai, S. Wall, V. Mitsialis, R. Conca, E. Unal, N. Karacabey, A. Mukhina, Y. Rodina, P. D. Taur, D. Illig, B. Marquardt, S. Hollizeck, T. Jeske, F. Gothe, T. Schober, M. Rohlfs, S. Koletzko, E. Lurz, A. M. Muise, S. B. Snapper, F. Hauck, C. Klein, and D. Kotlarz, **“CARMIL2 Deficiency Presenting as Very Early Onset Inflammatory Bowel Disease,”** *Inflamm. Bowel Dis.*, vol. 25, pp. 1788–1795, Oct. 2019.

- M. Hastreiter, T. Jeske, J. Hoser, M. Kluge, K. Ahomaa, M.-S. Friedl, S. J. Kopetzky, J. D. Quell, H.-W. Mewes, and R. Küffner, **“KNIME4NGS: A comprehensive toolbox for next generation sequencing analysis,”** *Bioinformatics*, vol. 33, pp. 1565–1567, May 2017.

Contents

Introduction 1

Although rare diseases affect only a limited number of patients individually, millions of people worldwide suffer from them as they encompass thousands of different disorders [1]. Thus, rare diseases represent a major global health issue that is easily missed when looking at statistics on the overall global disease burden that are headed by cancer, cardiovascular and infectious diseases [2]. The wide diversity makes differential diagnosis of rare diseases challenging and causal therapeutic options are often lacking. As genetic defects are the predominant cause of rare diseases and costs of sequencing have decreased considerably, exome sequencing became a standard diagnostic tool in clinical practice since its first successful application in 2011 [3]. Despite new insights generated through exome sequencing studies, the fraction of diagnosed patients is currently limited at approximately 40% [4, 5]. Apart from technical constraints, this is due to the fact that each patient carries thousands of genetic variants that need to be examined for their clinical relevance. To meet this challenge, bioinformatics software is needed to simplify the analysis of sequencing data, making full use of the available data and taking advantage of the growing volume of information on causal mutations and their corresponding phenotypes. Before presenting my contribution to this overarching challenge, I give a brief overview of related parts of human genetics, rare diseases in general and immunological diseases in particular, and the methods used to link these areas.

1.1 Genetic and phenotypic variation

The discovery of inheritance laws by Gregor Mendel in the 1860s was the first scientific milestone to understand how the genome determines the phenotype of an organism. However, Mendel could not be aware that the genome is the major building plan of organisms [6]. After the detection that chromosomes are the carriers of inherited information and that they are made of deoxyribonucleic acid (DNA), Francis Crick formulated the "central dogma" of molecular biology almost 100 years after Mendel's work [7]. It describes the flow of sequence information between the macromolecules DNA, ribonucleic acid (RNA), and proteins [8]. The sequence of DNA molecules can either be copied to another DNA molecule via replication, or transmitted to RNA molecules via transcription. The information in the nucleotide sequence of RNA molecules can then be translated to the amino acid sequence of proteins, but there is no biological pathway that transfers the sequence information of a protein to any of the three macromolecules. As proteins directly determine the phenotypic traits through their multiple functions as building blocks of the cell, catalysts of biological reactions, means of communication and transport, the "central dogma" established the path from DNA to RNA to protein as the major mechanism how phenotypic traits are influenced by the genome. Thus, the detection and interpretation of individual genetic variation that alters the sequence of the human genome is the key to understand how and to what extend the genome determines our personal phenotype in health and disease.

1.1.1 Types of genetic variation

The analysis of genetic variation in humans is a challenge because the more than 3 billion base pairs of each individual's genome cannot be read out as easily as a person's height or eye color, for example. Only recently it became technically possible to sequence human chromosomes as a whole [9]. So far and also in the foreseeable future, it is necessary to either assemble genomes from sequenced DNA fragments *de novo*, or to align sequenced fragments to an existing reference genome. *De novo* assemblies are computationally expensive and more complex to compare with each other as opposed to the comparison of alignments to the same reference. As genetic variation must be collected from hundreds and thousands of individuals to learn how phenotypic traits are influenced by the genome, *de novo* assemblies are infeasible in large scale. Hence, genetic variation in an individual is defined as the set of deviations from the reference genome and each deviation corresponds to a genetic variant.

The currently used versions of the human reference genome, Genome Reference Consortium Human Build 37 (GRCh37) and the more recent GRCh38, are string representations of the genomes derived from 20 individuals that were randomly sampled from the readers of the *Buffalo News*, a newspaper in Buffalo, New York, USA in 1997 [10]. The string-based representation allows the inclusion of variable regions in the human genome by providing alternate sequences for certain regions, but the majority of frequent genetic variation cannot be reflected. Furthermore, variation only prevalent in populations that are not covered by the 20 individuals who donated DNA is entirely absent in the reference genome. As a consequence of these limitations, existing collections of human genetic variants are used to create a human pan-genome that contains all observed variation in graph-based models, which should serve as a basis for future studies of human genetic variation [10].

Human genomes vary in many ways ranging from differences in the number of chromosomes to distinct single nucleotides. Genetic variants are classified according to their size and according to the way in which the genome is different compared to the reference genome, as shown in Figure 1.1. Substitutions of individual nucleotides, so called single nucleotide variants (SNVs), belong to small variants together with insertions or deletions

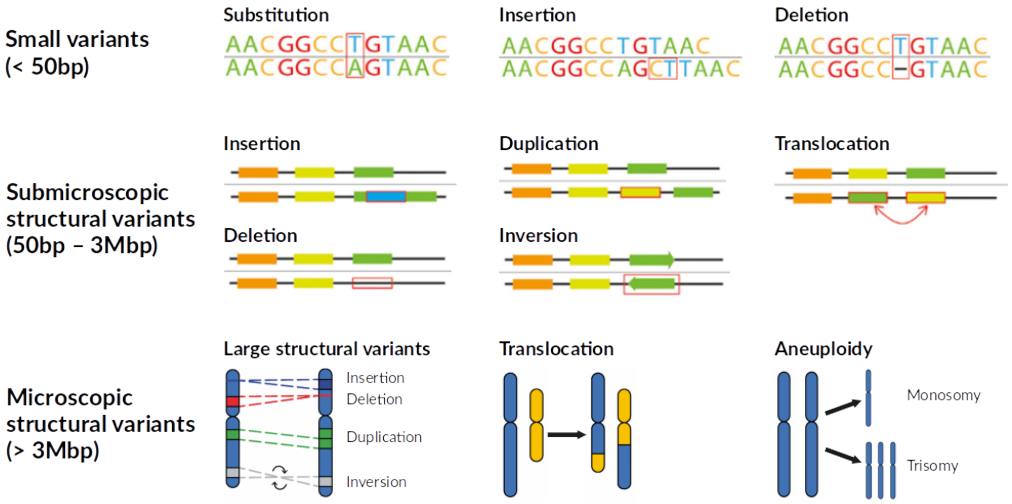


Figure 1.1: Main types of genetic variation. Illustrations of small variants and submicroscopic structural variants are adapted from [11]. For each variant type, the wild-type allele is shown above the gray line with the variant below. The visualization of microscopic structural variants is adapted from [12].

(InDel) shorter than 50 base pairs long. Apart from that, clusters of close SNVs on the same allele are defined as multi-nucleotide variants [13]. Variants longer than 50 base pairs are classified as submicroscopic structural variation [14]. Copy number variation is a sub-form that refers to changes of the number of a certain genomic region, which includes deletions and duplications. Combinations of the types of structural variation are also observed and referred to as complex structural variants. Genetic variation affecting more than three million base pairs forms the class of microscopic structural variation as it may be identified through a microscope [15]. This class includes large forms of submicroscopic structural variation and changes in the number of the chromosomes summarized as aneuploidies. Further massive chromosomal rearrangements are named chromoanagenesis and are commonly observed in degenerated cancer genomes [16].

1.1.2 Phenotypic consequences of genetic variants

Every human genome has 4.1 to 5.0 million sites that are different from the reference genome [17]. However, the influence of each genetic variant on the phenotype of an individual varies in a broad spectrum from having no impact on the phenotype at all, influencing phenotypic features, such as height or eye color, or causing diseases. Assessing the functional impact of a genetic variant in context of the genotype of an individual is crucial to estimate its phenotypic consequence, and to identify the few potentially pathogenic variants among the millions of genetic variants when analyzing patient sequencing data.

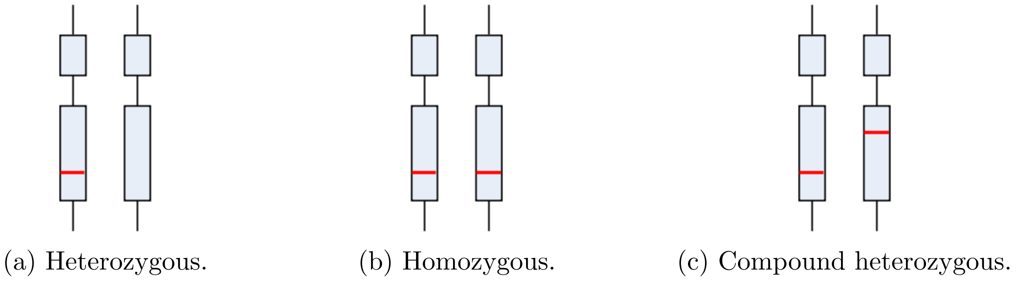
Functional impact Genetic variation can have a variety of different consequences on functions encoded in the human genome. Variants altering the amino acid sequence of proteins can modify all kinds of functions fulfilled by proteins, such as signaling, building cell structures, or executing biochemical reactions. Variants in other genomic regions can affect regulation of gene expression, for example, through the alteration of regulatory molecules like non-coding RNAs or by directly modifying DNA binding sites of transcription factors. The actual functional impact of a genetic variant depends on its type and its position in the genome. The more base pairs are affected, the more likely it is that functional elements in the genome will be altered, leading to changes of the molecular, cellular or physiological phenotype. If a variant is located at a site that encodes a critical functional element, the mutation will have a more pronounced effect than one in a less functionally constrained site. Approximately 80% of the human genome have functional elements assigned [18], but the interpretability of variants differs for variants located in loci encoding genes in comparison to intergenic variants. The functional mechanism of regulatory non-coding elements is often not well understood and subsequently it is unclear how genetic variation interferes with it. In contrast, the generation of protein products from coding genes is far better understood and the functional consequence of genetic variants is easier to predict. In the context of this work, variants in coding regions are grouped in three categories, namely protein-altering variants that modify the sequence of the encoded protein directly, variants that affect splicing by disrupting splice sites or other genetic regions that affect splicing, and variants in the 5' untranslated regions (UTRs) that create or disrupt upstream open reading frames (uORFs).

- **Protein-altering variants** Changes in the nucleotide sequence of coding regions can have various effects on encoded proteins reflected by a multitude of terms reflecting the functional impact of a variant. SNVs that introduce a premature stop codon are called stop-gain variants, SNVs that cause an exchange of one amino acid with another at a certain position are classified as missense variants, and SNVs that alter the start or stop codon are referred to as start- or stop-loss variants. An InDel in the coding sequence of a protein is termed an inframe insertion or deletion if its length is a multiple of three, otherwise it is termed a frameshift variant. Variants affecting the first or last dinucleotides of an intron are named splice donor or splice acceptor variants, respectively. Such splice site variants can lead to differences in the amino acid sequence by causing aberrant splicing, for example, through exon skipping or exon truncation [19]. In the context of ranking nucleotide changes by impact on encoded proteins, the term Loss-of-Function (LoF) is used to subsume splice site, stop-gain and frameshift variants because the disruption of splicing, the premature termination of the protein sequence, or the alteration of multiple amino acids likely render the encoded protein dysfunctional.

- **Other splicing variants** This group refers to mutations of branch points or variants in introns or exons that activate cryptic splice sites. As branch point and cryptic splice sites are not fully annotated for the human genome, such splicing defects are harder to identify as splice site variants, thus are classified as a separate group.

- **5' UTR variants** Approximately half of all human genes have a uORF in their 5' UTR [20]. These elements serve as regulators of translation as active translation of a uORF reduces the expression of its protein by up to 80%. A recent study observed strong negative selection for two distinct types of variants related to uORFs [21]. Variants generating a start codon for a new uORF and variants disrupting the stop codon of an existing uORF are the most deleterious variant types among all analyzed 5' UTR variants and are characterized by their ability to create reading frames that overlap with the coding sequence. Thus, such variants are promising candidates for the search for pathogenic variants.

Genotype The phenotypic consequence of a genetic variant is determined by the functional impact of the variant itself and its allelic composition or genotype. A variant occurring on one of both alleles of a locus is called heterozygous and results from a *de novo* mutation, or was inherited from either parent (Figure 1.2a). If the same variant is found on both alleles, it forms a homozygous genotype, as visualized in Figure 1.2b. A pair of heterozygous variants at the same locus is referred to as compound heterozygous if one of the variants is located on the maternal and the other on the paternal allele (see Figure 1.2c). Homozygous and compound heterozygous genotypes are also summarized as biallelic variants. While *de novo* mutations can also contribute to biallelic genotypes occasionally, homozygous variants and compound heterozygous variants usually result from one maternally and one paternally inherited variant allele. An accumulation of homozygous variants, which are otherwise rare in the population, indicates either related parents or an ethnicity that is not included in the reference genome.



(a) Heterozygous. (b) Homozygous. (c) Compound heterozygous.

Figure 1.2: Possible genotypes of genetic variants. A heterozygous genotype is given if the genetic variant is found either on the maternal or the paternal allele. It is a homozygous genotype if both alleles are affected by the same variant. A compound heterozygous genotype occurs if one of two different heterozygous variants is located on the maternal and the other on the paternal allele. The gray rectangles indicate exons. The red line visualizes the position of a genetic variant.

Recessively inherited traits or diseases are resulting from biallelic genotypes, while dominant traits are mediated by monoallelic variants. When considering the $4,695$ gene-disease phenotype relationships annotated in the Online Mendelian Inheritance in Man (OMIM) [22] data base by June 2020, the majority (51.84%) are inherited in autosomal recessive manner. Nevertheless, autosomal dominant inheritance still accounts for 41.70%, with the remaining 6.5% resulting from gonosomal inheritance. Indeed, there are several mechanisms how heterozygous variants can have an effect on the phenotype, including severe diseases. First, a heterozygous variant can have an effect equal to a homozygous variant if only the mutant allele is expressed. Monoallelic expression can be caused by heterozygous variants preventing the transcription or the translation of a functional gene product of one allele, or by X chromosome inactivation. For example, in women with Fabry disease, an X-linked inborn error of glycosphingolipid catabolism caused by mutations in the gene *GLA*, the direction and degree of X chromosome inactivation influences the expression of the mutant allele and thereby the severity of the disease phenotype [23].

Even if one of both alleles is properly transcribed and translated, a failure of the other allele by a deleterious variant may cause harmful consequences when the functionality of the gene product is dosage dependent. Genes encoding such proteins are called haploinsufficient, while genes for which the expression of one allele is sufficient to perform the encoded function are called haplosufficient. Genes that encode enzymes are enriched for haplosufficiency, while structural and regulatory proteins are enriched for haploinsufficiency [24]. The haplosufficiency of enzymes results from the fact that they do not act in isolation, but are part of larger systems of kinetically linked enzymes. In such systems a reduction to 50% enzyme activity as a consequence of a deleterious heterozygous variant is not expected to be detectable in the phenotype [25].

In addition to haploinsufficiency, a heterozygous variant can also have a dominant-negative effect when the mutant protein interferes with the function of the wild-type protein. *IKZF1* is a gene for which both, haploinsufficiency and dominant-negative alleles have been described as disease mechanisms. One study found heterozygous mutations in *IKZF1* that disrupt DNA binding of the encoded transcription factor IKAROS without inhibiting DNA binding of wild-type IKAROS [26]. Another study identified heterozygous mutations in *IKZF1* that also disturb the function of wild-type IKAROS by heterodimerization with mutant IKAROS [27]. The different disease mechanisms are also reflected in the disease phenotypes. The mutations in *IKZF1* underlying haploinsufficiency cause “common variable immunodeficiency”, while dominant-negative mutations cause a more severe type of immune defect called “combined immunodeficiency”.

In contrast to the reduction of functionality, so called gain-of-function mutations have a phenotypic effect as well. By improving or extending the encoded function of a protein they can cause imbalance in signaling cascades. An example is a heterozygous mutation in *JAK1* in patients with autoinflammation, immune dysregulation, and eosinophilia [28]. *In vitro* studies in patient cells verified a gain-of-function effect by showing increased JAK1 kinase activity resulting in enhanced STAT1/STAT3 phosphorylation. The translation of these findings initiated a treatment with ruxolitinib, an inhibitor of JAK1, which reduced signaling and improved the condition of the patients.

1.1.3 Challenges of genotype-phenotype correlation

When similar phenotypic features occur in related individuals, the question arises whether or which genetic loci are responsible for these characteristics. This is especially true for diseases that are characterized by high a degree of heritability, such as hearing loss [29]. Many challenges exist when trying to answer which genetic variant causes a certain phenotype or what phenotypic consequence does a certain variant have. Often, there is not a one-to-one relationship between genes and phenotypic traits, but rather polygenicity and pleiotropy are common, that is, multiple genes effect a trait or a single gene influences multiple traits respectively [30, 31]. In the context of gene-disease associations, these concepts are extended by the observation that the same clinical presentation can result from genetic variants in different genes, just as different variants in the same gene can be causative for distinct disease phenotypes in affected individuals [32]. In addition, incomplete penetrance can cause varying degrees of severity of the symptoms in individuals that carry the same pathogenic variant [33]. Finally, the interpretation of genetic variants must take into account the age of the individual as a genetic disease might only manifest when the patient grows older. Environmental factors are also important to consider, as, for example, variants that reduce response to certain pathogens have no effect in individuals who are not exposed to these pathogens [34].

The lack of a large-scale sequencing technology that can reliably identify small variants and structural variants of all sizes at the same time is another issue. Multiple experimental methods have to be applied to fully describe the genetic variation of an individual, which is too resource-intensive to be applied in large scale. Consequently, most genotype-phenotype correlation studies suffer from varying degrees of incomplete coverage of genetic variability. In general, there are three main study types used to establish genotype-phenotype correlations.

Case-control studies Case-control studies can be used to identify single genetic variants or loci that are significantly enriched in a group of cases in comparison to a control group. When searching for disease-associated loci, the case group consists of patients suffering from a specific disease, while the control group comprises healthy individuals. The proper selection of cases and controls is a key element in the design of case-control studies. As qualitative disorders can be interpreted as being the extremes of quantitative dimensions, the health of controls must be assessed thoroughly to avoid noise in genetic signals due to individuals being not affected yet or nearly asymptomatic but carrying the genetic predisposition [35]. A further prerequisite of case-control studies is a high number of cases and controls to generate sufficient statistical power for valid significance testing. The completion of the HapMap project [36] and the development of genotyping arrays made it possible to test for associations of millions of frequent SNVs with a variety of phenotypes in so-called genome-wide association studies [37]. As these studies are not limited to existing knowledge about the location of functional elements in the human genome, they are a powerful instrument to identify genetic associations and generate hypotheses of their functional relevance. However, the presence of genetic causality can only be established by subsequent studies, since the associated common SNVs are only markers for the region containing the causative genetic variation [38]. The advent of next-generation sequencing technologies made it feasible to also assess rare and potentially causal variation in large-scale studies. Together with the development of new statistical frameworks to test the association of single rare variants or their collective burden on gene-level, also rare and potentially causal variants can be explored in case-control studies [39]. However, there is still a conceptual limitation on common diseases as statistical power will suffer from very small and insufficient case group sizes when focusing on rare diseases.

Family studies Tracking the inheritance of genetic variants together with the expression of phenotypic traits of interest in multiple generations of one or more families is another approach to establish genotype-phenotype correlations. Similar to case-control studies, such so-called family studies can be used to associate genetic variants or loci with phenotypic traits without any prior knowledge about gene function. Likewise, it can also be a challenge to pinpoint the causative variant, depending on how precisely the region segregating with the phenotype can be narrowed down. In contrast to case-control studies, the collection of case and control groups is not necessary and a single affected pedigree is sufficient to perform a family study. Thus, this study type is also suited for rare phenotype conditions. Nevertheless, it is more powerful the more patients are included, either through analyzing several generations or multiple pedigrees. The inclusion of multiple generations is obviously limited by the number of living family members or frozen DNA samples. In the case of rare diseases, there might only be a few patients described with the disease phenotype worldwide and the genetic causalities might be different.

Proband-only studies Especially for genetic diagnosis of rare diseases in a clinical context, case-control or family studies are inefficient as a first tier attempt to identify the causative genetic variant It is more common to rely on existing knowledge and predictions of the functional impact of genetic variants to associate them with a given phenotype. But even if restricting to protein-altering variants, there are still 10,000 to 12,000 missense variants [17] and approximately 100 LoF variants per individual that have to be considered [40]. To differentiate benign from pathogenic variation, population allele frequencies are accounted for, as fully penetrant pathogenic variants cannot be more frequent than the prevalence of the disease of interest. In particular, there is a focus on rare variants as their low frequency indicates negative selection. However, rareness is not necessarily the result of negative selection. It can also result from non-random selection of individuals used to compute population allele frequencies. Large sequencing studies are usually biased towards European and American populations with an underrepresentation of, for example, Asian or Great Middle Eastern ethnicities [41]. Consequently, variants common in underrepresented or missing populations will seem rare when analyzing the DNA of such individuals. Finally, the assessment of a set of rare variants with regard to their potential pathogenicity tends to be biased towards variants in genes with known function as this allows the generation of intuitive hypotheses of the underlying pathomechanisms. This is a major problem because only 2,000 of the 19,000 human protein-coding genes are comprehensively functionally characterized due to the chemical, physical and biological properties of their encoded gene products that make them easier to analyze [42].

1.2 Definition of rare diseases

Although rare diseases affect patients in all parts of the world, there is no common prevalence threshold that defines a disease as rare. Instead, national definitions range from 5 to 80 patients per 100,000 individuals [1]. To overcome these differences and to compile a defined set of rare diseases, the Orphanet database was created in 1997 [43]. Since then Orphanet became an essential online resource that provides a rare disease nomenclature, an inventory of drugs to treat rare diseases, compiled lists of patient organizations and expert centers, and other services for patients and the scientific community. An analysis of Orphanet in October 2018 reports $6,172$ unique rare diseases [1]. Of these, 69.9% are manifesting in early childhood and 71.9% are genetic, including diseases with a known gene defect and diseases known or suspected to be familial, but for which an underlying gene defect has not yet been identified. While most genetic defects manifest in childhood, there are also genetic defects that cause rare diseases in adolescence or adulthood, such as Kennedy's disease. This spinal muscular atrophy is caused by a trinucleotide CAG repeat expansion in the androgen receptor gene on chromosome X that reduces the receptor's transcriptional activation activity [44]. Male carriers of the repeat expansion typically experience the onset of the disease between the ages of 30 and 50 while early symptoms of the disease can already be recognized in adolescence [45].

The distribution of the estimated prevalences of $5,304$ rare diseases in Orphanet shows that 84.5% of them affect less than 1 in $1,000,000$ individuals [1]. The majority of the burden of rare diseases in the population (77.3 - 80.7%) is caused by a small subset of 149 rare diseases with prevalences ranging from 10 to 50 in $100,000$. In total, the population prevalence of rare diseases is estimated between 3.5% and 5.9%, that is, there are 263 to 446 million rare disease patients world-wide. On average, this corresponds to the size of the population in the United States, which emphasizes the global burden caused by rare diseases.

1.3 Inborn errors of immunity (IEI)

Inborn errors of immunity (IEI) are a group of rare diseases defined by congenital defects in the human immune system causing a wide range of severe symptoms. Recurrent and severe infections are the result of a permanently weakened immune system, while autoimmunity or autoinflammation indicate an overactive immune system. Recently, the International

Union of Immunological Societies (IUIS) Export Committee has reported 416 IEI [32]. It has to be assumed that this set is far from being complete, as dozens of new immunodeficiencies have been described in each report over the last 20 years (see Figure 1.3). Together with the growing number of described IEI, the overall prevalence of this group of rare diseases is now estimated to affect between two and ten births in 10,000, which is ten times higher as assumed earlier. IEI differ greatly in how distinctly they can be defined, how well symptoms can be assessed and how specific these are to infer the underlying genetic defect. One example for a concisely defined disease sub-group with a wide phenotypic spectrum are severe congenital neutropenias (SCNs) identified by a low amount of neutrophils in the peripheral blood. Early-onset inflammatory bowel disease (IBD), characterized by chronic inflammation of the intestines, is an example of a pathology that can indicate an underlying immune defect, but can also have other etiologies.

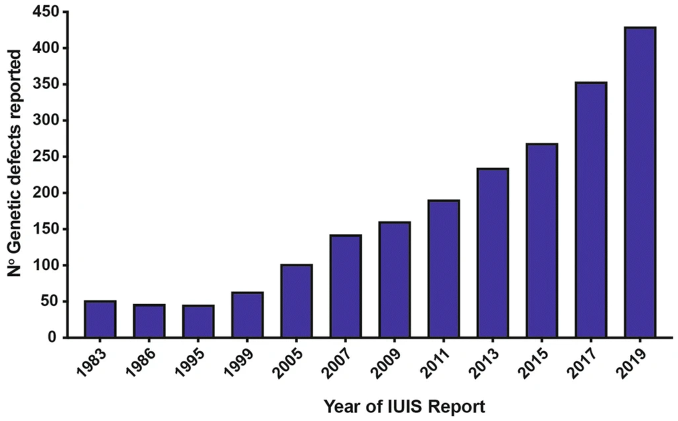


Figure 1.3: Number of immune defects in International Union of Immunological Societies (IUIS) reports since 1983. Since 1999 the number of described inborn errors of immunity is growing rapidly. Figure taken from [32].

.

1.3.1 Severe congenital neutropenia (SCN)

Congenital neutropenias are a group of rare diseases characterized by a decreased number of neutrophils in the peripheral blood. The absolute neutrophil count is used to divide neutropenias into mild ($1.0 - 1.5 \cdot 10^9$ cells/l), moderate ($0.5 - 1.0 \cdot 10^9$ cells/l) and severe ($< 0.5 \cdot 10^9$ cells/l) subtypes [46]. The latter form, referred to as SCN, is further defined by an accumulation of promyelocytes in the bone marrow as a consequence of arrested myelopoiesis during the differentiation of hematopoietic stem cells into neutrophils. In general, decreased numbers of neutrophils cause a higher susceptibility to infection, with patients with SCN suffering from severe, recurrent and often life-threatening infections setting on as early as the first months of childhood. SCN has an estimated prevalence between 3 to 8.5 cases per million individuals [47]. Besides non-syndromic forms, congenital neutropenia is often observed together with other features, such as oculocutaneous hypopigmentation, pancreatic insufficiency, metabolic disease, or it is part of the symptoms of other IEI [48]. In addition to the general prevention of microbial infections, the therapy of SCN consists of regular subcutaneous administration of granulocyte colony-stimulating factor (G-CSF), which increases the blood neutrophil count and improves quality of life and life expectancy [49]. A summary of several studies shows that patients with SCN are at risk to develop acute myeloid leukemia or myelodysplastic syndromes [47]. At present, congenital neutropenia can only be cured by allogeneic hematopoietic stem cell transplantation.

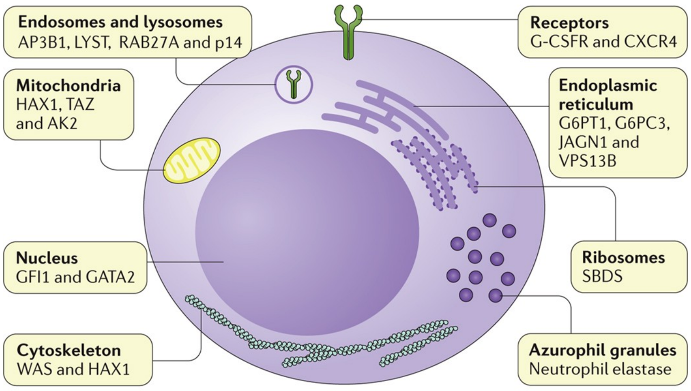


Figure 1.4: Proteins mutated in patients with congenital neutropenia. Neutrophil elastase is also referred to as ELANE. Figure adapted from [47].

The latest IUIS report lists 21 gene defects causing primarily congenital neutropenia together with several other IEI where neutropenia is one of the known associated features [32]. The described mutations affect different components of hematopoietic stem cells and myeloid cells resulting in a variety of pathogenic mechanisms, such as protein and vesicle mistrafficking, endoplasmic reticulum stress or disturbed energy metabolism. Almost half of the SCN patients carry autosomal dominant mutations in *ELANE*, which encodes neutrophil elastase, a protein playing a key role in innate immune defense [50]. The most frequent causes of autosomal recessive non-syndromic neutropenia are mutations in *HAX1*, which is critical for stabilizing the mitochondrial membrane potential [51]. Figure 1.4 gives an overview of mutated proteins and their cellular localization in hematopoietic stem cells and myeloid cells.

1.3.2 Very early onset inflammatory bowel disease (VEOIBD)

IBD is a group of heterogeneous diseases characterized by chronic or relapsing inflammation in the gastrointestinal tract. Depending on the affected parts of the digestive tract, IBD is classified as Crohn's disease, ulcerative colitis or IBD-unclassified. Very early onset IBD (VEOIBD) is a subtype of IBD defined by an onset in childhood below 6 years of age. The classification is motivated by the fact that a causative genetic defect is more likely in younger patients in contrast to adult IBD, which is influenced by general genetic predisposition and environmental factors [52]. Additionally, there is an increased likelihood that VEOIBD results from an underlying IEI. The close relationship of VEOIBD and immune defects is reflected by the observation that 20% of the known genetic defects causal for IEI are accompanied by IBD manifestations [53]. VEOIBD is often associated with a more severe disease course characterized by increased surgical interventions and growth failure [54]. Pediatric IBD is one of the more common rare diseases and its incidence is growing [52]. A Canadian epidemiological study observed an increase from 9.68 cases per 100,000 children in 1999 to 38.25 cases per 100,000 children in 2010 [55]. A similar trend has been observed in the United States where the prevalence has risen from 33 per 100,000 children in 2007 to 77 per 100,000 children in 2016 [56]. However, neither study was able to determine the extent to which increasing awareness and diagnosis of pediatric IBD and/or changing environmental conditions and dietary habits are responsible for the increased prevalence [55, 56].

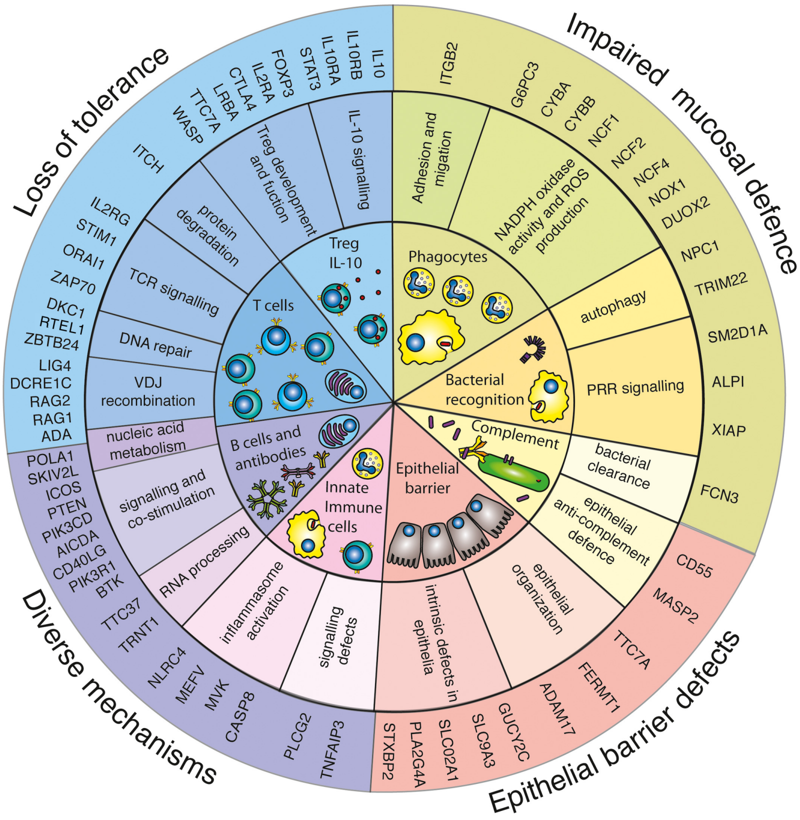


Figure 1.5: Pathomechanisms of monogenic inflammatory bowel disease. Molecular pathomechanisms are grouped into four categories: loss of immune tolerance, impaired mucosal defense, epithelial barrier defects and other mechanisms. The inner circle depicts involved cell types and cell components. The middle circle lists cellular pathways affected by genetic defects in the genes represented in the outer circle. The following abbreviations are used: Treg for regulatory T cells, TCR for T cell receptor, PRR for pattern recognition receptor and ROS for reactive oxygen species. Figure taken from [53].

A recent review lists 63 monogenic defects causing VEOIBD [52]. The proteins encoded by the genes affected play roles in various functions of the digestive tract. Some etiologies are partly overlapping and can be grouped together according to the affected cell type or cellular component, as shown in Figure 1.5. A group of defects involves a general deregulation of the immune system often through dysfunctional T and B cells. Other groups comprise mutations that impair the mucosal defense or the integrity of the epithelial barrier. Despite the growing number of identified genetic defects, the genetic etiology cannot be determined for the majority of VEOIBD patients. The proportion of genetically undiagnosed cases is strongly depending on the underlying patient cohort. A cohort study of VEOIBD patients younger than 6 years of age identified the causal gene defect for 32% [57], while another study based on a cohort of patients aged 0 to 18 years diagnosed 3% [58]. In addition to the growing prevalence of IBD, the results of these studies suggest that the role of environmental factors in IBD development increases with the age of the patients. The examination of the genetic background is nevertheless indispensable to choose an accurate therapeutic option [52]. In case an inborn error of immunity as cause of IBD can be determined, allogeneic hematopoietic stem cell transplantation provides a curative treatment approach [59]. Alternatively, the knowledge of the underlying pathomechanisms may guide targeted treatments with specific drugs, surgery or nutritional approaches, which can improve the life of the patients considerably.

1.4 Genetic diagnosis of rare diseases in clinical practice

In multi-factorial diseases, such as type 2 diabetes or schizophrenia, the exact cause of disease can hardly be determined for each patient due to the high dimensionality of genetic and environmental factors. In contrast, the majority of rare diseases is caused by single genetic defects. Thus, it is possible to pinpoint the exact cause of disease with current sequencing technologies. Additionally, the biological malfunction can be uncovered as there exists a defined cause-and-effect relation, if the genetic defect was identified. Once the biological mechanism leading from the genetic defect to the pathophenotype is understood, attempts can be made to correct it through targeted interventions. If such an attempt is successful, all patients with the same gene defect can be cured. Although the path from a genetic diagnosis to a targeted therapy can take decades, a recent study has shown that already the knowledge about the genetic cause of his or her disease can be of great value for the patient [60]. At best, of course, a targeted therapy is already available and the patient can be treated accordingly or even be cured. Otherwise, a genetic diagnosis ends at least the often long-lasting journey of visiting various doctors, trying diverse treatments, and undergoing medical tests without even knowing the exact disease. The diagnosis can be used to inform the patient about the expected progression of the disease and serve as basis for genetic counseling. Finally, the patient can play an active role in fighting the disease by being able to actively participate in projects performing basic or translational research on the identified genetic defect.

Whenever a genetic defect is the suspected cause of the pathophenotype of a patient, next-generation sequencing (NGS) techniques are increasingly used in clinics to perform the genetic diagnosis of patients suffering from rare diseases. Whole-exome sequencing

(WES) is a method to analyze variants in the 1-2% of the human genome that is coding for proteins, while whole-genome sequencing (WGS) covers genetic variants in the entire genome. Despite the more comprehensive coverage of genetic variance by WGS, WES has become the routine approach in clinical practice because it is cheaper and achieves a similar diagnostic rate [4, 5].

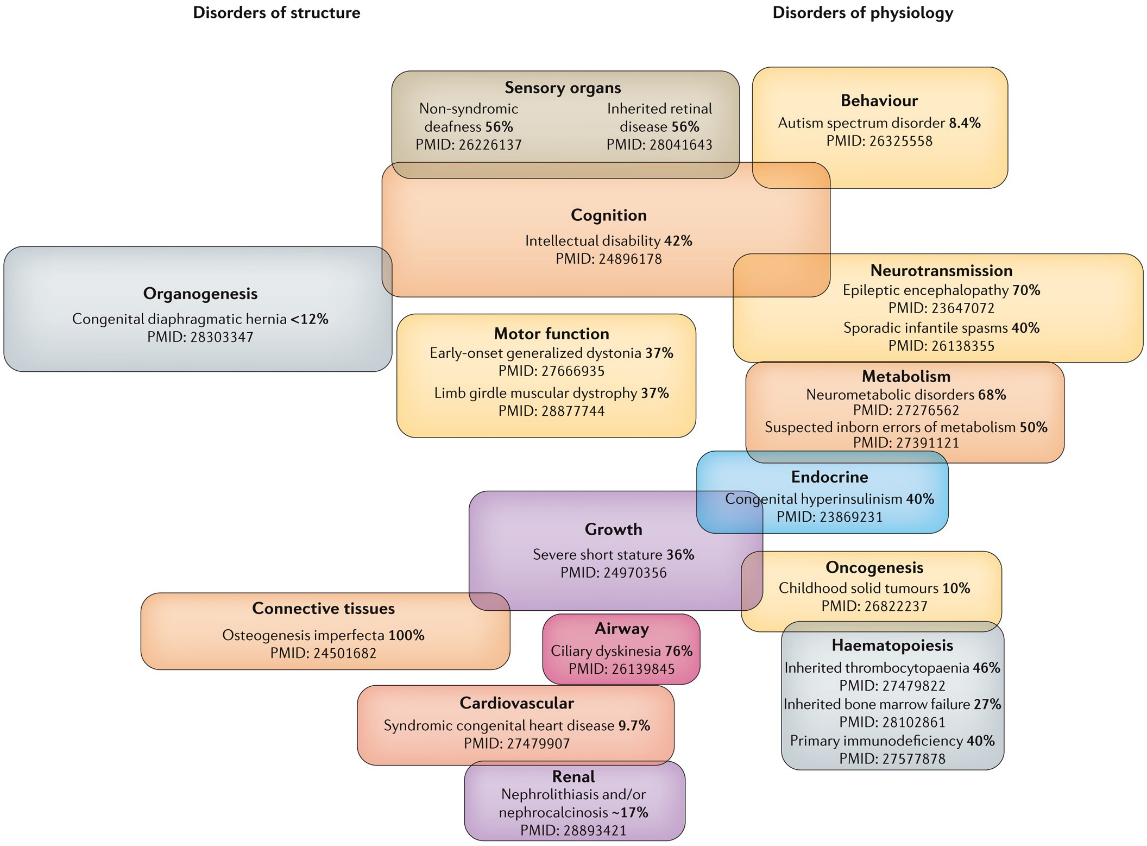

Disorders of structure
Disorders of physiology
Sensory organs
Non-syndromic deafness 56%
PMID: 26226137
Inherited retinal disease 56%
PMID: 28041643
Behaviour
Autism spectrum disorder 8.4%
PMID: 26325558
Cognition
Intellectual disability 42%
PMID: 24896178
Organogenesis
Congenital diaphragmatic hernia <12%
PMID: 28303347
Neurotransmission
Epileptic encephalopathy 70%
PMID: 23647072
Sporadic infantile spasms 40%
PMID: 26138355
Motor function
Early-onset generalized dystonia 37%
PMID: 27666935
Limb girdle muscular dystrophy 37%
PMID: 28877744
Metabolism
Neurometabolic disorders 68%
PMID: 27276562
Suspected inborn errors of metabolism 50%
PMID: 27391121
Endocrine
Congenital hyperinsulinism 40%
PMID: 23869231
Growth
Severe short stature 36%
PMID: 24970356
Oncogenesis
Childhood solid tumours 10%
PMID: 26822237
Connective tissues
Osteogenesis imperfecta 100%
PMID: 24501682
Airway
Ciliary dyskinesia 76%
PMID: 26139845
Haematopoiesis
Inherited thrombocytopaenia 46%
PMID: 27479822
Inherited bone marrow failure 27%
PMID: 28102861
Primary immunodeficiency 40%
PMID: 27577878
Cardiovascular
Syndromic congenital heart disease 9.7%
PMID: 27479907
Renal
Nephrolithiasis and/or nephrocalcinosis ~17%
PMID: 28893421

Figure 1.6: Diagnostic rates of whole-exome sequencing studies by disease phenotype. The size of the boxes reflects the prevalence of the disease phenotypes in pediatric practice. Figure taken from [5].

The general strategy of using DNA sequencing data is the definition and implementation of criteria to select the most promising potentially pathogenic genetic variants from the thousands of detected variants per patient. The result is a list of variants supported with varying level of evidence to be pathogenic. In order to prove that a selected variant causes the observed pathophenotype, it needs to be shown that the variant is not present in individuals without this phenotype, that it impairs the function of the encoded gene product, and that the phenotype can be recapitulated by introducing the mutant allele or can be rescued by introducing the wild-type allele in an appropriate model system [61]. As there are no methods available to verify or falsify the potential pathogenicity of dozens of candidate variants per patient automatically, stringent variant prioritization strategies are required. Despite advances in speed and quality of sequencing methods and more sophisticated bioinformatics tools, the results of multiple studies consistently indicate that the diagnostic rate is limited for most pediatric diseases, as visualized in Figure 1.6 [5].

There are different reasons why a genetic diagnosis may fail when using WES or WGS. The patient might not suffer from a genetic disease or a mosaic variant is causative, which is missed when DNA is taken from organs or tissues that do not harbor the variant. Further, the pathophenotype might be caused by genetic variation that is not detected. Due to the shortness of the sequenced DNA fragments most structural variation cannot be identified by NGS techniques. If WES is used, variants in the non-coding regions of the human genome are additionally not covered. Accurately predicting the functional impact of detected variants is another challenge, especially for variants in non-coding regions, such as UTRs, introns or intergenic regions. Even if a deleterious impact on protein-level can be confidently predicted, it often remains unclear if the gene has any relevance for the clinical phenotype. Finally, incomplete penetrance, digenic inheritance of pathogenic variants or more complex forms of multigenic etiology might easily be missed when assuming Mendelian inheritance.

1.5 Motivation and objectives

The Dr. von Hauner Children's Hospital of the Ludwig-Maximilians-Universität in Munich is part of an international network of research institutions collaborating in the research for rare diseases and personalized medicine [62]. Within this framework, the Dr. von Hauner Children's Hospital plays a central role as it houses the Care-for-Rare Laboratories [63] with their own NGS facility that allows immediate sequencing and analysis of patient samples taken on site at the hospital or sent in by international collaboration partners. A three-stage strategy is pursued to identify pathogenic variants that serve as the starting point for understanding the underlying disease mechanisms. First, the exome of the patient gets sequenced to screen for recessive effects in protein-coding and adjacent intronic regions as well as in untranslated regions. If none of the identified variants provides a conclusive explanation for the observed pathophenotype, WES of the parents is performed to search for potentially pathogenic *de novo* variants. If the trio analysis does not yield a promising candidate variant, patients are prioritized for additional studies, such as WGS, transcriptome or proteome analysis. Despite being highly informative, sequencing of parents cannot be done for all patients due to financial constraints and limited availability of blood or tissue samples of patients and healthy relatives.

My project at the Dr. von Hauner Children's Hospital aims to leverage the growing amount of collected WES data to extend the current knowledge on the genetic etiologies of IEI. For this purpose, I explore ways to improve the analysis of NGS data and establish a new workflow to discover novel candidates for disease genes by screening the entire WES data collection from 2, 451 individuals. As part of this, I develop new bioinformatic tools and routines to meet the special requirements of clinical research. By application of the implemented pipelines, I want to characterize the patient cohort thoroughly and gain new biological insights into the pathogenesis of IEI.

Material and Methods 2

My project is based on WES data of a cohort of $1,746$ patients with IEI together with 705 healthy relatives and on WGS data of a family with a child suffering from SCN. For the analysis, I've created several modules assembled from published tools and self-developed methods specifically tailored for clinical application. Additionally, several public data sets are integrated to support the analysis and interpretation of the sequencing data.

2.1 Clinical sequencing data

The NGS facility of the Care-for-Rare Laboratories of the Dr. von Hauner Children's Hospital is equipped with a NextSeq 500 and a NextSeq 550 sequencer (Illumina, San Diego, California) to analyze whole exomes of patients and healthy relatives. Prior to the acquisition of the NextSeq machines, a SOLiD sequencing platform was used that contributed to the exome data set as well. Exome enrichment was done using the Agilent SureSelect V5/V6+UTR kits (Agilent, Santa Clara, California) consisting of RNA probes targeting annotated exon regions while ensuring comparable hybridization conditions.

This work takes the entire WES data collection to evaluate genetic variants in known IEI genes and to search for yet unknown genetic defects underlying IEI. Moreover, it includes the analysis of externally generated WGS data of a family, for which preceding WES did not result in a reasonable candidate variant, to test whether a genetic diagnosis can be made by WGS in this case.

2.1.1 Whole-exome sequencing (WES) data collection

The main data set comprises sequencing data of $2,451$ individuals, $1,746$ of whom are patients and the other 705 are otherwise healthy parents, siblings or other relatives. Among the patients, there are 315 patients having both parents sequenced, referred to as trio patients. About 30% of the blood samples for sequencing were taken in Germany, approximately 20% were sent from Turkey and Iran respectively. The remaining part originates mainly from other countries in Europe and the Middle East. Patients with IBD or SCN as the main pathophenotype are stratified into two corresponding sub-cohorts together with their relatives. Table 2.1 gives an overview of the numbers of trio patients, non-trio patients and healthy relatives for the respective disjoint sub-cohorts.

Cohort	Trio patients	Non-trio patients	Healthy relatives
Inflammatory bowel disease	21	655	54
Severe congenital neutropenia	123	190	268
Other immune defects	171	586	383
Total	315	1,431	705

Table 2.1: Composition of the exome sequencing data collection. Trio patients are patients for whom both parents were sequenced. Non-trio patients were sequenced either as singletons, together with only one parent or with other healthy relatives. According to the main pathophenotype all patients and their relatives are stratified into three sub-cohorts.

2.1.2 Whole-genome sequencing (WGS) data of a family

Besides the WES data, one WGS data set of a Romanian family is part of the data underlying this thesis. The pedigree of the family, called SCN-1 in the following, is shown in Figure 2.1. The index patient II-3 presented with recurrent pulmonary infections, failure to thrive, SCN and Shwachman-Diamond syndrome (SDS) when she was five years old. Her brother II-2 also showed failure to thrive and skeletal abnormalities, but is not affected by neutropenia or pancreatic insufficiency.

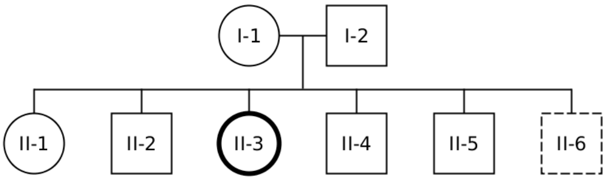


Figure 2.1: Pedigree of family SCN-1. The bold line highlights the index patient. Whole-genome sequencing was performed for all family members except II-6, visualized by the dashed line. The order of the children corresponds to the order of the dates of birth. Square nodes indicate male family members, circular nodes indicate females.

An initial WES analysis of the index patient, the parents and two brothers (II-2 and II-4) did not reveal any pathogenic variant in known SCN or SDS related genes. Because it was suspected that a non-coding variant underlies or modulates the symptoms of the index patient and her brother II-2, WGS was done for all family members at the HudsonAlpha Institute for Biotechnology in Huntsville, Alabama. The healthy brother II-6 did not undergo sequencing, because he was born later.

2.1.3 Ethics statement

All clinical sequencing data used in this work was generated as part of different studies that were approved by the Ethics Commission of the Medical Faculty of the Ludwig-

Maximilians-Universität in Munich. The reference numbers of the corresponding ethics votes are 346-11, 381-11, 387-11, 438-11, 486-11, 303-12, 187-13 BB, 353-13, 66-14, 501-14 and 806-16. The consent of the participants was obtained in written form.

2.2 Public data sets

Two types of information are crucial to distinguish pathogenic from benign genomic variation. Reliable data on the frequency of a variant allele is necessary to filter for rare alleles. Already existing knowledge about the pathogenicity of individual variants or entire genes helps to highlight known or potentially pathogenic variants. The following paragraphs introduce the integrated resources for the analysis and interpretation of the genetic variants identified in the sequencing data. Reference allele frequency data is obtained from the 1000 Genomes Project and the Genome Aggregation Database (gnomAD). ClinVar and the International Mouse Phenotyping Consortium (IMPC) are used as resources to inform about the pathogenicity of individual variants and entire genes.

2.2.1 1000 Genomes Project

Launched in 2008, the 1000 Genomes Project was the first project aiming to study genomic variation across different populations. It was finished in 2015 with the reconstruction of the genomes of $2,504$ individuals from 26 populations in East and South Asia, Europe, Africa and North and South America [17]. The analysis of low-coverage WGS, deep WES and dense microarray genotyping of all studied individuals showed that a typical genome has 4.1 to 5.0 million variant sites with only $40,000$ to $200,000$ rare variants having an allele frequency below 0.5%. Regarding variation in protein-coding regions, 149 to 182 protein-truncating and $10,000$ to $12,000$ protein-sequence-altering variants were reported per genome. The upper ranges of these figures were observed in African populations, which is consistent with an African origin of modern humans. Genetic variance could accumulate in African populations while serial founder effects have reduced genetic diversity in populations outside of Africa [64]. Although larger sequencing data sets have become available in recent years, the 1000 Genomes Project remains a highly valuable resource due to the multitude of included populations and its fully published data set.

2.2.2 Genome Aggregation Database

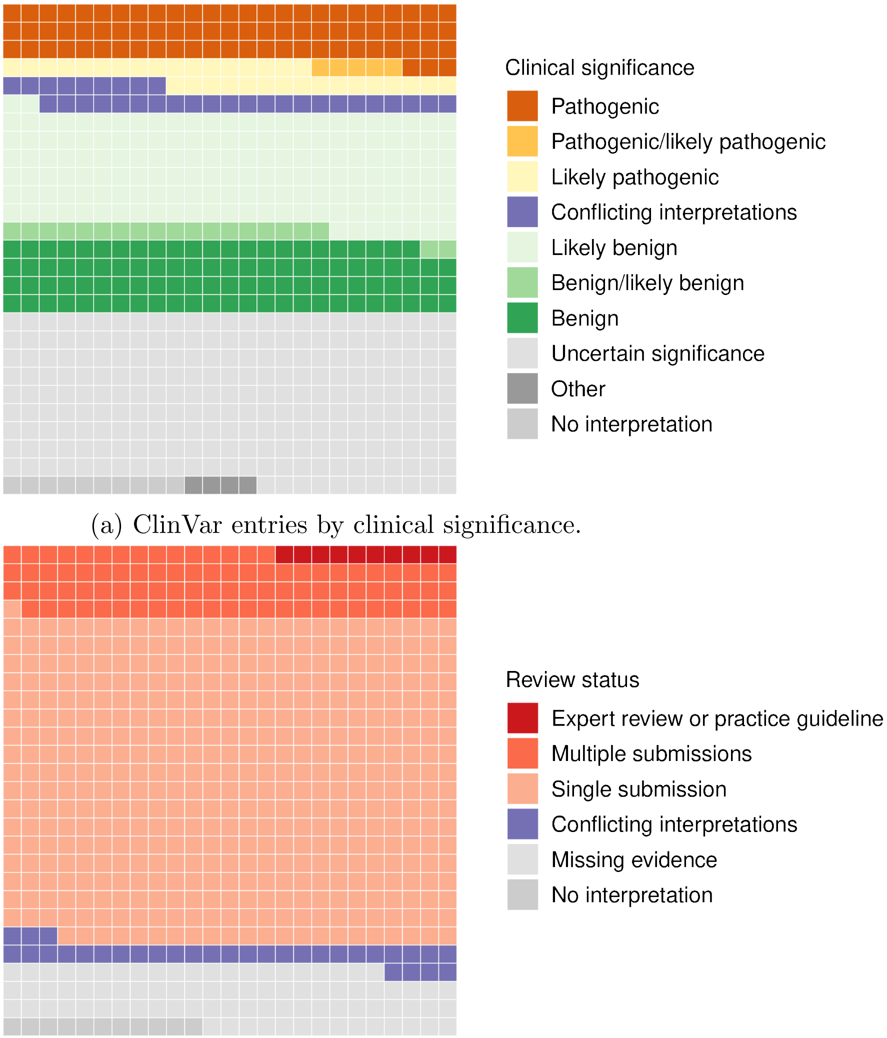
Following the 1000 Genomes Project, the Exome Aggregation Consortium (ExAC) has set the next major milestone for the analysis of human genomic variation [13]. The consortium has assembled WES data of $60,706$ human individuals and consistently processed all data to generate high-quality variant calls. The successor of the ExAC is the gnomAD data collection, which has expanded the data set to WES data of $125,748$ individuals and WGS data of $15,708$ individuals [65]. Because of its tremendous size, the frequency spectrum of the identified variants is resolved in much higher resolution in comparison to previous sequencing projects. Therefore, gnomAD has become the most important source of allele frequencies to filter for variants potentially causing rare pediatric diseases as individuals suffering from severe pediatric diseases are not part of the data set. However, since it cannot be completely excluded that the included individuals suffer from other diseases that are investigated in the respective integrated studies, the gnomAD consortium defined a subset of control individuals that are healthy with regard to the studied diseases. This control set comprises $54,704$ exome and $5,442$ genome data sets.

The size of the gnomAD data set enables the systematic comparison of the number of observed LoF variants per gene to the expected number inferred from a human mutation rate model. The median ratio of observed to expected LoF variants is 0.48, which shows that LoF variants are subject to negative selection in most genes. Thus, in addition to allele frequencies on variant-level, gnomAD provides a metric to estimate the deleteriousness of LoF variants on gene-level. Although gnomAD is essential for the analysis of clinical sequencing data, there are two evident limitations of the data set. First, the majority of the individuals has European ancestry while Middle Eastern populations are almost completely absent, which limits the filtering capability for variants common in Middle Eastern populations but rare in other populations. In contrast to the 1000 Genomes Project, the individual genotypes are not part of the published data set, which makes it impossible to study the composition of genotypes in single individuals. This information would be necessary, for example, to assess the frequency of compound heterozygous variant pairs.

2.2.3 ClinVar

ClinVar [66] is a public resource that collects interpretations of clinical significance of genetic variants indicated by their impact on certain disease phenotypes. Although the term "clinical significance" does not represent statistical significance but rather reflects the effect size of a variant, it is commonly used to describe the pathogenicity of a variant [67, 68, 69]. Identified genotype-phenotype relationships including the associated mode of inheritance and other details of evidence can be submitted by clinical and research institutions as well as other qualified groups. These reports are then aggregated on variant level to support variant interpretation through the type, consistency and quality of the provided evidence [70]. ClinVar classifies variants as pathogenic or likely pathogenic, benign or likely benign or as variants of uncertain significance. The reliability of the assigned interpretations is indicated by the review status ranging from missing or contradicting evidence, reports of only one submitter, consistent information of multiple submissions, to interpretations of expert panels or practice guideline-providing groups.

The February 2020 release of ClinVar contains $1,339,085$ variant interpretations, of which $674,527$ are annotated on the human reference genome assembly GRCh37 and all others on GRCh38. Figure 2.2 shows the distribution of the GRCh37 assembly based ClinVar entries across the different classes for clinical significance and review status. Re-



(a) ClinVar entries by clinical significance.

(b) ClinVar entries by review status.

Figure 2.2: Stratification of ClinVar entries by clinical significance and review status. Each row of the charts contains 25 squares with each square corresponding to 1,000 ClinVar entries. The squares are arranged according to the order of the corresponding legends starting with the squares representing the last legend entry in the lower left corner and then adding squares from left to right, from bottom to top for each legend entry. The interpretations released in February 2020 and based on the GRCh37 assembly were used to create the diagrams. (a) The five main categories for classifying the clinical significance of genetic variants are Pathogenic, Benign, Likely pathogenic, Likely benign and Uncertain significance. The sub-categories Pathogenic/likely pathogenic, Benign/likely benign and Conflicting interpretations are a consequence of submissions reporting different interpretations. (b) The review status of a variant interpretation can be divided into four quality levels. The highest reliability is achieved by expert panels or by practice-guideline providing groups. Multiple concordant submissions are less reliable, but still more trustworthy than single submissions. Interpretations without evidence or conflicting interpretations are the least trustworthy.

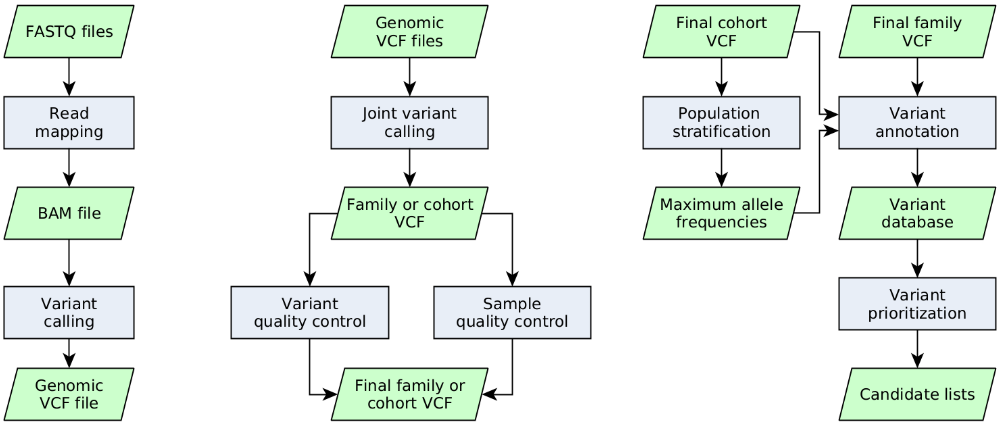
garding the clinical significance, variants of uncertain significance account for the largest proportion equivalent to 34%. The vast majority of interpretations come from individual submissions, indicating that the majority of the entries in the database should be treated with caution in the absence of independent confirmation of the classification. In total, there are 13,481 variants that have been classified as pathogenic with medium to high reliability through multiple submissions or expert review or practice guidelines. This subset is the most relevant for clinical sequencing as it enables fast and reliable detection of known disease-causing variants. Although its size seems rather small compared to the total number of entries based on GRCh37, ClinVar is nevertheless a highly valuable resource to quickly explore whether certain genetic variants have already been investigated in other studies.

2.2.4 International Mouse Phenotyping Consortium

The IMPC was founded in 2011 with the aim to create a comprehensive catalog of systematically phenotyped mono- and biallelic knockouts of 20,000 mouse genes [71]. When finished, the resource will represent the first fully functional description of a mammalian genome. In February 2020, the IMPC published release 11.0 containing 75,844 significant phenotypic traits in knockouts of 6,440 genes generated by the 24 consortium members worldwide. In addition to the new insights into the function of individual genes in mice, analyses across all genes yielded further important results. A study of the first 1,751 gene knockouts in 2016 showed that one third causes embryonic defects and that the underlying genes are enriched for the 3,302 human disease genes reported in the Human Genome Mutation Database [72] at the time of the study [73]. Furthermore, the authors of the study found that incomplete penetrance is frequently observed in these mouse strains despite their defined genetic background as a consequence of stochastic variability in gene expression of functionally redundant genes [74]. Another subsequent study on 2,186 gene knockouts in 2017 found that a large proportion of phenotypic traits is influenced by sex [75]. An analysis of the first 3,328 null mutants in the same year showed that 90% of all reported gene-phenotype relationships were hitherto unknown and that 1,092 genes were functionally described for the first time [76]. Although gene-phenotype relationships in mice knockout strains are generally not fully transferable to humans, the IMPC nevertheless provides a valuable resource to support the interpretation of genetic variation in human orthologs of already phenotyped mouse genes.

2.3 Analysis of genomic sequencing data

The search for pathogenic variants in the exome and genome sequencing data presented in Section 2.1.1 and 2.1.2 involves a multitude of steps starting from the raw sequencing reads. Figure 2.3 shows the three major parts of the overall workflow beginning with the identification of genetic variants for each individual separately (a), then using the variants to generate a comprehensive matrix of genotypes including only high quality variants and samples (b), and finally selecting potential pathogenic variants and candidate genes for disease associations (c).



(a) Separate variant identification.

(b) Joint variant calling and quality control.

(c) Generation of candidate list.

Figure 2.3: Main steps for processing sequencing data. The left part (a) shows the workflow to generate genomic VCF files from raw FASTQ files that is performed separately for each sequenced individual. The middle part (b) illustrates the combination of multiple gnomic VCF files to a single VCF file containing the genotype information of all individuals and the subsequent quality control of the identified variants and included samples. The right part (c) depicts the workflow steps required to create lists of candidate variants or genes. Input files, intermediate data and results are indicated as green nodes, gray nodes indicate processing steps. The abbreviation VCF stands for variant call format.

2.3.1 Variant calling and quality control

The Genome Analysis Toolkit (GATK, version 3.8-1) [77] and accompanying best practice guidelines [78, 79] are applied to detect genetic variants in two main steps as indicated in Figure 2.3a and 2.3b. They include mapping of raw reads and the improvement of the read alignments followed by single and joint variant calling and subsequent quality control measures.

Mapping raw reads Before identifying genetic variants, millions of short sequencing reads generated by WES and WGS have to be aligned or mapped to the human reference genome. The algorithm BWA-MEM [80] of the tool Burrow-Wheeler Aligner (BWA, version 0.7.15) [81] is used for mapping the reads to the human genome assembly GRCh37. While the original BWA algorithm is restricted to align the full sequence of a read, BWA-MEM decides automatically whether to generate local or end-to-end alignments, thus performing better for read lengths of modern NGS machines. BWA-MEM returns a file containing all input reads together with their mapping position in the reference genome, the quality of the mapping and further information defined in the Sequence

Alignment/Map (SAM) format specifications [82]. The accompanying tool set SAMtools (version 1.3.1) is then applied to compress the SAM files to Binary Alignment/Map (BAM) files.

Improving read alignments Variant calling algorithms search for genetic variants by identifying mismatches in the alignments of the sequenced reads. The quality of the resulting variant calls critically depends on the ability of the algorithm to differentiate between sequencing errors and evidence for real genetic variants. In general, the more reads contain a differing nucleotide, the more likely the mismatch is not caused by a sequencing error, but reflects a genetic variant in the sequenced genome. Additionally, the quality of the sequenced bases is taken into account relying on the base quality score computed by the sequencing machine for each base of each read. Both the number of reads and the base quality score can be distorted by systematic errors that need to be corrected by the following two procedures to avoid false-positive or false-negative variant calls.

The number of reads supporting a variant at a certain position can be inflated by duplicate reads that are originating from the same genomic DNA fragment. Duplicates are either introduced during library preparation when polymerase chain reaction is used to amplify DNA fragments before sequencing, or the sequencing machine mistakenly reports the sequence of one DNA fragment as two identical sequences because the readout algorithm considers it as two different fragments. In order to prevent false-positive variant calls resulting from duplicated reads they are marked by the tool MarkDuplicates of the Picard toolkit [83] (version 2.5.0). It identifies duplicate reads by comparing the 5' positions of two reads and marks them if the are identical. Because duplicate reads are ignored during variant calling, the read having the highest quality score within a set of duplicates is not marked as such.

Algorithms used by sequencing machines are prone to systematic technical errors that result in over- or under-estimated base quality scores. Although the exact errors depend on the sequencing machine and the reagents used for the sequencing reaction they are detectable by analyzing covariation in the data. GATK provides tools to perform so-called Base Quality Score Recalibration (BQSR), which identifies bias and corrects the base quality scores while taking into account known genetic variation. Both steps, marking duplicate reads and BQSR are important to improve sensitivity and specificity of the subsequent variant calling step.

Calling variants Variant calling is implemented as a two step process as shown in Figure 2.3a and 2.3b. First, the GATK HaplotypeCaller [84] computes genotype likelihoods for variant positions separately for each individual, then the GATK GenotypeGVCF tool creates a matrix of genotypes for the whole family or cohort at every variant site. Besides being a widely applied method for variant calling, for example in the gnomAD project, especially the use of the GATK HaplotypeCaller offers several advantages over other tools [84]. The authors report that GATK HaplotypeCaller scales efficiently to large sample

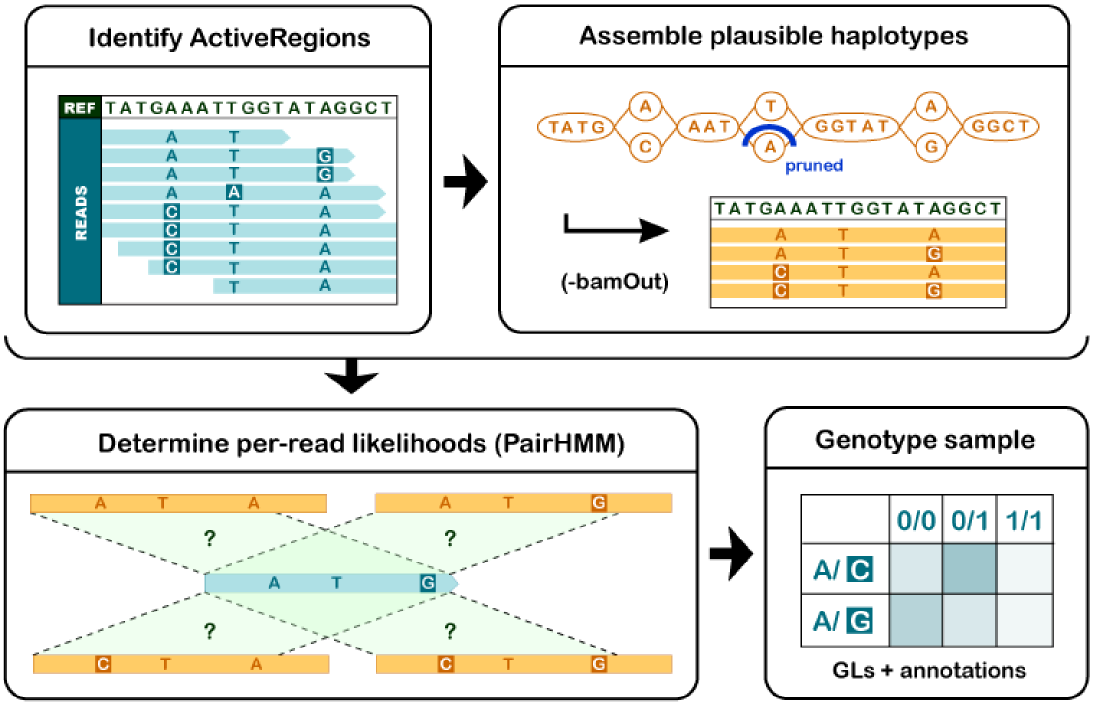


Figure 2.4: Main steps of the HaplotypeCaller algorithm of the Genome Analysis Toolkit. First, the algorithm identifies ActiveRegions in the mapping data where read alignments provide evidence for genomic variation. In the next step, possible haplotypes are created that could have generated the read data in the ActiveRegions. Then, likelihoods of the haplotypes are computed based on the observed read data. Finally, for each identified genetic variant, the most likely genotype is assigned according to the maximum likelihood, which is illustrated by the darkest shade in the matrix in “Genotype sample”. The abbreviation PairHMM stands for Pair-hidden Markov model and GL indicates genotype likelihood. The figure is taken from [84].

sizes without loosing accuracy and that the accuracy of calling insertions and deletions is superior in comparison to other algorithms.

As illustrated in Figure 2.4 the GATK HaplotypeCaller algorithm consists of four main steps. First, it identifies regions varying from the reference genome, so-called ActiveRegions, based on information on mismatches, insertions or deletions in the read alignments. For each ActiveRegion, the algorithm assembles possible haplotypes underlying the mapped reads by building a De Bruijn-like graph. Each haplotype is then realigned by the Smith-Waterman algorithm to detect potentially variant sites. Then, the program performs a pairwise alignment of each read against each haplotype using a pair-hidden Markov model resulting in a matrix of likelihoods of haplotypes given the read data. For each potentially variant site, the program uses this matrix to compute likelihoods of alleles per read, applies Bayes' rule to infer posterior likelihoods of each genotype given the observed read data and then assigns the most likely genotype. As a result, the GATK HaplotypeCaller produces a file in the genome variant call format (gVCF) that contains genotype likelihoods for each variant position and the non-variant regions in between.

The GATK GenotypeGVCF tool takes gVCF files of multiple samples as input and performs joint variant calling to create a matrix of genotypes containing all genomic positions where at least one individual of a family or a cohort carries a non-reference allele. In contrast to single sample variant calling, the resulting genotype matrix is the major advantage of joint variant calling, as it distinguishes for non-variant positions of a sample whether the individual carries two copies of the reference allele or whether no genotype could be determined due to poor sequencing quality. Moreover, variant identification is the more accurate the more samples are available during the variant calling step. The specificity is improved as a higher amount of samples provides a greater ability to detect systematic errors to filter out false-positive variant calls. Beyond that, there is a higher chance to find evidence for low-frequency variants increasing the sensitivity. The GATK GenotypeGVCF tool generates a file in the variant call format (VCF) [85] that represents the genotype matrix with all identified variant positions as rows and samples as columns. In addition, the family or cohort VCF file includes quality scores computed on variant and genotype level.

Variant quality control After calling variants, it is necessary to conduct quality control steps to remove false-positive variants from the obtained call set. Instead of directly using the variant quality scores computed by the GATK GenotypeGVCF tool, best practice guidelines recommend a filtering technique called Variant Quality Score Recalibration (VQSR) [78]. This machine learning based method takes the provided quality scores as features to model profiles of known variants in the call set. Subsequently, a meta-score is calculated, which provides a continuous estimate of the probability that a variant is true. The sensitivity and specificity of the variant call set can then be controlled by setting a sensitivity threshold value that specifies the percentage of known variants that are to be retained after filtering. Choosing a higher percentage as threshold value improves sensitivity, while a lower percentage increases specificity. Figure 2.5 illustrates the concept of VQSR based on artificial data. For the analysis of the WGS family data set and the WES cohort data set a threshold of 99.5% is chosen, which means that 99.5% of all known variants are contained in the filtered set. The threshold is chosen close to 100% because in the clinical context it is more serious to miss a variant classified as a false-negative than to report false-positive variants.

In order to further improve the quality of the variant call set, the quality of individual genotypes is assessed by means of the depth of coverage (DP) and the genotype quality (GQ) value. The DP value indicates the number of reads at the variant position. The GQ value is a quality score ranging from 0 to 99 defined as the difference between the Phred scaled genotype likelihood of the second most likely genotype and the Phred scaled genotype likelihood of the most likely genotype. Following the recommendations of Carson *et al.* [86] genotypes are discarded if $\mathrm{GQ} < 20$ to achieve a 99% confidence of the remaining genotype calls. Genotypes are also rejected if $\mathrm{DP} < 8$ to ensure that the probability of a heterozygous genotype appearing as a homozygous genotype by random chance is smaller

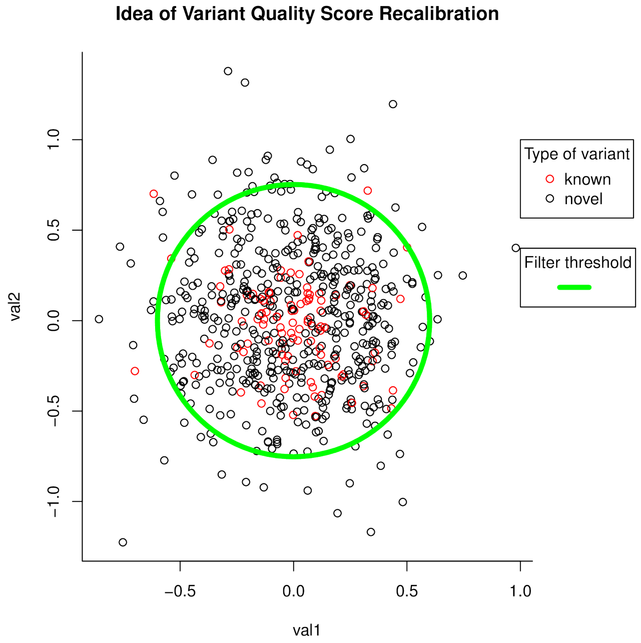


Figure 2.5: Visualization of the concept of Variant Quality Score Recalibration. The dots represent identified variants. They are plotted according to the values of two variables, $val1$ and $val2$, which correspond to two quality metrics. For this figure, the values of $val1$ and $val2$ are following a normal distribution centered at 0 with a standard deviation of 0.3 and 0.4 respectively. Dots are highlighted in red if a number drawn at random from an equal distribution between 0 and 1 is lower than 0.4 when multiplied with $val1$ and when multiplied with $val2$. Thus, dots in the center are more likely to be colored red, which corresponds to known genetic variants whose values tend to be located at the center of the distribution of the respective quality metric. The green circle visualizes the meta-score threshold corresponding to the chosen sensitivity threshold. It separates the set of variants in a way that the proportion of known (red) variants within the circle is equal to the selected sensitivity threshold. All variants within the circle will be kept and all variants outside the circle will be filtered out.

than 1% when assuming a two-tailed binomial model where the reference and the alternate allele have a 50% chance of being in each read.

The reliability of identified insertions and deletions decreases with their length especially in repetitive regions of the genome because sequencing reads are only 150 base pairs long. To avoid that false-positive calls of InDels cause false-positive signals when searching for new disease genes, all insertions equal or longer than ten base pairs and all deletions removing more than nine base pairs are only kept in the cohort analysis when occurring in known disease-related genes.

2.3.2 Sample quality control

With the many steps from taking a blood sample of a patient to readily processed sequencing data, a multitude of different errors risks to be introduced into the data set. DNA samples might be contaminated with DNA of laboratory staff or other patient samples. Samples might be assigned to other family members or completely different individuals or two samples of the same individual are sequenced as samples of two different persons. Library preparation of the sample DNA might fail or other technical errors occur during the sequencing reaction resulting in low quality sequencing data. It is important to remove low quality samples or fix the underlying issue to prevent corruption of the downstream analysis. To detect and resolve inconsistencies in the sequencing data, the tool Peddy (version 0.4.2) [87] is used to assess the quality of the samples after generating the initial variant call set. Based on a set of $23,770$ frequent biallelic SNVs that were identified in the 1000 Genomes Project, Peddy computes several quality measures.

Sequencing depth and call rate Peddy determines the median sequencing depth of the preselected SNVs and summarizes the fraction of SNVs with sufficient read coverage to call a genotype as the call rate. Samples having a median coverage below 30 or a call rate below 90% indicate low quality of the underlying sequencing, which increases the risk of missing genetic variants when keeping them in the analysis. Sequencing of such samples should be repeated if no disease-causing variant was found so far and if sufficient patient material is available. For my analyses, I remove low quality samples from the WES cohort. The distribution of sequencing depth and call rate in the WES data collection and the number of failed samples are presented in Section 3.3.1.

DNA contamination Potential contamination of DNA is detected by assessing the distribution of the fraction of alternate alleles per heterozygous genotype. A deviation of a binomial distribution with $p = 0.5$ indicates that more than two alleles were sequenced and thus the DNA is likely contaminated with other DNA. To measure the difference from the expected distribution, Peddy computes the interdecile range of the observed distribution. The more the binomial expectation is violated, the higher is the interdecile range value. To identify DNA contamination in the WES data set, I select all samples having an interdecile range value higher than 0.3 for further examination by the more sophisticated but also more time consuming tool VerifyBamID [88]. Its algorithm computes likelihoods that varying contamination levels have generated the observed sequencing data and then determines the most likely contamination level. When a sample exhibits a substantial level of DNA contamination the data is not included in further analyses and, whenever possible, a new blood sample is taken, as it is usually not possible to determine when the contamination occurred.

Relationship coefficient and sex prediction In order to inform about relationships between the individuals of a family or a cohort, Peddy computes pairwise relationship coefficients using the following formula:

$$\frac{Het_{i,j} - 2 \cdot N_{IBS0}}{min(Het_i, Het_j)} \tag{2.1}$$

where i and j represent the indices for each individual, Het_i is the count of sites where individual i is heterozygous, Het_j is the count of sites where individual j is heterozygous, $Het_{i,j}$ is the count of sites where individuals i and j are both heterozygous, and N_{IBS0} is the count of sites at which individuals i and j share no alleles, referred to as zero identity-by-state (IBS0) sites. The coefficient can adopt 1.0 as the maximum value in the case of genetically identical individuals, which is the result either for identical twins or for duplicate sequencing of the same individual. In order to differentiate parent-offspring from sibling-sibling pairs, both having an expected coefficient of 0.5, the count of IBS0 sites has to be taken into account. Since both alleles of a child originate from either father or mother, there are no positions where no allele is shared with either parent and thus the count of IBS0 sites should be 0. For sibling-sibling pairs, IBS0 sites exist, for example, when both parents are heterozygous and one sibling inherits the reference alleles and the other the alternate alleles.

In addition to the relationship coefficient, Peddy examines the ratio of heterozygous to homozygous variant genotypes on the non-pseudoautosomal regions of the X chromosome to predict the sex of the sequenced individuals. Due to the hemizygosity of males all variant calls in these regions are reported as homozygous variants resulting in an expected ratio of 0. For females, a minimum ratio of 1.0 is expected, which can be lower if there is some degree of consanguinity in the family. Using the computed relationship and the predicted sex assignment, several anomalies in sequencing data can be identified, resolved, and also be leveraged for further analysis.

- **Unexpected sex** If the predicted sex deviates from the sex stated in the metadata of the individual, the correctness of the metadata should be validated first. However, it is also possible that the sex prediction is incorrect. I have observed ratios of 0.33 and higher, which could clearly be assigned to female individuals although they were classified as males by Peddy. If both, the sex prediction and the metadata of the individual are correct, identified relationships to other family members can help to identify the individual that was actually sequenced. Otherwise, the sequenced data should be discarded and replaced by sequencing a new sample of the individual of interest.

- **Duplicate samples** If a relationship coefficient close to 1.0 is observed for two individuals that are not identical siblings, the reason for the duplication needs to be investigated. In case of an intentional repetition of the sequencing, the two read sets can be pooled to increase sequencing depth, otherwise it has to be found out beforehand who has been sequenced twice. If the two individuals that should have been sequenced are a male and a female, the sex prediction can help to determine which of the individuals was sequenced twice given that not samples of a third individual were involved in the duplication. However, sequencing another sample of one of the both individuals that might be duplicated or a sample of a close relative of one of them is the most accurate way to resolve the origin.

- **Inconsistent pedigrees** Relying on the relationship coefficient and the count of IBS0 sites, all expected parent-offspring and sibling-sibling pairs can be verified in trios or more complex pedigrees. For this purpose, a maximum count of IBS0 sites of 50 is set as threshold to identify parent-offspring pairs when the relationship coefficient is close to 0.5. Although no IBS0 sites are expected, they can result from sequencing artifacts that are targeted by variant quality control described in Section 2.3.1. Pedigree inconsistencies can, for example, arise through unintentional swapping of the patient's mother and father, through swapping the child for one of its parents, or through the assignment of unrelated individuals as parents. While accidental permutations can be corrected by swapping the labels of the underlying data sets, individuals wrongly assigned as parents must be excluded from further analysis to prevent incorrect variant filtering by segregation.

- **Unknown relationships** As Peddy computes relationship coefficients for all pairs of individuals in a cohort, it can also identify unknown relationships. A minimum coefficient of 0.2 is required for assuming a familial relationship. If related individuals are patients with similar disease phenotypes they are grouped together in a subanalysis as there is an increased chance that the same genetic defect is causative, which can be found more easily due to the increased power of segregation filtering.

2.3.3 Population-specific allele frequencies

A considerable proportion of the patients in the WES cohort lives in or comes from Middle Eastern countries (see Section 2.1.1). As Middle Eastern populations are underrepresented in the 1000 Genomes Project and the gnomAD, candidate filtering will fail to remove genetic variants that are rare in European populations but common in Middle Eastern populations. To prevent such population-specific and thus non-pathogenic variants from inflating candidate lists, I have developed a routine to detect and remove them. First, the cohort is stratified into different populations without any prior knowledge, then the maximum allele frequency across all detected populations is assigned to each variant.

Population stratification The tool ADMIXTURE (version 1.3.0) is applied to stratify the WES cohort according to the different ethnic origins of the individuals [89]. For a given number of populations, its algorithm estimates the proportion of ancestry from each contributing population over the individual's entire genome. As ADMIXTURE requires unrelated individuals as input, a maximum subset of unrelated individuals is extracted from the cohort before determining the number of underlying populations and assigning each individual of the cohort to one of the identified populations.

To produce a set of unrelated individuals, a relationship network is created with individuals as nodes connected by edges if they are related to each other. If the relationship coefficient computed by Peddy as described by equation (2.1) is 0.2 or higher, two individuals are considered as being related with each other. A maximum subset of unrelated

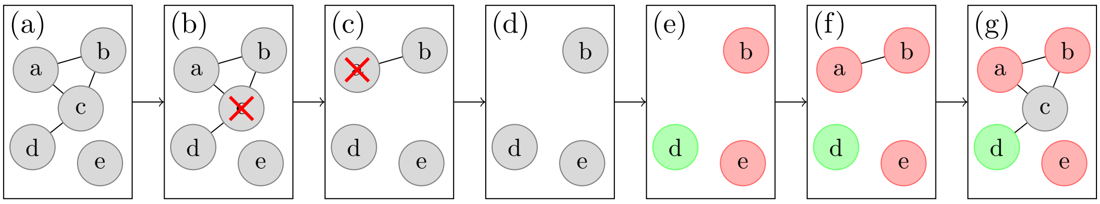


Figure 2.6: Steps of population stratification. Panel (a) to (d) visualize the generation of the maximum set of unrelated individuals for an exemplary cohort of five individuals. Panel (e) illustrates the initial population stratification through ADMIXTURE for two identified populations: individual d belongs to the green and individual b and e belong to the red population. Panel (f) and (g) demonstrate the expansion of the population assignment along the edges of the relationship network. Individual c cannot be assigned to one of the two populations as it is related to individuals of the red and the green population.

individuals is then determined by removing nodes with the highest degree until the highest degree equals 0. Figure 2.6 (a) shows the relationship network for an exemplary cohort of five individuals. According to the described method individual c and a are removed resulting in a network consisting of three unrelated individuals (see Figure 2.6 (b) to (d)).

Before carrying out ancestry estimation by ADMIXTURE, a set of representative genetic variants is selected from all identified variants in the unrelated cohort. For this purpose, VCFtools (version 0.1.14) [85] first extracts biallelic SNVs with a minimum allele frequency of 5%. Then, PLINK (version 1.9) [90] is applied to the variant set to remove all variants in linkage disequilibrium (LD). To identify variants in LD, PLINK screens for pairs of variants with a squared correlation r^2 greater than 0.1 in windows of 50 kilobases length by shifting this window by ten variants in each step.

For different numbers of assumed populations in the input cohort, ADMIXTURE estimates the corresponding ancestry proportions for each individual, and then computes a cross-validation error across all iterations [91]. The number of populations resulting in the minimum cross-validation error represents the estimated number of contributing ancestries. By k-means clustering of the ancestry proportions, each individual is assigned to one of the identified populations as visualized in Figure 2.6 (e). The population background of individuals initially excluded from the analysis is derived from the relationship network as shown in Figure 2.6 (f) and (g). Individuals with relationships to individuals of different populations remain unassigned.

Maximum allele frequency calculation All variants surpassing a given allele frequency threshold in at least one identified population are to be filtered out as population-specific variants defined by:

$$\underset{p \in P}{\exists} AF_p > AF_{max} \tag{2.2}$$

where p is a population of the identified set of populations P, AF_p is the allele frequency of the variant in p and AF_{max} is the chosen maximum allele frequency threshold. For this purpose, allele frequencies of each genetic variant are computed separately within each detected populations. A call rate of at least 10% is required to generate a valid value for a population. The maximum of the computed allele frequencies per variant is then stored in a separate VCF file.

Ethnicity prediction Available data on the origin of some of the individuals is used to gain insight into the ethnicity of the identified populations. For each country that is the reported origin of at least ten individuals, it is examined whether more than half of the individuals belongs to one of the identified populations. These population-specific countries are then used to manually assign each population to a world region. The population to region assignment enables the comparison of the reported to the predicted ethnic origin in order to identify potentially mixed up DNA samples.

2.3.4 Variant annotation

After removing low quality variants and samples from the variant call set of a family or an entire cohort, several features are annotated to each variant, which serve as criteria to filter for potentially pathogenic variants. To obtain a comprehensive and informative description of the variants, annotations of several different tools are incorporated. The Variant Effect Predictor (VEP, version 95) [92] and GEMINI (version 0.30.1) [93] annotate the functional impact of a variant and its allele frequency. The deleteriousness of variants is determined by the tool InterVar (version 2.0.2) [94] that aims to classify variants as pathogenic or benign and by the Combined Annotation Dependent Depletion (CADD, version 1.4) [95] framework that assigns a pathogenicity score to each variant.

Variant Effect Predictor and GEMINI At its core, the Ensembl VEP determines the functional consequence of a variant with respect to the gene or genes at the locus where it is situated [92, 96]. Furthermore, it supports the annotation of a multitude of other predefined or user-specified features. GEMINI transforms the resulting annotated VCF file into a database file, which accelerates and simplifies the subsequent selection of variants of interest.

Figure 2.7 shows an overview of the main functional consequences that are annotated by the VEP. Variants affecting splice-sites, causing non-synonymous amino acid changes, a shift of the reading frame or a premature stop codon likely alter the protein sequence and are thus most relevant when searching for pathogenic variants. Such potentially deleterious variants are usually rare while technical error rates are uniformly distributed resulting in an increased effective error rate especially for LoF variants [97]. To adjust for the increased effective error rate, the VEP plugin Loss-Of-Function Transcript Effect Estimator (LOFTEE) is used to assess the impact of LoF variants [40, 65]. LOFTEE incorporates various types of information including the location of the variant in the transcript, the strength of splice sites and the ancestral state of the allele to flag variants either as low or

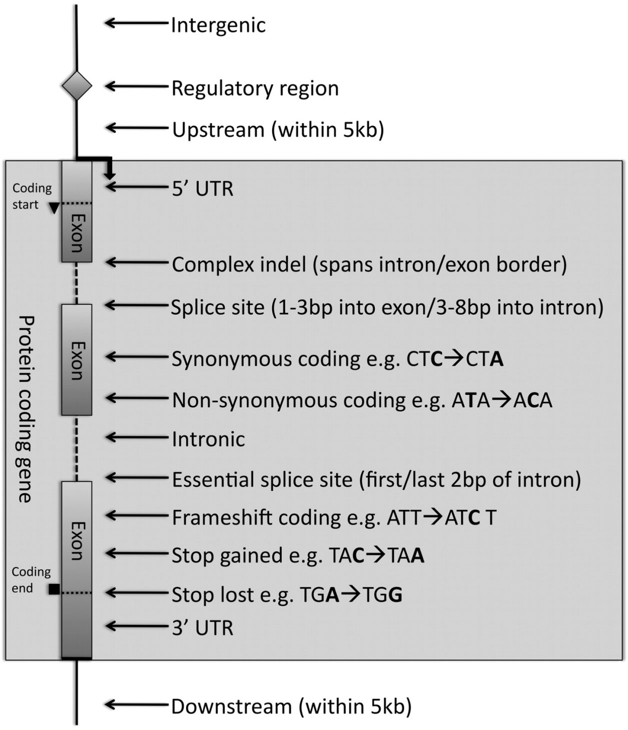

Intergenic
Regulatory region
Upstream (within 5kb)
5' UTR
Coding start
Exon
Complex indel (spans intron/exon border)
Splice site (1-3bp into exon/3-8bp into intron)
Exon
Synonymous coding e.g. CTC→CTA
Non-synonymous coding e.g. ATA→ACA
Protein coding gene
Intronic
Essential splice site (first/last 2bp of intron)
Exon
Frameshift coding e.g. ATT→ATC T
Stop gained e.g. TAC→TAA
Coding end
Stop lost e.g. TGA→TGG
3' UTR
Downstream (within 5kb)

Figure 2.7: Main variant consequences annotated by the Variant Effect Predictor. The consequence of a variant is determined in relation to the gene in which the variant is located or with respect to the closest gene. The abbreviation UTR stands for untranslated region, indel refers to an insertion or deletion, kb is kilobases, bp is base pairs and miRNA is micro ribonucleic acid. Figure adapted from [96].

high confidence LoF. To facilitate the interpretation of stop-gain variants, their annotation is completed by a score that is computed by the NMDetective and reflects the probability of triggering nonsense-mediated mRNA decay (NMD) [98].

Variants not directly changing the amino acid sequence are harder to interpret regarding their functional impact. The SpliceAI plugin extends the search for splicing defects to intronic and synonymous variants based on a deep neural network that predicts splice junctions from pre-messenger RNA (mRNA) transcript sequences [99]. Variants in 5' UTRs can have an effect on transcription and translation when they affect uORFs. To find such effects, additional information is added to variants creating a new upstream start-codon or removing an existing upstream stop-codon by the 5' UTR annotator plugin for the VEP [21]. After the annotation by the VEP, GEMINI creates a variant database and adds allele frequencies provided by the 1000 Genomes Project and the gnomAD consortium. For the cohort analysis, GEMINI further annotates the maximum allele frequency across the identified populations as described in Section 2.3.3.

InterVar In 2015, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) [69] published guidelines to standardize the interpretation of genetic variants. For this purpose, the authors defined a set of 28 criteria representing different levels of evidence either for the pathogenicity or benign impact of a variant. Additionally, rules are provided for combining the criteria to classify genetic variants as pathogenic, likely pathogenic, likely benign, benign or having uncertain significance.

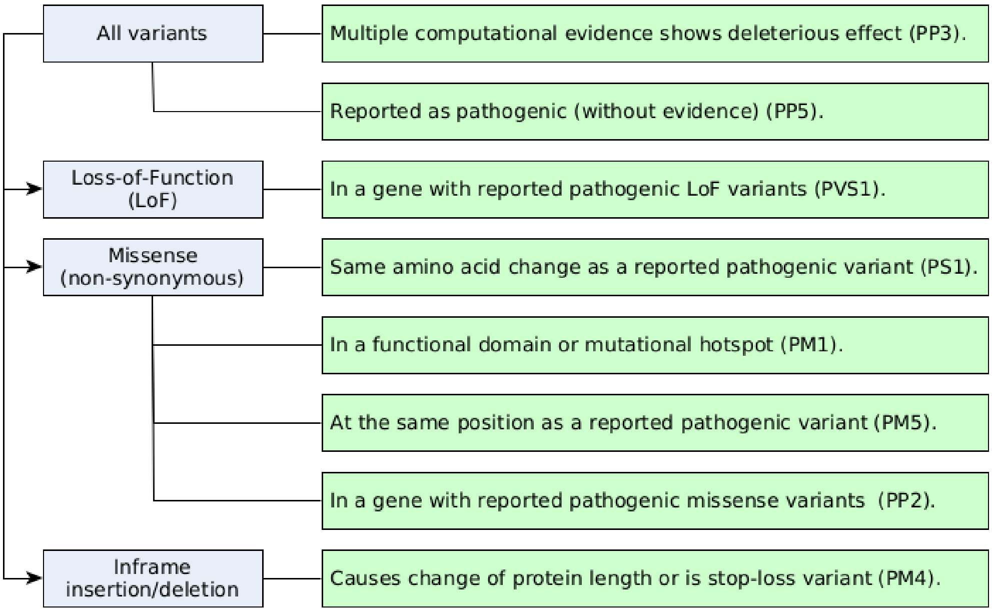

All variants
Multiple computational evidence shows deleterious effect (PP3).
Reported as pathogenic (without evidence) (PP5).
Loss-of-Function (LoF)
In a gene with reported pathogenic LoF variants (PVS1).
Missense (non-synonymous)
Same amino acid change as a reported pathogenic variant (PS1).
In a functional domain or mutational hotspot (PM1).
At the same position as a reported pathogenic variant (PM5).
In a gene with reported pathogenic missense variants (PP2).
Inframe insertion/deletion
Causes change of protein length or is stop-loss variant (PM4).

Figure 2.8: Subset of the ACMG/AMP criteria annotated by InterVar. Some criteria are only assigned to specific variant types. The abbreviations of the criteria are taken from the original ACMG/AMG publication [69]

.

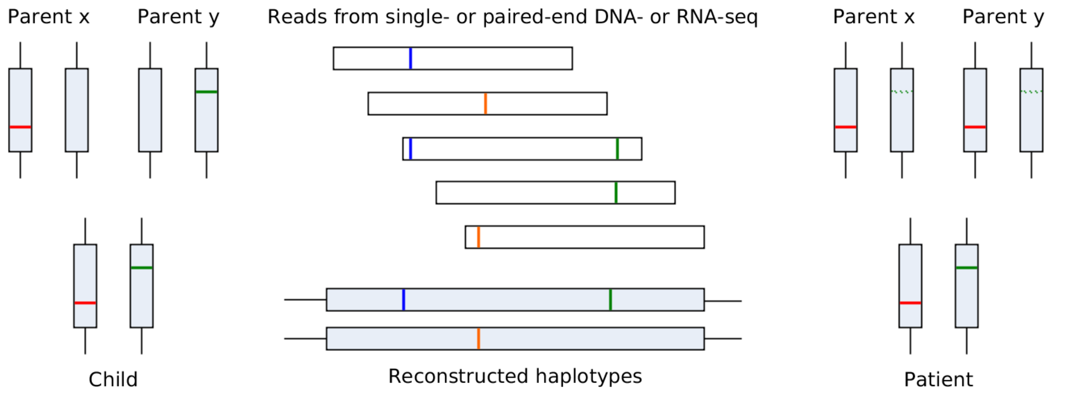
The tool InterVar [94] automates the annotation of 18 of the 28 criteria by extracting information from public data sets, such as ClinVar and gnomAD. Figure 2.8 shows an overview of the eight criteria for pathogenicity that are implicated in the variant prioritization approach described in Section 2.3.6. The criteria PS1, PM5, PP2 and PP5 are based on the ClinVar database, PM1 uses information given by dbNSFP [100, 101] and InterPro [102], and PM4 relies on the RepeatMasker track [103] of the UCSC Genome Browser [104]. Criterion PVS1 results from a combination of the LoF tolerance estimation provided by the gnomAD (see Section 2.2.2) and the presence of pathogenic LoF variants reported in the ClinVar database. Although automation through InterVar simplifies the ACMG/AMP classification while increasing standardization, it depends heavily on the availability and quality of entries in the underlying databases. Therefore, the annotated criteria should rather be considered as an indication of the variant impact then as a final classification.

Combined Annotation Dependent Depletion CADD is a framework that assigns scores to genetic variants to estimate their deleteriousness [95, 105]. Its underlying machine learning model is based on more than 60 genomic features and was trained to differentiate between frequent benign variants and simulated *de novo* variants that could be neutral or deleterious. The pure statistical approach gives CADD an advantage over data based methods, such as InterVar, which are heavily relying on curated sets of pathogenic and benign variants. CADD does not suffer from implicit ascertainment bias nor the varying quality and limited scope of positive and negative evidence for pathogenicity as discussed for ClinVar in Section 2.2.3. In contrast to other scores, CADD scores can be computed for all types of variants and can thus rank any given variant set by predicted deleteriousness. However, it was shown that the CADD framework is most powerful to evaluate protein-coding variants while the scoring of non-coding is less reliable [106]. Although, CADD does not explicitly consider the impact of coding variants on intra-chain amino acid interactions or on the overall protein structure, it performs only slightly worse in a comparison to a tool called VIPUR that performs structural modeling to predict deleteriousness [107]. On a set of 950 human variants with 664 deleterious and 286 neutral variants, VIPUR achieves an accuracy of 0.782 while CADD achieves 0.781. The difference is more pronounced for the balanced accuracy reported with 0.752 for VIPUR and 0.695 for CADD. The inferior performance when weighting the performance on neutral variants equally is in accordance with the result of another study that showed that CADD scores tend to overestimate the deleteriousness of benign variants while accurately predicting pathogenic variants [108]. Taken together, CADD scores are well suited to prioritize variants in clinical exome sequencing data because any genetic variant can easily be scored, pathogenic variants are accurately identified and the tendency to report false-positives prevents disease-causing variants from being systematically missed.

2.3.5 Search for compound effects

In order to find the genetic defect underlying an autosomal recessive disease in a non-consanguineous family, pairs of harmful heterozygous variants affecting one allele each are screened first [109]. Their compound effect is able to impede the function of the gene product in the same way as homozygous deleterious variants, which are rather found in families with a history of consanguinity. However, the identification of these so-called compound heterozygous variant pairs is not readily possible in NGS data, because information on the parental origin of the sequencing reads is usually lost during library preparation. Thus, algorithms are required to evaluate whether two heterozygous variants are located on the same or opposite alleles referred to as phasing.

As the number of possible compound heterozygous combinations increases quadratically with the number of rare heterozygous variants per gene, there can be many more pairs of potentially compound heterozygous variants then single hetero- or homozygous candidate variants per patient. The difference is the more pronounced the more rare heterozygous variants are found, primarily depending on the specified maximum allowed allele frequency. Therefore, it is not only relevant to identify compound heterozygous pairs, but also to reduce the set of candidate pairs by excluding those located on the same allele or those being non-pathogenic according to the observed inheritance pattern in the family. Targeting phasing of rare variants in regions of interest and the combination of different phasing strategies to achieve best possible results are additional requirements in the clinical context. As existing phasing tools do not meet all of the required features, Paul Hager and I developed and published a new tool called SmartPhase [110, 111].



(a) Trio phasing. (b) Read-based phasing. (c) Innocuous labeling.

Figure 2.9: Strategies of SmartPhase for resolving heterozygous variant pairs. Variant pairs are phased using parental genotypes (a) or sequencing reads (b) or are labeled as innocuous (c). The dashed green lines in panel (c) show the possible genotypes of the parents assuming that the child is compound heterozygous for the exemplary variant pair.

Figure 2.9 visualizes the implemented strategies to resolve heterozygous variant pairs. SmartPhase is able to perform trio phasing and read-based phasing and can combine both strategies if both genotypes of the parents and reads from single- or paired-end DNA or RNA sequencing are given (see Figure 2.9a and 2.9b). If parental genotypes are available and both parents are healthy, SmartPhase can additionally identify non-pathogenic pairs when the inheritance pattern contradicts the observed phenotypes (see Figure 2.9c). This is especially useful in the case of variants heterozygous in both parents and its offspring, which cannot be phased, but can be classified as innocuous because there is either compound heterozygosity in one of the parents too, or both variants were inherited on the same allele. Furthermore, SmartPhase is able to incorporate the haplotypes generated by the GATK HaplotypeCaller during variant calling (see Section 2.3.1) to phase variants not resolved by trio or read-based phasing or innocuous labeling. SmartPhase returns a bitflag and a confidence score for each input variant pair to inform in detail about the phasing result and its reliability.

SmartPhase was validated on simulated data and a subset of the WES cohort described in Section 2.1.1. We showed that SmartPhase generates error-free predictions when using a threshold of 0.34 for the confidence score to discard low quality phasing predictions. In comparison to WhatsHap [112], another phasing tool combining trio and read-based phasing, we demonstrated that SmartPhase is markedly faster and resolves more pairs when parental genotypes are provided.

2.3.6 Variant prioritization

While the identification and quality control of genetic variants is a standardized process to a large extent, the prioritization of variants according to their potential pathogenicity depends much more on the underlying disease type and the research objective. For rare diseases, especially if they manifest in early childhood, it is generally assumed that a single rare genetic defect causes the disease in a fully penetrant manner [113]. In common diseases, single disease-causing variants are rarely observed and it is more likely that a mixture of common and rare variants determines the predisposition to a disease, which then manifests itself through mechanistically diverse interactions with environmental factors. Regarding the research objective, exploration of specific diseases in firmly defined cohorts primarily aims to describe the complete diversity of genetic findings to develop or extend a comprehensive picture of the underlying pathomechanisms. In contrast, clinical cohorts are consistently growing by including new patients and rather aim to find genetic variants that can be translated to treatment after verifying pathogenicity. As resources for functional validation are limited regarding time and cost, criteria for prioritization tend to be chosen more strictly to select variants and genes whose further evaluation is promising. For my project, which focuses on the variety of rare IEI in clinical context, I am thus searching for rare deleterious variants as monogenic and fully penetrant causes of the observed pathophenotypes. The following paragraphs describe the criteria to screen for recessive effects in the form of homozygous variants or compound heterozygous variant pairs and for dominant effects either inherited as heterozygous allele or caused by a *de novo* mutation. Further, it is specified how the criteria are applied to discover so far unknown candidates for disease genes in the patient cohort.

Frequency criteria Allele frequencies and genotype counts of the 1000 Genomes Project and the gnomAD data set are used for filtering for rare alleles. This step depends on the assumption that the individuals included in these data sets do not suffer from IEI. The assumption is supported by the fact that both data sets aim to represent the healthy population and that it is unlikely that individuals with IEI have been recruited as healthy participants due to the severity of the pathophenotype. To ensure that variants that are pathogenic but not fully penetrant are not overlooked, the control subset of gnomAD is used for applying the following filter criteria in genes causative for IEI, otherwise the whole data set is used.

Based on the assumption that heterozygous variants causing dominant effects are not part of the reference data sets, all heterozygous variants identified in the 1000 Genomes Project or gnomAD are filtered out. Homozygous variants are kept as potential candidate if no homozygous genotype is reported in gnomAD and the allele frequency is below 5% in the 1000 Genomes Project and below 0.5% in gnomAD. The threshold is less stringent for the 1000 Genomes Project because of the much smaller number of sequenced individuals it is based on. As it is not possible to assess the prevalence of variant pairs because frequencies are given separately for each genotype, the filter criteria for homozygous variants are equally applied for heterozygous variants that are part of potentially compound heterozygous pairs. This choice is based on the assumption that heterozygous variants in compound heterozygous loci would have the same effect as individual homozygous variants. For the WES cohort, the population-specific allele frequencies computed as described in Section 2.3.3 are used to remove all alleles more frequent than 10% in any of the identified populations.

Impact criteria Using the annotations described in Section 2.3.4, genetic variants are selected as potential pathogenic candidates when they directly alter the protein sequence or likely affect splicing, transcription or translation. Protein-altering variants, in homozygous configuration or heterozygous either as *de novo* variant, dominantly inherited or part of a compound heterozygous pair, are considered as candidates when their CADD score is equal or higher than 15 and InterVar annotated at least one of the pathogenicity criteria shown in Figure 2.8. LoF variants in genes not known to cause IEI are excluded when they are flagged as low confidence variants by LOFTEE. The CADD and InterVar criteria are not applied for intronic or synonymous variants for which SpliceAI predicted a splicing defect with high probability, and not for 5' UTR variants that have an effect on a uORF. Each uORF created or disrupted by a heterozygous or homozygous variant in the 5' UTR is instead filtered for having a moderate or strong consensus with the Kozak sequence, a frequent sequence motif in eukaryotes framing the start-codon AUG [114, 21]. Moreover, such newly created or disrupted uORFs must not have a stop-codon in the UTR, thus generating overlapping reading frames, which may reduce the expression of the encoded protein.

Segregation criteria Three criteria ensure that variants are excluded that do not segregate with observed pathophenotypes under the assumption of dominant-recessive inheritance. Heterozygous variants are excluded as candidates for dominant effects or as part of potentially compound heterozygous pairs, if a homozygous genotype of the variant allele occurs in any of the healthy individuals in the cohort. Based on the assumption that a disease affecting multiple individuals of a family is caused by the same underlying genetic defect, all patients must be carries of a potential candidate variant, otherwise it is excluded. Heterozygous variant pairs, which are on the same allele or labeled as innocuous in any affected individual are excluded to limit the number of heterozygous variant pairs to those where all affected individuals could be compound heterozygous. For homozygous variants or heterozygous variant pairs in trios, it is not necessary that each parent inherited one of the variant alleles as deviations from this, so-called Mendelian errors, can be the consequence of incorrect genotype calls in the parents resulting in false-negatives when discarding such variants.

Candidate gene selection In order to discover gene defects that have not yet been described to cause IEI, additional impact and segregation criteria are applied after aggregating annotations on variant level. To include only the most promising protein-altering variants, a variant is selected if it is part of the top 15 variants in at least one of the individuals carrying the variant when ranking by CADD score. A pair of heterozygous variants must be phased by SmartPhase (see Section 2.3.5) as being compound heterozygous in at least one of the individuals in which the variant combination was found.

After the additional variant filtering, genes are sought that are hit by deleterious mutations in at least two unrelated patients. This criterion strengthens the hypothesis for a disease association and is therefore used to restrict lists of potentially relevant genes to the most promising candidates for subsequent long-term research projects that aim to link the gene mechanistically to the suspected disease. It is implemented as a filter for genes where the same or different, recessive or dominant candidate variants are found in patients from at least two different families. Additionally, it is required that none of the patients has been diagnosed genetically or that only variants in that gene have been reported as genetic diagnosis. If such genes harbor variants of healthy individuals that fulfill the defined frequency and impact criteria, it is no longer considered a candidate gene, however. Genes are only considered as candidates for dominant effects if at least one of the selected heterozygous variants is *de novo* according to the sequencing data in at least one trio or was inherited at least once from diseased parents. The remaining candidate genes are finally annotated with known gene-disease relationships taken from OMIM and immune mouse phenotypes identified by the IMPC to quickly identify genes potentially associated with immune diseases.

Results 3

The comprehensive analysis of the entire collection of WES data at the Dr. von Hauner Children's Hospital and the analysis of a single family with WGS data yielded a wide range of results presented in the following five sections. First, the workflows are described that were implemented and extended in the course of the sequencing data analysis (Section 3.1). The next section covers the analysis of the WGS data, which resulted in the discovery of the disease-causing variant in the index patient of the family (Section 3.2). The remaining three parts present the results on the WES data divided into an examination of the intermediate results (Section 3.3), a characterization of the generated candidate lists (Section 3.4) and a detailed description of identified pathogenic variants that provide new insights into the etiology of IEI (Section 3.5).

3.1 Implemented workflows

To implement the analysis steps described in Section 2.3 as structured and reusable pipelines, I used the Konstanz Information Miner (KNIME). This open source platform is designed to create flexible analysis pipelines for various applications in data science [115]. KNIME provides a graphical user interface that enables users to easily create workflows from individual data processing modules, so called nodes. The user community extends the standard set of nodes for basic operations on tabular data continuously to make KNIME usable for various applications. To also make the standard steps of NGS pipelines available as KNIME nodes, I developed the extension KNIME4NGS in collaboration with many other colleagues. The subsequent steps of variant prioritization, which are specifically tailored to the analysis of the entire WES data set, are also implemented as a KNIME workflow, referred to as Candidate Identification Pipeline.

3.1.1 KNIME4NGS

The KNIME4NGS extension [116] was developed to enable the straight-forward generation of workflows for versatile NGS data analyses using the intuitive and user-friendly KNIME platform. We implemented a set of 42 nodes each providing a certain functionality for building customized NGS data analysis pipelines. Many of the nodes are wrappers of software, such as BWA or GATK, to make these commonly used tools available in the KNIME environment. The input and output of all developed nodes are in tabular form to enable the use of our nodes together with the default nodes of KNIME or nodes of other extensions. In addition to the nodes, KNIME4NGS offers a dedicated binary manager to simplify the installation of the underlying software binaries. Since the analysis of NGS data involves the application of multiple interdependent tools on large data sets, the probability of spurious premature termination of individual processes increases. In order to minimize manual control and intervention, we have designed the High-Throughput Executor (HTE) as an extension of the standard KNIME node model. For each analysis step, the HTE records its completion state in a database and restarts the process if it failed for reasons like insufficient memory or other randomly occurring errors. The established database can then be used to identify and eliminate error-prone steps in the workflows. The HTE further ensures that successfully completed analysis steps will not be re-executed when the entire workflow is run again.

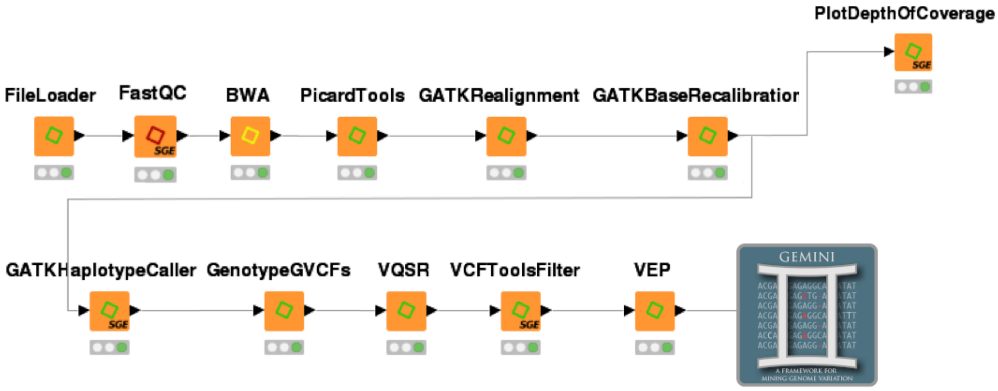


Figure 3.1: KNIME pipeline for whole-genome sequencing data analysis. The pipeline covers all steps from the import of FASTQ files to the generation of the GEMINI database. All KNIME nodes shown are part of KNIME4NGS. Figure taken from [117].

In the context of my thesis, KNIME4NGS was used to process the WGS data (see Section 2.1.2) as shown in Figure 3.1. The WES data was processed by an already established script-based pipeline at the Dr. von Hauner Children's Hospital that covers the same steps up to variant calling. In the following section, I will introduce how I extended this pipeline with a KNIME pipeline for variant prioritization.

3.1.2 Candidate Identification Pipeline

Periodic re-analysis of sequencing data increases the diagnostic rate as new algorithms and growing knowledge help prioritizing variants [118]. Nevertheless, the diagnostic rate is still limited, which motivates the search for deleterious mutations in genes that are not yet associated with the observed disease phenotype, so called candidate genes. As it is hardly feasible to review the variant lists for each patient separately after each re-analysis iteration, I created a workflow that takes a cohort VCF as input and generates two types of overviews of relevant genetic variants. First, the workflow generates lists of all rare potential deleterious variants in known disease-associated genes that serve as a reference to review the cohort for variants especially in recently identified novel disease-associated genes. Second, it reports novel candidate genes that are hit by potentially pathogenic variants in multiple undiagnosed patients. The latter list serves as a starting point to collect evidence for each gene that supports or contradicts an association with the pathophenotypes of the affected patients in order to decide for which candidate genes further functional studies should be performed. As described in Section 2.3.6, the prioritization strategy for variants within known disease genes and within potentially relevant genes are similar regarding the information that is used for filtering, but thresholds are more stringent when searching for candidate genes.

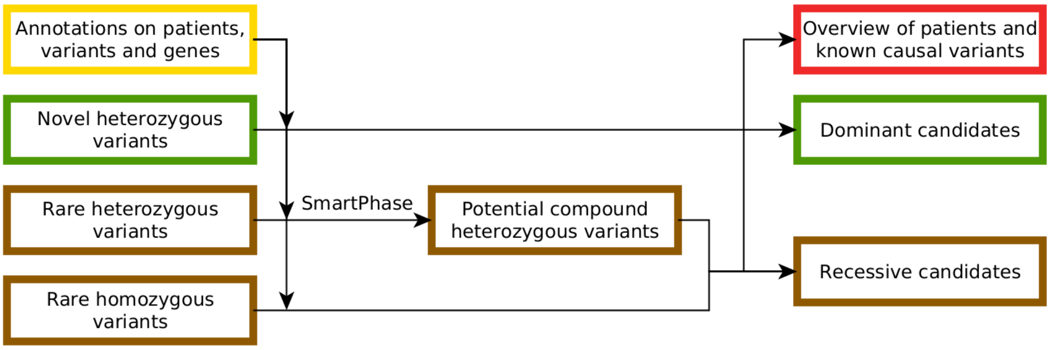

Annotations on patients, variants and genes
Novel heterozygous variants
Rare heterozygous variants
Rare homozygous variants
SmartPhase
Potential compound heterozygous variants
Overview of patients and known causal variants
Dominant candidates
Recessive candidates

Figure 3.2: Schematic overview of the Candidate Identification Pipeline. Each box represents a group of nodes in the underlying KNIME workflow that performs a specific function. Nodes in green and brown boxes are responsible for the identification of dominant and recessive candidate genes. Nodes in the yellow box import additional annotations and nodes in the red box generate supplemental overview lists of patients and known pathogenic variants. A detailed view of all parts of the workflow and its 238 nodes is given in Section 6.1 in the Appendix (Chapter 6).

The developed KNIME workflow, named Candidate Identification Pipeline (CIP), consists of 238 nodes. A schematic overview that groups the nodes by their functionalities is shown in Figure 3.2 complemented by a more detailed description in Section 6.1 in the Appendix (Chapter 6). Nodes in the yellow box import the origin, the phenotype and the diagnostic state of the patients, the CADD scores and the InterVar classification of the filtered variants, and information on genes from the IMPC database and from OMIM. To search for dominant and recessive effects, the CIP takes frequency-filtered variants occurring hetero- or homozygous in at least one patient as input. In addition, a list of heterozygous variants is passed to the workflow, from which it creates potentially compound heterozygous variant pairs that serve as input for SmartPhase. After performing the variant prioritization steps described in Section 2.3.6, the resulting candidate variants are reported in four separate lists containing putative dominant and recessive effects in known and novel disease genes. The green boxes of the workflow generate the monoallelic candidates and the brown boxes are responsible for the biallelic candidates. Supplementary to the lists of candidates, nodes in the red box of the CIP compile a list of all information available for the patients in the WES cohort and a list with detailed information on variants that have previously been reported as diagnosis.

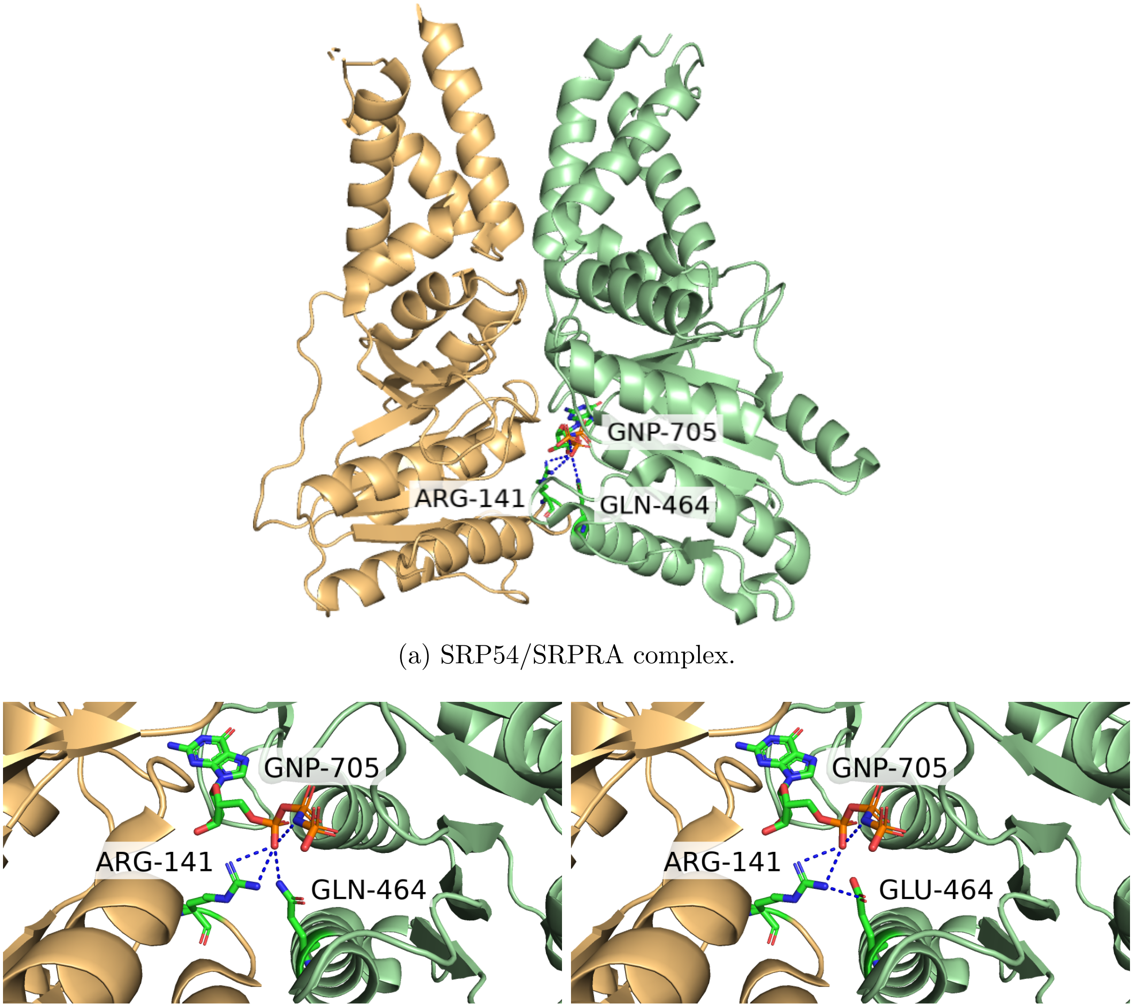
3.2 Analysis of the WGS data of family SCN-1

The analysis of family SCN-1 was performed to compare the quality of exome and genome sequencing and to search for a genetic variant that explains the pathophenotype of the index patient. It was performed by Kaarin Ahomaa as her master's thesis project under my supervision [117].

We compared the WGS data to WES data of the index patient SCN-1pa (II-3), her parents (I-1 and I-2) and two of her brothers (II-2 and II-5) (see Figure 2.1). In order compute the coverage of the exomes of the sequenced individuals, we defined a reference exome by merging all $208,979$ human protein-coding exons defined in Ensembl release 85 [119]. WGS resulted in an average sequencing depth of 42x over the entire genome, while WES achieved an average coverage of 117x for the exomes of the core trio, which were captured with the Agilent SureSelect V5+UTR kit, and 137x for the exomes of the brothers, for whom the more recent Agilent SureSelect V6+UTR kit was used (Agilent, Santa Clara, California). Although the WES coverage is much higher for both kits, the percentage of protein-coding regions reliably covered with at least 20 reads is considerably lower for both, the brothers with 86% and the core trio with 92%, in contrast to an average of 98% for all family members in the WGS data. This result highlights the superiority of WGS in comparison to WES when considering the completeness of the coverage of protein-coding regions.

We identified a total of $6,018,305$ variants in the WGS data set with an average (range) of $74,511$ ($66,424$ - $86,294$) SNVs and $10,568$ ($9,409$ - $12,122$) InDels per patient in the regions targeted by Agilent SureSelect exome capturing kits. Variant filtering and prioritization according to the procedures described in Section 2.3.6 revealed no promising segregating variant but two *de novo* variants with CADD scores higher than 25. Both of them are missense mutations, one in the gene *METTL26* (also known as *C16orf13*, chr16:g.686265C>T, ENST00000397666:c.26G>A, p.Arg9Gln) and the other in *SRPRA* (also known as *SRPR*, chr11:g.126134989G>C, ENST00000332118:c.1390C>G, p.Gln464Glu). While less is known about the function of human METTL26, *SRPRA* is a promising candidate gene as there are two studies that report that variants in *SRP54*, a direct interaction partner of SRPRA, induce a highly similar SCN phenotype including SDS-like features [120, 121].

SRPRA encodes for the α subunit of the heterodimeric signal recognition particle (SRP) receptor. The SRP complex and its receptor are a universally conserved cellular machinery that targets proteins cotranslationally to the endoplasmatic reticulum (ER), thus playing a



(a) SRP54/SRPRA complex.

(b) SRPRA wild-type closeup.

(c) SRPRA mutation closeup.

Figure 3.3: Visualization of the *SRPRA* mutation in patient SCN-1pa. The upper part (a) shows SRP54 in yellow on the left side and SRPRA in green on the right side in cartoon view. The glutamine (GLN-464) affected by the mutation in the patient, a close by arginine residue (ARG-141), and bound phosphoaminophosphonic acid guanylate ester (GNP-705), a non-hydrolyzable analog of guanosine triophosphate, are shown in stick representation together with hydrogen bonds between them (dashed blue lines). The lower left part (b) shows a closeup of the wild-type amino acids and the GNP. The lower right part (c) shows the same protein region with GLN-464 mutated to glutamic acid (GLU-464). Additionally, the hydrogen bond to GNP-705 is lost and one to ARG-141 is gained. The visualization was created with PyMOL(TM) (2.3.2) based on model 5L3Q in the Protein Data Bank [122].

central role for proper subcellular protein localization [123]. The identified *de novo* variant causes a change of glutamine to glutamic acid in the GTPase domain of SRPRA and has a CADD score of 27.4 indicating a deleterious mutation. Figure 3.3 shows that the mutated position is close to the guanosine-5'-triphosphate (GTP) binding pocket. Modeling the side chain substitution with PyMOL(TM) indicates that a hydrogen bond to GTP is lost, which might hinder GTP hydrolysis and SRP complex function. This would most likely lead to impairment of transmembrane protein targeting to the ER. The pathogenicity of the variant was validated in induced pluripotent stem cell (iPS) cells, which were differentiated into bona fide neutrophil granulocytes [124]. The introduction of the patient mutation in wild-type iPS cells reduced the capacity to differentiate into neutrophil granulocytes. Additionally, the *in vitro* generated neutrophil granulocytes were more susceptible to apoptosis and an increased activation of the unfolded protein response could be observed.

In retrospect, the *SRPRA* variant was part of the candidate variant set that resulted from the initial WES data analysis, but could not be linked to the phenotype of the patient as the pathogenicity of heterozygous variants in *SRP54* was not yet known. This case supports the argument that WES data of undiagnosed patients should first be re-evaluated before performing WGS to make use of the growing number of described genotype-phenotype associations [125]. Although WGS offers a better coverage of the protein-coding regions of the exome, the analyzed data shows that still approximately 90% of protein-coding regions are reliably covered by WES. Additionally WGS returns millions of variants that are hard to interpret because information about non-coding genetic elements and their influence on the phenotype is scarce. This reasoning is part of the motivation for the analysis of the entire WES data collection at the Dr. von Hauner Children's Hospital, the results of which are presented in the following sections.

3.3 Analysis of the WES data collection

The analysis of the entire WES data collection at the Dr. von Hauner Children's Hospital includes an initial quality control of the sequenced samples and the set of identified variants. The individuals are subsequently stratified according to their predicted ethnic origin as a prerequisite for computing population-specific allele frequencies before prioritizing potentially disease-causing variants. The following sections present the results of these major steps for WES data of $1,746$ patients and 705 healthy relatives (see Table 2.1).

3.3.1 Quality control of variants and samples

After the generation of the cohort VCF as described in Section 2.3.1 quality control procedures attempt to exclude false-positive variants by first removing low quality variants and then discarding all variants from samples with poor sequencing quality. The transition/transversion (Ti/Tv) ratio is a commonly used metric to measure the success of removing false-positive signals from variant call sets [86, 126]. Transversions are SNVs that exchange pyrimidine with purine bases or vice versa (A↔C, A↔T, C↔G, G↔T), while transitions are exchanges within pyrimidine or purine bases (C↔T, A↔G). If substitutions would occur at random, a Ti/Tv ratio of 0.5 would be expected as there are two possible transitions and four possible transversions. Because transversions are energetically unfavorable due to the structural difference of the nucleotides, they are observed less frequently resulting in Ti/Tv ratios around 3.0 for variants in coding regions and 2.0 in non-coding regions [126].

Stage of variant call set	Individuals	Variants	Transition/Transversion ratio
Initial variant call set	2,451	4,011,777	1.93
After variant quality control	2,451	3,675,680	2.06
After sample quality control	2,312	3,484,964	2.09

Table 3.1: Results of quality control steps applied to the exome sequencing data collection. The number of individuals in the cohort, the number of genetic variants, and the transition/transversion ratio is shown directly after variant calling, after variant quality control, and after sample quality control.

Table 3.1 shows that the variant quality control steps based on the overall quality of a variant and the quality of individual genotypes has a greater impact on the number of variants in the call set and on the Ti/Tv than the sample quality control. Summarized over both steps, 526,813 or 13.13% of all called variants are discarded including 346,878 SNVs with a Ti/Tv ratio of 1.00. The increase of the Ti/Tv ratio of the remaining variant set from 1.93 to 2.09 and the low Ti/Tv ratio of the removed variants indicates that the quality of the variant set was increased. Computing the Ti/Tv ratio of the filtered variant set separately for SNVs in coding and non-coding regions results in 2.66 and 1.94, respectively. The reason for the pronounced difference in the coding variants is that 75.64% of all identified SNVs are located in non-coding regions. This is because the used library preparation kits (Agilent SureSelect V5/V6+UTR, Agilent, Santa Clara, California) also capture UTRs and intronic regions around the exons. Despite considering the Ti/Tv ratios separately for coding and non-coding variants, they are lower as the expected with 2.66 versus 3.0 for coding and 1.94 versus 2.0 for non-coding regions. This is a consequence of the chosen filter setting, which was optimized to achieve a high sensitivity at the cost of some false-positive variants retained in the variant set.

Cohort	Trio Patients	Non-trio patients	Healthy relatives
Inflammatory bowel disease	18	645	49
Severe congenital neutropenia	93	185	231
Other immune defects	149	588	354
Total	260	1,418	634

Table 3.2: Composition of the exome sequencing data collection after quality control. Trio patients are patients for whom sequencing data of both parents passed the quality control measures. Patients are counted as non-trio patients if these are singletons, or only sequencing data of healthy relatives passed quality control, which do not form a complete trio. According to the main pathophenotype all patients and their relatives are stratified into three sub-cohorts.

Figure 3.4 shows an overview of the median sequencing depth and the rate of called genotypes for predefined frequent SNVs as computed by the tool Peddy. A total of 138

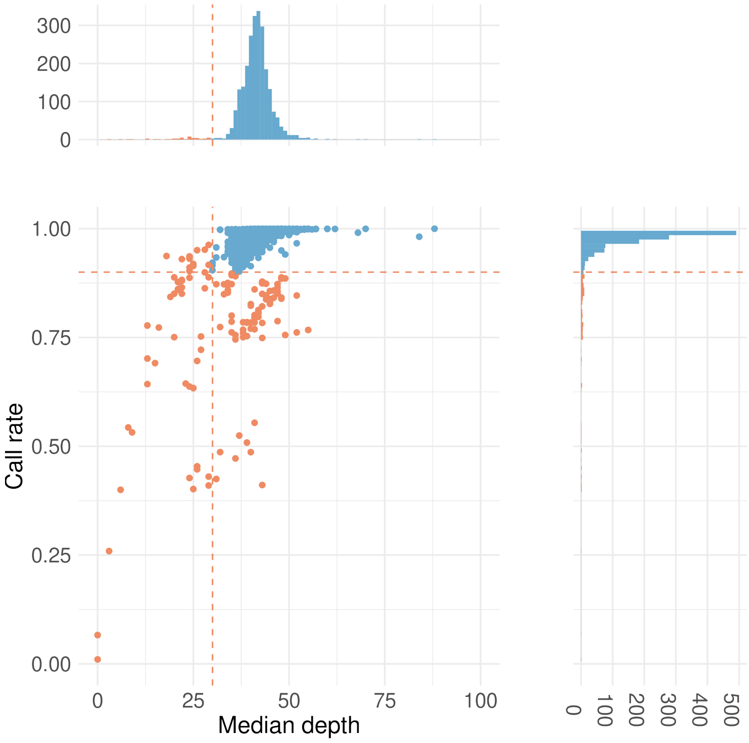


Figure 3.4: Sequencing depth and call rate of the exome sequencing data collection. The tool Peddy computes the sequencing depth and the call rate of each sequencing data set using frequent reference single nucleotide variants as described in Section 2.3.2. Each dot represents the sequencing data set of an individual. The orange dashed line represents the respective filter threshold value for both quality metrics. Samples having a median coverage below 30 and a call rate below 90% are highlighted in orange and are considered as low-quality samples. The bin size of the histogram of the median depth is 1 and 0.01 for the histogram of the call rate, respectively.

samples fail at one or both filter threshold values. Almost half of them (66 or 47.83%) are parents that were purposely sequenced at a lower coverage to improve filtering for segregation in the corresponding patient analyses at low financial cost. Nevertheless, they are omitted in the cohort analyses as their low-coverage and potentially false-positive variants could interfere when evaluating whether any healthy proband has deleterious variants in potential candidate genes. The histograms illustrate that the vast majority of the samples have significantly better quality than required to pass the filter threshold values. Of all samples, the median of the median depth is 41 and the median call rate amounts to 0.99. In addition to the 138 low quality samples, one sample was found that was sequenced twice as patient and father. The duplicate sample is not related to the mother and is male according to the WES data although the patient of the family is female. Since these indications suggest that the father was sequenced twice, the WES data set labeled as patient was excluded from subsequent analyses. Table 3.2 shows the number of trio patients, non-trio patients and healthy relatives remaining after quality control stratified by major pathophenotype. Compared to the cohort composition before the quality control measures (see Table 2.1), the number of trio patients is noticeably reduced by 55 or 17.46% because these are now counted as non-trio patients due to the removal of parents sequenced at low coverage.

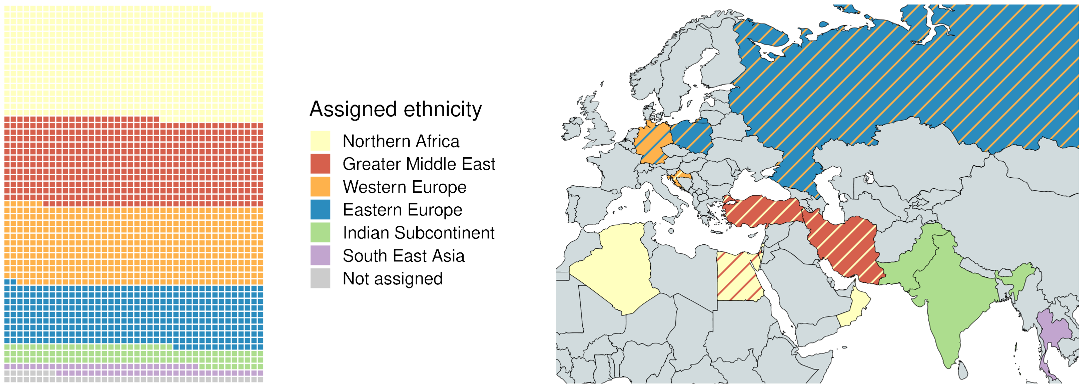
3.3.2 Population stratification and ethnicity prediction

Stratifying the individuals in the WES cohort at the Dr. von Children's Hospital by ethnic origin pursues two aims. First, the knowledge of the ethnic background of an individual can be used to reveal mislabeling of DNA samples. Second, stratifying the cohort by the ethnic origin of the probands enables the computation of population-specific allele frequencies, which is required for the identification of population-specific variants that seem rare in the overall cohort, although they are frequent in the population of the variant allele carriers. As these variants are most likely benign they can be removed during variant prioritization.

To determine the ethnic origin of the probands, the maximum subset of unrelated individuals was first extracted from the entire WES cohort as described in Section 2.3.3. The resulting set comprises $1,861$ individuals. Then, ADMIXTURE was applied to identify a stratification resulting in the minimum cross validation error by iterating from two to ten populations, which suggested an optimum number of six populations. By spreading the initial assignment of the $1,861$ unrelated individuals along the edges of the created relationship network, $2,264$ or 97.23% of the individuals of the cohort could be assigned unambiguously to one of the six identified populations. The remaining 48 individuals were not assigned, because they are related to individuals in different populations. Figure 3.5a visualizes the number of assigned individuals per population. The two largest populations cover more than half of all individuals, while only 7.9% of the cohort are assigned to the two smallest populations.

For the assignment of the identified populations to ethnicities, countries were sought that are the annotated origin of at least ten individuals with at least half of them being assigned to the same population. The identified population-specific countries are highlighted in the map in Figure 3.5b. The countries Algeria, Egypt, Israel and Oman were used to assign the largest population to Northern Africa (NAFR) ethnicity. The less specific Greater Middle East (GME) ethnicity is defined by Iran and Turkey. Europe is divided into Western Europe (WEUR) and Eastern Europe (EEUR) ethnicities characterized by the countries Croatia and Germany for the former, and Poland and Russia for the latter. India and Pakistan are allocated to Indian Subcontinent (ISC) ethnicity and Thailand represents South East Asia (SEAS) ethnicity.

As expected, the map shows that many countries cannot be assigned to ethnic groups unambiguously. There is a mixture of NAFR and GME ethnicities in Iran, Egypt and Turkey, just as WEUR and EEUR ethnicities are found in Germany, Poland and Russia. Thus, a contradiction between an individual's country of origin and its predicted ethnicity is not a clear indication whether the underlying DNA sample was mislabeled. It may be the consequence of overlapping ethnicities in neighboring countries, migration of the patient



(a) Population stratification.

(b) Ethnicity assignment.

Figure 3.5: Population stratification and assigned ethnicities. The left chart shows the distribution of the 2, 312 individuals in the WES cohort on the six populations identified by ADMIXTURE as described in Section 2.3.3. Each row of the chart contains 40 squares corresponding to 40 individuals. The squares are arranged according to the order of the assigned ethnicities in the legend starting with unassigned individuals in the lower left corner and then adding squares from left to right, from bottom to top for each ethnicity. Countries in the right map are colored if they are the origin of at least ten individuals in the cohort and more than half of them belongs to one of the six identified populations. These population-specific countries were used to define the ethnicities Northern Africa (NAFR, 688 individuals), Greater Middle East (GME, 538 individuals), Western Europe (WEUR, 484 individuals), Eastern Europe (EEUR, 376 individuals), Indian Subcontinent (ISC, 116 individuals) and South East Asia (SEAS, 62 individuals). The main color of the highlighted countries is determined by the primary ethnicity, which makes up more than 50% of the individuals from this country. The color of the stripes indicates secondary ethnicities that account for more than 10% of the individuals from the country. Population-specific countries are Algeria (22 individuals, 95.5% NAFR), Croatia (13 individuals, 76.9% WEUR, 15.4% NAFR), Egypt (42 individuals, 78.6% NAFR, 14.3% GME), Germany (618 individuals, 56.5% WEUR, 28.6% EEUR), India (32 individuals, 100% ISC), Iran (318 individuals, 53.5% GME, 39.9% NAFR), Israel (73 individuals, 75.3% NAFR, 16.4% GME), Oman (24 individuals, 100% NAFR), Pakistan (32 individuals, 100% ISC), Poland (44 individuals, 56.8% EEUR, 36.4% WEUR), Russia (16 individuals, 81.2% EEUR, 12.5% WEUR), Thailand (10 individuals, 100% SEAS) and Turkey (345 individuals, 59.5% GME, 34.7% NAFR).

or its family to a country with another predominant ethnicity, or the annotation of the country of origin refers to the origin of the DNA sample rather than the ethnic origin of the patient. Thus, the ethnicity prediction can be used as an additional information when there is a suspect of mislabeled DNA or to help to resolve such issues, however, its primary use is to identify and filter out population-specific variants.

3.3.3 Intermediate results of variant prioritization

In addition to the necessary variant prioritization steps based on annotations described in Section 2.3.4 and 2.3.6, frequency filtering using population-specific allele frequencies (see Section 2.3.3) and targeted search for potentially pathogenic compound heterozygous variant pairs using SmartPhase (see Section 2.3.5) are two novel features that were implemented for the analysis of the WES data collection. In order to estimate their influence on variant filtering, intermediate results of variant prioritization are presented in the following paragraphs.

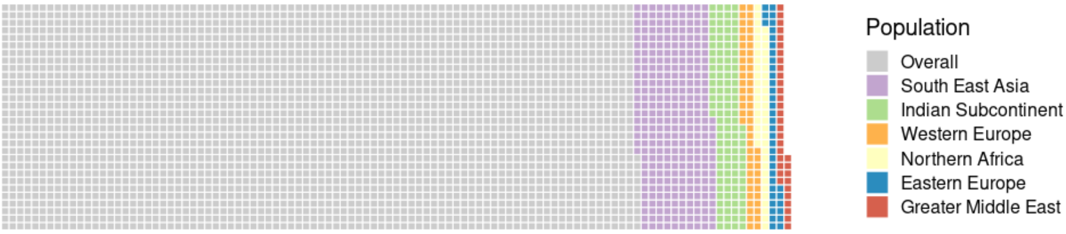


Figure 3.6: Frequency filtering combining overall and population-specific allele frequencies. Each square of the chart corresponds to 100 variants of the total set of 312, 959 filtered variants. The squares are arranged according to the order of the populations in the legend starting in the lower left corner with variants with an allele frequency (AF) equal to or greater than 0.1 in the overall cohort (253, 038 variants or 80.85%). Then squares are added from bottom to top, from left to right for each population for variants with an AF equal to or greater than 0.1 in the corresponding population and an AF smaller than 0.1 in the overall cohort (30, 503 variants or 9.75% for South East Asia, 11, 927 variants or 3.81% for Indian Subcontinent, 5, 667 variants or 1.81% for Western Europe, 4, 584 variants or 1.46% for Northern Africa, 3, 883 variants or 1.24% for Eastern Europe and 3, 357 variants or 1, 07% for Greater Middle East). The population with the maximum AF was counted if the AF was higher than 0.1 in multiple populations.

Population-specific allele frequencies The first step of variant prioritization is the exclusion of frequent variants as these cannot be the sole cause of a rare disease. The gnomAD database is the main resource for frequency filtering as it includes the largest collection of human exome and genome sequencing data. However, it is mainly based on individuals with European ancestry while the majority of the individuals in the WES cohort originates from Northern Africa and the Greater Middle East with some Asian populations (see Figure 3.5a). The differing underlying population structures limit the capability of filtering variants that are frequent in the populations included in the WES cohort but are rare in European ancestries. Therefore, frequency filtering in the generated variant call set is not only relying on public data sets but also on population-specific allele frequencies calculated separately for each population. Variants with an allele frequency equal or greater than 10% in the overall cohort or in any of the populations are assumed to be too frequent as a fully penetrant disease-causing allele and are discarded. Figure 3.6 shows the increase in the ability to filter variants using population-specific allele frequencies in addition to the allele frequencies as calculated over all individuals in the cohort. Additional 59, 921 variants can be filtered out corresponding to an increase of 23.68% and resulting in a total of 312, 959 excluded variants. As larger populations have a stronger impact on the overall allele frequency than smaller ones, the populations that dominate the cohort, namely NAFR, GME, WEUR and EEUR, contribute only to a minor extent to the additionally filtered variants. Consequently, the two smaller Asian populations ISC and SEAS have the highest impact through 42,430 variants that have an allele frequency equal to or greater than 10% only in one of both populations.

Division of the variant set into potentially dominant and recessive effects After having filtered by cohort and population-specific allele frequencies, the allele counts and frequencies provided by the 1000 Genomes Project and the gnomAD database were used to compile three lists of rare variants according to the thresholds defined in Section 2.3.6. The first list contains 1,319,177 variants that are heterozygous in at least one individual in the WES cohort, but that do not occur in either of the two reference data sets. The second list comprises 97,330 variants that were found to be homozygous in at least one individual in the cohort, but no homozygous genotypes are reported in the gnomAD data set, and the allele frequency is below 5% in the 1000 Genomes data set and below 0.5% in gnomAD. The third list includes 2,683,830 variants having a heterozygous genotype in at least one individual of the cohort and passing the same allele frequency thresholds as the variants in the second list. The CIP takes the first list as input to search for dominant effects, while the second and the third list are used to screen for recessive effects occurring as homozygous and compound heterozygous variants, respectively.

Search for compound effects To identify potential compound heterozygous variant pairs in the WES cohort, the prefiltered list of 2,683,830 rare heterozygous variants was further filtered for having an effect on the protein sequence or affecting splicing. Then, pairs of variants were created where both variants have been found in the same individual and both are located in the same gene resulting in a list of 99,300 variant pairs with 34 pairs per individual in the median. SmartPhase was able to resolve 30,848 of these pairs as being on the same allele, 2,041 pairs as being compound heterozygous and 460 as being non-pathogenic according to the segregation pattern in trio patients. All 30,848 pairs on the same allele and 460 non-pathogenic variants were discarded. To further reduce the amount of variant pairs that remain unresolved, all variant pairs were removed for which a pair with identical positions and alternate alleles exists in the set of the 31,308 variant pairs already discarded. As a result, 40,559 variant pairs remained, which were then concatenated with the list of 97,330 homozygous variants for joint processing as recessive candidate variants resulting in the lists described in Section 3.4.

3.3.4 Sensitivity of variant calling and frequency filtering

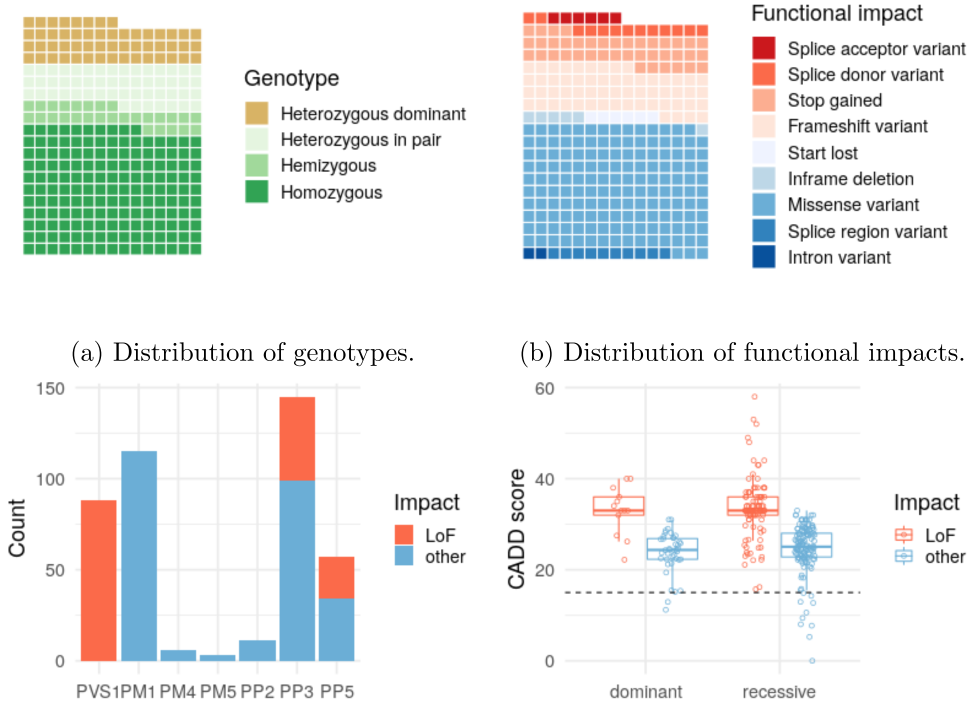
The sensitivity of variant calling and the subsequent quality control and frequency filtering steps is estimated by the amount of rediscovered known pathogenic variants. In total, 343 gene defects explain the pathophenotype of 337 of the 1,678 patients in the WES cohort.

There are six patients with two faulty genes and 27 of the gene defects are found in at least two patients. While 15 of the 343 gene defects are caused by larger deletions, 328 include small variants that should be found by the implemented cohort analysis pipeline. However, 14 variants were not found, because two of them are not covered by WES reads, seven are filtered out during the quality control of the variant call set and five are reported as homozygous genotypes in gnomAD. The latter are therefore not considered as fully penetrant pathogenic variants but rather as risk alleles. While it is impossible to identify the two variants not covered by sequencing reads, the other twelve variants could be kept in the variant call set by using more lenient threshold values when assessing the quality or frequency of a variant. As this would inevitably lead to longer lists of candidate variants and genes, the false-negative rate of 4.27% is accepted in the search for pathogenic variants in the entire cohort to limit the amount of false-positives in the underlying variant call set.

3.3.5 Reliability of variant impact annotations

As shown in Section 3.3.4, the cohort analysis pipeline rediscovered 314 known pathogenic gene defects in the WES cohort. Of these, 27 are found in more than one patient, which results in 293 distinct genetic variants when counting the heterozygous variants in 32 compound heterozygous pairs separately. Figure 3.7a shows that 81.91% of them are included in recessive inheritance patterns either as a homozygous or hemizygous variant, or as a heterozygous variant part of a compound heterozygous pair. The majority of the 293 variants has a direct effect on the protein-sequence (262 or 89.42%), a minor percentage directly affects splicing sites (29 or 9.90%) and only two variants are located in introns. The variants can further be divided in 117 LoF and 176 other variants as highlighted by the red and blue colors in Figure 3.7b. Since the pathogenicity of this set of variants is already proven, it offers the possibility to evaluate the reliability of variant annotations used for variant prioritization.

SpliceAI SpliceAI predicts a highly likely effect on splicing for 30 variants. Based on the variant annotation by the VEP (see Section 2.3.4), these include three frameshift, one missense and one intron variant in addition to twelve splice donor, seven splice region and six splice acceptor variants. Although the prediction of a splicing defect is expected for the latter three variant types, there are three splice region variants for which SpliceAI did not predict a splicing defect when considering only predictions with a probability above 80%. For two of them the probability of a splicing defect is close to the recommended threshold value (69% and 75%). For the third variant, SpliceAI correctly predicted the loss of a donor site as the existence of a splicing defect could be verified experimentally, but SpliceAI estimated the probability of the prediction to be far below the threshold at only 18%. Additionally, there is one deletion of ten bases affecting a splice donor site, but the used SpliceAI plugin provides no prediction, because only deletions up to a length of four bases are supported. Despite the chosen filter value aiming at high precision, SpliceAI



(a) Distribution of genotypes.

(b) Distribution of functional impacts.

(c) Number of ACMG/AMP criteria annotated by InterVar.

(d) Distribution of CADD scores by mode of inheritance.

Figure 3.7: Characterization of pathogenic variants discovered in the exome sequencing data collection. Chart (a) illustrates the distribution of the genotypes of the 293 distinct genetic variants. Green colors indicate recessive inheritance in contrast to dominant inheritance in ocher. Chart (b) shows the functional impact of the disease-causing variants ordered by severity as proposed by the Variant Effect Predictor [92]. Red colors distinguish Loss-of-Function (LoF) variants from other variants. Each square represents one variant in (a) and (b). The squares are arranged according to the order of the corresponding legend starting with the squares representing the last legend entry in the lower left corner and then adding squares from left to right, from bottom to top for each legend entry. The bar plot (c) indicates the number of criteria defined by the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) that apply to LoF and other variants as annotated by InterVar. PVS1 stands for "LoF variant in a gene where LoF is a known mechanism of disease", PM1 is "missense variant in a functional domain or mutational hotspot", PM4 is "inframe insertion or deletion that changes the length of the encoded protein or causes the loss of a stop codon", PM5 is "missense variant at the same position as a reported pathogenic variant", PP2 is "missense variant in a gene where missense variants are a common mechanism of disease", PP3 indicates that "multiple computational evidence shows deleterious effect", and PP5 states "variant reported as pathogenic" (see Figure 2.8). The box plot (d) compares the distribution of Combined Annotation Dependent Depletion (CADD) score values of LoF and other variants divided into those that contribute to dominant and those that contribute to recessive inheritance. The horizontal dashed line highlights the CADD score threshold value 15 used to filter for deleterious variants.

has a reasonable recall (89.66%) on pathogenic variants that cause splicing defects. The effect on splicing of the three frameshift, the missense and the intron variant has not yet been verified experimentally, thus the precision of SpliceAI cannot be assessed on this set of variants.

LOFTEE Of the 117 LoF variants, 98 pass all LOFTEE filters while 19 are labeled as low confidence LoF variants. Eight of the high confidence LoF variants are flagged as being located in an exon that does show the pattern of conservation of a protein-coding exon according to PhyloCSF [127]. One splice donor variant is classified as a low confidence LoF variant because it only affects splicing of the 5' UTR. Another splice region variant has low confidence because a rescuing splice acceptor site is located less than 15 base pairs away. All other low confidence LoF variants are filtered because they are close to the end of a transcript and are otherwise not deleterious according to LOFTEE [65]. The precision of LOFTEE regarding the ability to identify high confidence LoF variants cannot be reasonably determined on a set of known pathogenic variants. However, the recall is 83.76%, which indicates that LOFTEE should not be used to filter LoF variants when aiming to diagnose single patients as there is a chance of more than 15% that a pathogenic LoF variant is ignored as a low confidence variant. Nevertheless, a classification as a high-confidence LoF variant provides strong evidence supporting the deleteriousness of a variant when searching for new disease-causing gene defects. Therefore, the LOFTEE classification into high and low confidence LoF variants is used as a criterion for prioritizing variants in genes not related to IEI.

InterVar According to the variant classification scheme of the ACMG/AMP and the criteria annotated by InterVar [69, 94], 43 of the 293 known pathogenic variants are classified as pathogenic, 72 as likely pathogenic, one as likely benign and all others as variants with uncertain significance. As only 39.25% of the variants are correctly classified as pathogenic or likely pathogenic, but 60.41% remain unclassified and one variant is even misclassified, filtering by InterVar pathogenicity classification would result in many false-negatives. However, InterVar has annotated one of the eight pathogenicity criteria shown in Figure 2.8 for 254 or 86.69% of the known disease-causing variants. Figure 3.7c shows how often each criterion is annotated. Overall, 57 variants are listed as being pathogenic in ClinVar (PP5). No differing missense variant causing the same amino acid change as a reported pathogenic variant was found (PS1) and only three missense mutations at the same position as a reported variant are part of the 293 known pathogenic variants (PM5). Inframe deletions in non-repeat regions or stop-loss variants (PM4) are rare because these impacts are rare themselves (see Figure 3.7b). Missense variants in a functional domain or a mutational hotspot (PM1) are quite frequent, but only eleven missense variants are located in genes where missense variants are a common mechanism of disease (PP2). Regarding the 117 LoF variants, 88 of them are located in genes where LoF is a known mechanism of disease (PVS1). Finally, for almost half of the pathogenic variants (145 or 49.49%) multiple computational evidence suggests a deleterious effect. While the lack of these eight criteria does not implicate that a variant is benign, the presence of any of them represents an evidence of deleteriousness. The finding that 86.69% of all known pathogenic variants fulfill at least one of the eight criteria, motivates the use of this feature as a filter criterion in the search for deleterious variants in genes not related to IEI.

CADD In contrast to InterVar, CADD scoring does not depend on curated data sets of pathogenic and benign variants and therefore it is able to estimate the pathogenicity of every genetic variant. Figure 3.7d shows the distribution of the CADD scores of all 293 known disease-causing variants separately for those contributing to dominant and recessive inheritance (ochre and green colors in Figure 3.7a) subdivided into LoF and other variants (red and blue colors in Figure 3.7b). As expected, CADD scores are generally higher for LoF variants, and there is no difference between the scores of dominant and recessive variants reflecting that CADD scores estimate the pathogenicity of a variant independent of the mode of inheritance of its associated phenotype. Figure 3.7d also highlights the threshold value of 15 that is used to classify variants as potentially pathogenic or likely benign. As only twelve of the 293 known pathogenic variants are below this threshold value, CADD score based variant classification achieves a recall of 95.90% on this variant set.

Taken together, the results from the list of known pathogenic variants provide evidence that LOFTEE, InterVar and CADD generate reliable annotations as the filter criteria based on them achieve recall values of at least 84%. Section 3.4.2 presents how many genes known to cause IEI are rediscovered when applying these and all other filter criteria defined in Section 2.3.6.

3.4 Resulting lists of candidate variants and genes

Based on the three frequency-filtered input lists described in Section 3.3.3, the CIP generates four lists of candidate variants. Separated into mono- and biallelic variants, two lists contain variants in genes with known defects causing IEI and two contain variants in genes not associated with IEI. While the first two lists are only filtered to ensure that the variants are rare and affect only patients, the other two lists result from the application all the criteria described in Section 2.3.6 to generate a set of promising candidate genes.

3.4.1 Variants in genes associated with IEI

In 380 of all 408 genes reported to be causative for IEI [32], the CIP identified 1,852 homozygous variants and 383 pairs of heterozygous variants with 40 of them being compound heterozygous according to SmartPhase. In 101 of the 102 IEI dominant genes, 5,257 heterozygous variants were found including 185 *de novo* variants according to WES data of the parents. Although the second list comprises only a quarter of the genes of the first list, it is almost three times longer because filtering heterozygous variants by segregation is limited due to the low proportion of trios in the cohort (see Table 3.2). Of all 1,678 patients, 1,496 harbor a heterozygous variant and 936 are affected by a homozygous or a heterozygous variant pair in one of the IEI genes. In total, a rare genetic variant in at least one of the IEI genes can be found in 94.46% or 1,585 of the patients in the WES cohort.

Consequence	Heterozygous	Homozygous	Heterozygous in pair
Splice acceptor variant	11	4	10
Splice donor variant	12	11	15
Stop gained	29	22	18
Frameshift variant	34	45	36
Stop lost	4	-	-
Start lost	4	2	1
Inframe insertion	8	1	6
Inframe deletion	14	5	14
Missense variant	635	274	579
Protein-altering variant	-	-	1
Splice region variant	82	26	86
Synonymous variant	265	114	-
5' UTR variant	384	93	-
3' UTR variant	1,220	367	-
Non-coding transcript exon variant	-	1	-
Intron variant	2,348	811	-
Upstream gene variant	137	58	-
Downstream gene variant	70	18	-
Total	5,257	1,852	766

Table 3.3: Consequences of variants found in genes associated with inborn errors of immunity. Biallelic variants are separated into homozygous variants and heterozygous variants in potential compound heterozygous pairs. Synonymous, 5' UTR, 3' UTR, intron, upstream and downstream gene variants were not used to create pairs of heterozygous variants which explains the absence of these consequences in the fourth column of the table. The order of the consequences corresponds to the severity rating provided by the Variant Effect Predictor [92]. The abbreviation UTR stands for untranslated region.

Although almost all patients have a rare variant in one of the IEI genes, in most cases it probably has no functional impact. As Table 3.3 shows, the majority of the variants in the lists of heterozygous and homozygous variants are located in non-coding regions, such as UTRs and introns. Nevertheless, these might affect gene expression regulation by creating or disrupting uORFs, or splicing by activating cryptic splice sites as detailed in Section 1.1.2.

Of the 384 heterozygous and 93 homozygous 5' UTR variants, 32 and five are annotated as uORF affecting variants respectively. Only two of the heterozygous variants fulfill the criteria of being potentially deleterious by creating uORFs with a moderately strong Kozak sequence [114] and a reading frame overlapping and shifted towards the main reading frame. Both variants are found in patients suffering from IBD, one in the gene *PIK3R1* and the other in *SAMD9*. For both patients, the variants were excluded as sole cause of disease. The variant in *PIK3R1* occurs in four of the non-control individuals in gnomAD and the patient with the variant in *SAMD9* has already been diagnosed with a homozygous splice donor variant in *IL10RB*, known as a monogenic cause of VEOIBD [128].

Within the intronic and synonymous variants, SpliceAI predicts the gain of splice sites for three heterozygous and two homozygous intron variants and for one homozygous synonymous variant. One of the heterozygous and one of the homozygous intron variants can be excluded as monogenic cause of disease as the pathophenotypes of the affected patients are already explained by mutations in other genes. The homozygous synonymous variant is also considered as non-pathogenic as Sanger sequencing showed that the genotypes of the healthy parents are homozygous, too. The second heterozygous intron variant could be a false-positive variant, as the locus is covered only by 15 sequencing reads, twelve supporting the reference allele and three the mutant allele. The third heterozygous intron variant affects *ELANE* and the second homozygous intron variant affects *DOCK8*. Both variants are promising explanations for the diseases of the patients, one suffering from SCN and the other from hyperimmunoglobulin E syndrome, both known to be caused by mutations in the respective genes [32]. The heterozygous variant in *ELANE* is a *de novo* mutation according to the WES data and is predicted to create a donor splice site, which elongates exon 4 of 5 by 83 nucleotides (chr19:g.855879G>T, ENST00000263621:c.598-79G>T), thereby potentially introducing additional amino acids and a shift of the reading frame. SpliceAI also predicts the gain of a donor splice site for the homozygous variant in *DOCK8*. The variant is located 76 nucleotides after exon 23 in a transcript having 35 exons (chr9:g.429930A>G, ENST00000382329:c.3027+76A>G), which might result in an elongation of the coding sequence by 75 nucleotides. For both variants, the predicted elongations of the mRNAs have yet to be confirmed by gel electrophoresis of complementary DNA (cDNA) isolated from patient cells.

The most deleterious variants according to their functional impact are LoF variants. Among the homozygous variants there are 82 LoF variants, among the heterozygous ones there are 86 LoF alleles and the list of heterozygous pairs includes 17 pairs of LoF variants. Many of these variants have already been confirmed to be disease-causing. The proportion is highest for the homozygous variants, where 69 of the 82 variants are reported as genetic diagnosis. For heterozygous variant pairs, seven of the 17 pairs are considered as disease-causing, while pathogenicity is only proven for nine of the 86 heterozygous LoF variants. For most of them, segregation has not yet been tested and functional studies have not yet been conducted. Therefore, it cannot be assessed whether the considerably lower fraction of confirmed diagnoses among monoallelic LoF variants is due to the limited possibilities in filtering heterozygous variants as a consequence of the low amount of sequenced parents, or rather due to the primary focus on biallelic variants when searching for pathogenic variants.

In total, the two lists of possible dominant and recessive variants in IEI genes contain the genetic diagnoses of 260 patients. The number will certainly grow once segregation and functional studies are completed. However, this figure also shows that only a small proportion of all $1,678$ patients can be explained by variants in known genes. This motivates the search for deleterious variants in genes that are not yet associated with IEI, the results of which are presented in the following section.

3.4.2 Disease gene candidates

In the WES cohort, the CIP discovered 205 candidate genes each hit by deleterious variants in patients of at least two families. Recessive inheritance is assumed for 155 genes as these harbor biallelic variants, while dominant inheritance is supposed for 49 genes affected by heterozygous variants. In addition one gene on the Y chromosome has deleterious variants in two male patients. For each gene Table 6.1 indicates the number of identified variants and the range of their CADD scores, the number of affected patients and their genotypes, as well as phenotypic information taken from IMPC and OMIM.

Sensitivity of candidate gene selection In order to check the recall of candidate gene discovery, all genes serve as positive control that harbor known pathogenic mutations in at least two unrelated patients of the cohort. This set comprises 58 genes, 49 of which are causative for recessively inherited diseases in 181 patients and nine genes causative for dominantly inherited diseases in 47 patients. The list of 155 recessive candidate genes contains 35 of the 49 disease genes and the list of 49 dominant candidate genes includes five of the nine disease genes corresponding to an overall recall of 67.24%. There are a numerous reasons why 19 of the genes in the truth set are not rediscovered. Three genes are excluded as candidates because variants in healthy individuals fulfill the deleteriousness criteria, too. Eleven genes are missed because one or no patient remains after removing variants that fail one or more of the impact criteria. This happens four times because no pathogenicity criterion is annotated by InterVar, once because the CADD score is below 15 and once because a LoF variant has low confidence according to LOFTEE. For the remaining five genes a mixture of these criteria results in the exclusion of the respective gene as a candidate gene. Another gene is not found, because the two affected patients carry variants that occur in the gnomAD data set. In one more gene, one of two known pathogenic variants affects only one patient of a pair of diseased siblings with different pathophenotypes. As a consequence, the gene is not reported as a candidate, because there is a contradiction to the expectation that all patients of a family suffer from the same genetic defect, and there is just one other patient remaining with a deleterious variant in the same gene. One of the recessive disease genes is not found, because only one patient remains after excluding the compound heterozygous variant pair in the other patient, which could not be phased by SmartPhase. Two dominant disease genes are not identified as candidate genes, because dominant inheritance or a *de novo* event could not be confirmed in any of the four affected patients due to the lack of parental sequencing data. This filter criterion explains why the true positive rate is much higher for recessive candidates (71.43%) than for dominant candidates (44.44%). As sequencing data of the parents is only available for 260 of all 1,678 patients, strict filter criteria are required to reduce the number of false-positives among dominant candidates, which however will also cause more false-negatives than among recessive candidates. The mixture of filter criteria that cause genes to be missed shows that the chosen criteria are not redundant and there is no bias towards one single criterion.

Evaluation of the candidate gene list Reviewing the 205 candidate genes for biological plausibility with regard to the phenotypes of the patients is an ongoing process. We started with the eleven genes with potentially pathogenic variants in five or more patients. Seven of these genes, namely *HAX1*, *SRP54*, *G6PC3*, *IL10RB*, *BTK*, *ZAP70*, and *WAS* have already been associated with IEI [32]. One of the four remaining genes, given the pseudonym *C_GENE_1*, stands out because one of two potentially pathogenic variants in this gene was identified as a heterozygous genotype in seven unrelated patients (see Table 6.1). Whether the variants are inherited or *de novo* mutations could not be assessed for these patients, because exome sequencing data of the parents has not been generated to date. The identified missense variant causes a change of valine to glycine and has a CADD score of 23.3. As the parents of the affected families are healthy, the pathogenicity of the variant could only be explained by seven independent *de novo* events. While this already seems rather unlikely, the quality of the variant genotypes gives rise to further doubts. Only between 7.69% and 29.41% of the reads at the variant position support the alternate allele in the seven patients, which indicates false-positive variant calls. In addition, a heterozygous missense variant affecting the same amino acid position by exchanging the valine with an alanine is reported in 72 individuals in gnomAD, which contradicts a strong pathogenic potential of this amino acid position. The variant observed in gnomAD has a CADD score of 19.3 indicating deleteriousness. However, the prevailing evidence for a benign effect of this variant given by multiple healthy carriers suggests that CADD scores of amino acid changes at the affected position tend to be overestimated. Consequently, the variant observed in the patients was not analyzed further, as it is probably a false-positive signal. Finally, the described pathophenotypes for *C_GENE_1* in OMIM indicate autosomal recessive as well as autosomal dominant inheritance with strongly overlapping phenotypic descriptions. The annotated phenotype for autosomal dominant inheritance, which could be relevant for the seven patients, includes some of their symptoms, but is also characterized by prominent dysmorphic facial features, which were not reported for these patients. Based on these considerations *C_GENE_1* was discarded as candidate gene.

Another candidate gene, given the pseudonym *C_GENE_2*, is currently the focus of further functional studies. *C_GENE_2* is located on chromosome X and encodes a cytoplasmic protein that plays a role in cell signaling. The gene is a particularly promising candidate for a disease association as there are six male patients carrying six different hemizygous missense variants (see Table 6.1). All variants hit functional domains of the gene and the CADD scores range from 18.95 to 23.8. Interestingly, the pathophenotypes of the patients differ and include IBD phenotypes, inherited bone marrow failure, and vasculitis. This could be the result of a yet unknown wide-ranging functional diversity of the encoded protein, but several further experiments are required to explore the full spectrum of its biological functions and to evaluate the pathogenicity of the identified variants.

The following section provides a detailed description of the variants in the two remaining genes with at least five affected patients, namely *EPCAM* and *SLC5A1*. Additionally, pathogenic variants in eight other genes identified during my work on the WES cohort are presented.

3.5 Discovered pathogenic variants

The results of my work include not only the CIP itself and the generated lists of candidate variants and genes, but also the explanation of the genetic etiology of multiple patients. Some diagnoses resulted directly from the lists, others could be found after thorough review of all candidate variants of individual patients. The following sub-sections present 26 pathogenic variants whose discovery goes beyond routine diagnosis by having direct implications for the understanding of the diversity of IEI. The segregation of all variants was confirmed by Sanger sequencing, seven of them were published in four papers, and five are part of a manuscript in preparation.

Gene & Variant Position	Impact	CADD v1.6	InterVar v2.0.2	Type	Patients
EPCAM (AR)					
2-47600604-T-C	missense	25.4	PP3	hom	DIAR5-1pa
2-47600989-C-G	stop gained	33.0	PVS1, PP3	hom	DIAR5-2pa
2-47602386-G-T	stop gained	58.0	PVS1, PP3	hom	DIAR5-3pa
2-47604152-G-T	splice acceptor	32.0	PVS1, PP3	hom	DIAR5-4pa
2-47604202-A-G	missense	15.5	-	hom	DIAR5-5pa
2-47604217-G-C	splice donor	33.0	PVS1, PP3	hom	DIAR5-6pa
2-47606078-A-G	intronic	14.5	PP5	hom	DIAR5-7pa
SLC5A1 (AR)					
22-32462939-TG-T	frameshift	32.0	-	hom	GGM-1pa
22-32480526-C-G	missense	26.2	PP3	hom	GGM-2pa
22-32480629-T-C	missense	26.8	PP3	hom	GGM-3pa
22-32480964-G-T	missense	22.7	PM1, PP3	hom	GGM-4pa
22-32481011-T-A	missense	26.6	PP3	hom	GGM-5pa
22-32487745-G-T	stop gained	40.0	PP3	hom	GGM-6pa

Gene & Variant Position	Impact	CADD v1.6	InterVar v2.0.2	Type	Patients
CARMIL2 (AR)					
16-67681199-CAG-C	splice acceptor	33.0	-	hom	VEOIBD-1pa
16-67683038-CAT-C	frameshift	26.4	PVS1	hom	VEOIBD-2pa
16-67683871-GGTGGGCGTCC-G	splice donor	24.8	-	hom	VEOIBD-3pa
FOXP3 (XLR)					
X-49111955-CCTT-C	inframe deletion	20.7	PM4, PP5	hemi	VEOIBD-4pa
G6PC3 (AR)					
17-42152399-C-T	missense	29.8	PM1, PP3	hom	VEOIBD-5pa
SRP54 (AR)					
14-35465950-C-T	missense	25.3	PM1	hom	VEOIBD-6pa
RTEL1 (AR)					
20-62326972-G-A	missense	24.1	PM1, PP5	hom	VEOIBD-7pa
DKC1 (XLR)					
X-153993690-T-G	intronic	9.4	-	hemi	DKC-1pa
SRPRA (AD)					
11-126134989-G-C	missense	26.8	PM1, PP3	het	SCN-1pa
SRP54 (AD)					
14-35476564-G-T	missense	31.0	PM1, PP3	het	SCN-2pa
14-35476574-AAAC-A	inframe deletion	22.5	PM4, PP5	het	SCN-[3-7]pa
14-35477982-G-C	missense	27.0	PM1, PP3	het	SCN-8pa
SRP19 (AR)					
5-112200230-G-A	splice region	14.3	PP3	hom	SCN-9pa[1-4]

Table 3.4: Discovered pathogenic variants. The order of the genes corresponds to the order of their description in Section 3.5. The division by horizontal lines corresponds to the subsections 3.5.1, 3.5.2, 3.5.3, and 3.5.4. The abbreviation in brackets after the gene name indicates the mode of inheritance (AR is autosomal recessive, AD is autosomal dominant, XLR is X-linked recessive). The chromosome, position, reference and alternate allele of the variants are concatenated as a string with dashes as separator. The position is based on the human reference genome assembly GRCh37. The column “InterVar” lists criteria defined by the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) as annotated by InterVar. PVS1 stands for “Loss-of-Function (LoF) variant in a gene where LoF is a known mechanism of disease”, PM1 is “missense variant in a functional domain or mutational hotspot”, PM4 is “inframe insertion or deletion that changes the length of the encoded protein or causes the loss of a stop codon”, PP3 indicates that “multiple computational evidence shows deleterious effect”, and PP5 states “variant reported as pathogenic” (see Figure 2.8). The column “Type” specifies the genotype of the identified variant (hom is homozygous, hemi is hemizygous, het is heterozygous). The patient identifier consists of an abbreviation of the major pathophenotype, the number of the affected family and the number of the patient if the variant occurs in multiple patients of a family. “Diarrhea 5, with tufting enteropathy, congenital” is abbreviated as DIAR5, “glucose-galactose malabsorption” as GGM, “very early onset inflammatory bowel disease” as VEOIBD, “dyskeratosis congenita” as DKC, and “severe congenital neutropenia” as SCN. Square brackets are used to list affected patients or families.

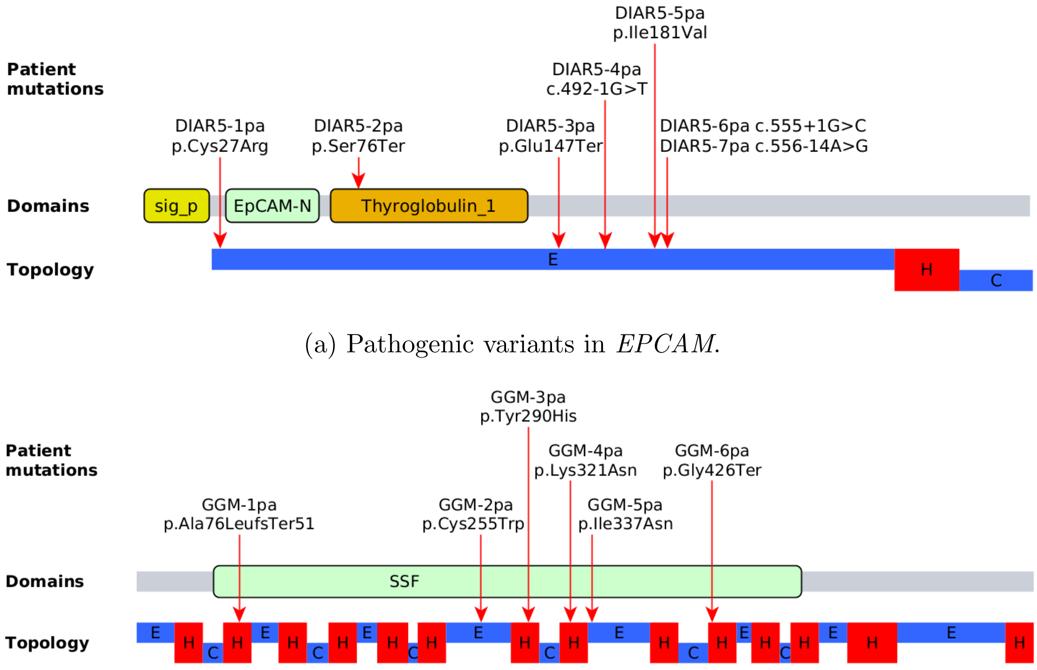
3.5.1 *EPCAM* and *SLC5A1* - Disease diagnosis driven by variant prioritization

Among the genes that are reported by the CIP to be hit by deleterious biallelic variants, *EPCAM* and *SLC5A1* appear as the strongest candidate genes. In addition to the fact that at least five variants were found in each gene, the phenotype of all affected patients is consistently characterized by chronic diarrhea and/or IBD-like phenotypes.

The protein product of *EPCAM* is the epithelial cell adhesion molecule, which is suspected to mediate physical interaction of intestinal epithelial cells and intraepithelial lymphocytes, thereby playing a role in the defense against mucosal infection [129]. Further, it is known that mutations in the gene can cause diarrhea-5 with congenital tufting enteropathy (DIAR5) [130], which matches the described pathophenotypes of the affected patients.

The solute carrier family 5 member 1 encoded by *SLC5A1* is responsible for the active transport of glucose and galactose across the brush border membrane of the cells within the gastrointestinal tract [131]. Mutations in *SLC5A1* can cause glucose-galactose malabsorption (GGM) in early childhood, which is characterized by severe chronic diarrhea and failure to thrive fitting to the phenotypes of the affected patients [132]. Although GGM may be lethal if untreated, it is possible to control the disease by replacing lactose, sucrose and glucose by fructose-based nutrients. Thus, the proper diagnosis of a *SLC5A1* defect is crucial for the patient as its quality of life can be improved tremendously and permanently through nutritional adjustment.

Table 3.4 lists the 13 identified variants in *EPCAM* and *SLC5A1* occurring as homozygous genotypes in 13 unrelated patients. The variants in the patients DIAR5-5pa and GGM-1pa were only found after a targeted search for variants in both genes, as InterVar did not annotate any of the required pathogenicity criteria for either variant. The pathogenicity of the variants was not verified experimentally, but can be derived from their annotated features as discussed in the following paragraphs.



(a) Pathogenic variants in *EPCAM*.

(b) Pathogenic variants in *SLC5A1*.

Figure 3.8: Positions of the pathogenic variants in *EPCAM* and *SLC5A1*. The domain and topology information is taken from Pfam [133] and UniProtKB [134] entries P16422 for *EPCAM* and P13866 for *SLC5A1*. The signal peptide of *EPCAM* is abbreviated as sig_p, the EPCAM N-terminal domain as EpCAM-N, and the Thyroglobulin type-1 repeat as Thyroglobulin_1. The sodium-solute symporter family domain of *SLC5A1* is abbreviated as SSF. The topology E stands for extracellular, C for cytoplasmic and H for helical transmembrane segments. The positions of the variants in *EPCAM* with respect to the cDNA and protein sequence are based on ENST00000263735 and ENSP00000263735. The positions in the protein sequence of SLC5A1 are based on ENSP00000266088.

EPCAM Figure 3.8a shows that all seven pathogenic mutations in *EPCAM* affect the extracellular part of the encoded protein. Because the missense variant in patient

DIAR5-1pa is located between the signal peptide and the N-terminal domain of *EPCAM*, it is likely that the functionality of one or both domains is impaired as indicated by the CADD score of 25.4 and other computational evidence of deleteriousness (PP3). CADD estimates the pathogenicity to be lower for the second missense variant in patient DIAR5-5pa, but the secondary structure of human EPCAM provides an indication of its impact. The affected isoleucine is located at the start position of a beta strand after a helix [135]. Although it is replaced by valine, an amino acid with similar physicochemical properties, it is nevertheless reasonable to assume that even minor changes at this structurally critical position in the protein can lead to misfolding and a dysfunctional protein product. The pathogenicity of the two stop-gain variants in the patients DIAR5-2pa and DIAR5-3pa results directly from the truncation of the protein together with the activation of NMD predicted by the NMDetective. SpliceAI predicts the creation of a splice acceptor site by the intron variant in patient DIAR5-7pa. Schnell *et al.* identified the same variant in a patient with congenital tufting enteropathy and verified its pathogenicity experimentally [136]. They predict that the variant introduces a frameshift and causes the truncation of the protein sequence after amino acid 191 (p.Tyr186PhefsTer6). Additionally, they observed *in vitro* that the mutant EPCAM is retained in the ER resulting in a loss of EPCAM at the cell surface. The same study analyzed a variant disrupting the same splice site as the variant in patient DIAR5-4pa and showed again that the mutant EPCAM is not present at the cell surface. Similar to the intron variant the authors predict that the loss of this splice acceptor site causes the truncation of the protein due to exon skipping and the introduction of a frameshift resulting in a termination codon after amino acid 188 (p.Ala165MetfsTer24). Under the assumption that the splice donor variant in patient DIAR5-6pa causes intron retention the protein would be truncated after amino acid 187 (p.Tyr186ValinsTer2). Although other splicing defects are conceivable, the variant is considered to cause disease because of the reported pathogenicity of other splice site disrupting variants in *EPCAM* [136].

SLC5A1 As depicted in Figure 3.8b all variants in *SLC5A1* are located in the sodium-substrate symporter family domain, which is required for glucose and galactose uptake in the small intestine. The four missense variants in the patients GGM-[2-5]pa are additionally characterized by CADD scores higher than 20 indicating deleterious amino acid exchanges that likely disrupt the transport function. Since the frameshift variant in patient GGM-1pa modifies the amino acid chain from position 76 on and causes a truncation after amino acid 126, it is apparent that the protein product is dysfunctional. The stop-gain variant in GGM-6pa truncates the protein closer to its end, but triggers NMD according to the NMDetective, which explains a complete loss of function.

Taken together, the identification of the variants in *EPCAM* and *SLC5A1* highlights that the developed CIP is not limited to detect genetic defects causing IEI, but can be applied more generally to identify mutations causing rare diseases. Furthermore, both genes demonstrate that even if the initially described pathophenotypes of patients indicate a certain disease, the analysis of genetic data should consider all discovered rare and deleterious variants as potential cause of disease.

3.5.2 *CARMIL2*, *FOXP3*, *G6PC3*, *SRP54* and *RTEL1* - New insights into VEOIBD as a feature of IEI

Besides searching for yet unknown disease genes, the CIP screens for harmful variants in known disease genes in order to learn more about the phenotypic spectrum of defects in these genes. With a focus on patients suffering from VEOIBD, the following paragraphs present pathogenic variants in five IEI genes, which are particularly interesting because the pathophenotypes of the seven patients involved extended the known phenotypic spectra of the underlying genetic defects.

CARMIL2 Three homozygous deletions were identified in *CARMIL2* encoding the capping protein regulator and myosin 1 linker 2 (see Table 3.4). CARMIL2, also known as RLTPR, is essential for CD28 costimulation in T cells and the development of regulatory T cells [137]. The central role of CARMIL2 in the immune system is reflected by the fact that mutations in the gene can cause a wide rage of immune-related phenotypes including recurrent bacterial, viral, and fungal infections, Epstein-Barr virus associated lymphoproliferative malignancy or atopy [138]. Patient VEOIBD-1pa and VEOIBD-3pa are children from consanguineous parents who both have a sibling each carrying the same genotype. All five patients suffer from pancolitis together with failure to thrive, abdominal pain and diarrhea classified as VEOIBD. Immunophenotyping of peripheral blood mononuclear blood cells of the patients showed CD28-dependent functional defects of T cell activation and proliferation confirming the presence of a CARMIL2 deficiency in the patients [139]. Consequently, the newly identified mutations demonstrate that CARMIL2 deficiency can manifest with IBD-like symptoms, which adds VEOIBD to the phenotypic spectrum of gene defects in *CARMIL2*.

FOXP3 A deletion of one amino acid in *FOXP3* was found in a male patient suffering from VEOIBD since its first weeks of life (ENST00000376197:c.748_750del, p.Lys250del, see Table 3.4). The forkhead box P3 encoded by *FOXP3* is a transcriptional regulator that plays an important role for the development of CD4+CD25+ regulatory T cells, which in turn are essential for the active suppression of autoimmunity [140]. Mutations in *FOXP3* are known to cause a syndromic disease characterized by Immunodysregulation, Polyendocrinopathy, Enteropathy, and X-linked inheritance abbreviated as IPEX [141]. As IPEX fits to the pathophenotype of the patient, the identified mutation was considered as causative and the patient underwent hematopoietic stem cell transplantation, which improved his condition considerably. Li *et al.* deciphered parts of the pathogenic mechanism of the same deletion and another adjacent deletion of an glutamic acid (p.Glu251del) [142].

They report that both mutations disrupt the leucine zipper motif, which is required for FOXP3 homooligomerization and heteromerization with FOXP1, thereby impairing proper DNA-binding. In addition to the known features of the mutation in patient VEOIBD-4pa, we describe for the first time ganulomas, small nodular collections of macrophages, in the lung and duodenum as a pathophenotype of IPEX [143]. Further, the example of the patient demonstrates that *FOXP3* should be considered as a candidate gene in patients suffering from VEOIBD.

G6PC3 & SRP54 A recent study of 25 Iranian patients with VEOIBD identified an underlying genetic defect in 14 patients suffering from IBD [144]. The two homozygous missense variants in the genes *G6PC3* and *SRP54* shown in Table 3.4 are particularly interesting because they provide new insights into the relationship of IBD and neutropenia. *G6PC3* encodes glucose-6-phosphatase, which catalyzes hydrolysis of glucose-6-phosphate to glucose and phosphate in the ER [145]. Gene defects in G6PC3 cause an autosomal recessive form of SCN associated with structural heart defects and urogenital abnormalities [146]. Symptoms compatible with the latter two pathophenotypes are also observed in patient VEOIBD-5pa, but the patient does not suffer from neutropenia. Despite this contradiction to the expected pathophenotype, the homozygous missense variant is considered pathogenic as it is predicted to cause a deleterious amino acid exchange in the enzymatic type 2 phosphatidic acid phosphatase domain of G6PC3 (ENST00000269097:c.479C>T, p.Ser160Leu). We hypothesized that the variant causes a functional defect in neutrophils that disrupts their role in the intestinal immune defense without increased apoptosis of neutrophils, which explains the IBD phenotype without neutropenia [144].

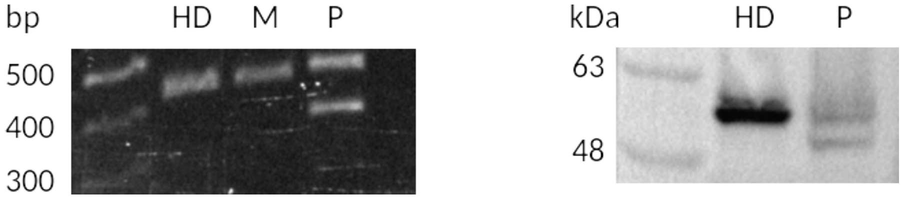
The signal recognition particle 54 encoded by *SRP54* is one of the components of the SRP that recognizes the signal peptide of secretory proteins and interacts with SRPRA of the SRP receptor to target them to the ER [147]. Two studies have shown that heterozygous mutations in *SRP54* can cause a dominant form of SCN associated with SDS phenotype [120, 121]. In contrast to the reported patients, patient VEOIBD-6pa carries a homozygous variant in *SRP54*, suggesting an autosomal recessive mode of inheritance for this gene as well. Interestingly, heterozygous carriers in this family seemed to be mildly affected. The predicted deleteriousness of the variant in the N-terminal helical bundle domain of *SRP54* (ENST00000216774:c.35C>T, p.Ser12Leu) gave rise to the hypothesis that the numeral defect in neutrophils hinders the formation of a normal gut barrier against microbiota, which predisposes the patient to IBD [144]. As a relationship between mutations in *SRP54* and IBD has never been reported before, further functional studies will be performed to evaluate this hypothesis. Together with the variant in *G6PC3*, both variants suggest that despite both genes are primarily associated with neutropenia, disruptions in both genes can also result in an IBD phenotype with no or less pronounced neutropenia.

RTEL1 The regulator of telomere elongation helicase 1 encoded by *RTEL1* is crucial for the maintenance of telomeres and therefore for general genomic stability [148]. Biallelic mutations in *RTEL1* have been described to cause dyskeratosis congenita (DC) and its more severe form Hyeraal-Hreidarsson syondrome (HHS) characterized among others by bone-marrow failure, developmental delay and recurrent infections [149, 150]. Based on functional studies showing that formation of telomeric circles is increased in patients with *RTEL1* mutations compared to controls, Walne *et al.* postulate that impaired processing of telomere loops during DNA replication causes telomere shortening, which manifests as DC and HHS [150]. Recently, we reported a homozygous missense variant in *RTEL1* in an Ashkenazi Jewish patient that presented with failure to thrive and infantile ulcerative colitis before severe immunodeficiency evolved as the patient grew older [151]. The variant in the patient (see patient VEOIBD-7pa in Table 3.4) causes an amino acid exchange from arginine to histidine at the C terminus of the protein (ENST00000360203:c.3791G>A, p.Arg1264His). Reports of patients with the same variant and similar symptoms prove the pathogenicity of the variant [149, 150]. Additionally, we observed shortened telomeres in leukocytes of the patient suggesting that the initial manifestation of the IBD phenotype is likely the result of aberrant telomere function in both immune and epithelial cells [151]. The variant is particularly prevalent in the Ashkenazi Jewish population with a carrier frequency between 0.45% and 1% and was presumably introduced by a common founder [152]. In addition to the extension of the phenotypic spectrum of *RTEL1* mutations, we therefore conclude that especially patients with IBD-like symptoms who originate from the Ashkenazi Jewish population should be screened for the identified variant in *RTEL1* [151].

Summary The identified mutations in the patients VEOIBD-1pa to VEOIBD-7pa extend the phenotypic spectrum of mutations in the genes *CARMIL2*, *FOXP3*, *G6PC3*, *SRP54* and *RTEL1*. Additionally, the described defects demonstrate that IBD-like symptoms can indicate underlying immunodeficiencies by co-occurring with other immmunophenotypes, or being the first or most pronounced manifestation of a defect in a gene associated with IEI. More generally, the results show that any rare and deleterious variant should be evaluated for its potential pathogenicity, regardless of whether the affected gene is associated with the observed pathophenotype.

3.5.3 *DKC1* - A novel intronic branch point mutation as cause of disease

A novel hemizygous variant in the second intron of the gene *DKC1* was found in a male patient clinically diagnosed as suffering from DC with an extended HHS phenotype (see patient DKC-1pa in Table 3.4). *DKC1* codes for dyskerin, which plays a critical role in ribosome biogenesis and telomere maintenance [153, 154]. Mutations in different exons of *DKC1* have been described to cause DC with HHS phenotype following X-linked recessive inheritance [155]. The variant identified in the patient is located in the intron after the second exon and 29 bases away from the start position of the third exon (ENST00000369550:c.85-29T>G). Due to the matching pathophenotype and the close location to an intron-exon boundary, it was hypothesized that the mutation causes DC and HHS through a splicing defect of *DKC1*. The Human Splicing Finder predicts the creation of an exonic splicing enhancer site in the intron but no probable impact on splicing [156]. BPP, an algorithm dedicated to detect branch points indicated that the mutation switches off a branch point [157]. Branch points are nucleotides within a conserved sequence upstream of the 3' splice site that are critical for splicing process [158].



(a) Gel electrophoresis of cDNA.

(b) Western blot.

Figure 3.9: Experimental validation of the splicing defect in *DKC1* in patient DKC-1pa. (a) Gel electrophoresis of the cDNA of *DKC1* generated from samples of a healthy donor (HD), the mother of DKC-1pa (M) and DKC-1pa (P). The leftmost column serves as a reference for the length of the separated fragments in base pairs (bp). (b) Western blot of dyskerin in peripheral blood mononuclear blood cells derived from samples of a healthy donor (HD) and DKC-1pa (P). The leftmost column indicates the length of the proteins in kilodaltons (kDa). Figures provided by Dr. Ido Somekh.

To test the hypothesis of a splicing defect, Dr. Ido Somekh analyzed the cDNA of *DKC1* of the patient, its mother, and a healthy donor by gel electrophoresis. Figure 3.9a shows one band for the mother and the healthy donor, but two bands for the patient. The upper band indicates cDNA with the same length as in the mother and the healthy donor. Through sequencing the patient cDNA within the lower band and the cDNA of the upper band in the mother, it was found that exon 3 is missing. The loss of exon 3 on protein-level was proven by the Western blot in Figure 3.9b, which shows two bands for the patient and only one band for a healthy donor. Both experiments support the hypothesis of a splicing defect resulting in two transcript isoforms transcribed from the same allele. The mutant form lacks exon 3, but is transcribed and translated to a similar extent as the wild-type form. Exon 3 seems crucial for the proper function of dyskerin as several mutations in exon 3 of *DKC1* are known to cause DC with the extended HHS phenotype [155]. Taken together, it was concluded that the hemizygous mutation is pathogenic by rendering approximately half of the translated dyskerin dysfunctional as a consequence of the deletion of the amino acids encoded by exon 3.

3.5.4 *SRP54* and *SRP19* - Novel variants and a new gene causing SCN

Following publications of autosomal dominant genetic defects in *SRP54* [120, 121] and the recognition of the pathogenicity of the *de novo* variant in *SRPRA* in patient SCN-1pa (see Section 3.2), the WES data collection was screened for variants in these and other genes that form the SRP and its receptor. Thereby, I found one of the published variants in

SRP54 in five patients, two novel variants in *SRP54* and a variant in the gene *SRP19* that was so-far not functionally related to any disease phenotype (see Table 3.5).

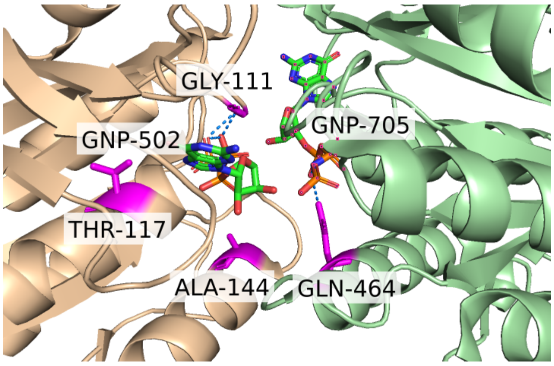


Figure 3.10: Residues in the SRP54/SRPRA complex affected by mutations in the families SCN-1 to SCN-8. The figure shows a closeup of the GTPase domains in the SRP54/SRPRA complex that is also shown in Figure 3.3a. SRP54 on the left side is colored yellow, SRPRA on the right side is colored green. Stick representation is used for the amino acids affected by the identified mutations and bound phosphoaminophosphonic acid guanylate ester, a non-hydrolyzable analog of guanosine triophosphate (GNP-502, GNP-705). Amino acids mutated in the patients SCN-1pa (GLN-464), SCN-2pa (GLY-111), SCN-[3-7]pa (THR-117) and SCN-8pa (ALA-144) are highlighted in magenta. Hydrogen bonds between GLY-111 and GNP-502 and between GNP-705 and GLN-464 are indicated by dashed blue lines. The visualization was created with PyMOL(TM) (2.3.2) based on model 5L3Q in the Protein Data Bank [122].

SRP54 Figure 3.10 shows that the variants identified in *SRPRA* and all variants identified in *SRP54* affect residues that are located in the GTPase domains of the interacting SRPRA and SRP54. Glycine at position 111 (GLY-111) is mutated to tryptophan in patient SCN-2pa by a heterozygous missense variant (ENST00000216774:c.331G>T, p.Gly111Trp). Similar to the disappearance of the hydrogen bond between GLN464 and GTP (represented as GNP-705) in SCN-1pa (see Figure 3.3b and 3.3c), the mutation in SCN-2pa most likely alters both the hydrogen bond to GTP (represented as GNP-502) and the overall structure of the GTPase region due to the structural difference of glycine and tryptophan. As the SCN phenotype of SCN-2pa fits to the pathophenotypes caused by reported mutations in *SRP54* this novel mutation is considered as pathogenic although its segregation could not be evaluated due to missing sample material of the parents.

The deletion of the threonine at position 117 (THR-117) was already reported in one patient by Carapito *et al.* [120] and in 14 patients by Bellanne-Chantelont *et al.* [121]. The mutation was identified as a heterozygous inframe deletion in the patients SCN-3pa to SCN-7pa (ENST00000216774:c.349_351del, p.Thr117del). As Figure 3.10 shows that the deleted amino acid THR-117 is located in an helix, Carapito *et al.* and Bellanne-Chantelont *et al.* assume that the mutation affects GTP binding through rearrangement of the three-dimensional structure of the binding pocket. In accordance with the published pathogenicity of the variant, all five patients suffer from SCN and myeloid maturation arrest could be shown in four patients (Y. Mizoguchi, S. Hesse, M. I. Linder, *et al.*, manuscript in preparation). Exocrine pancreatic insufficiency as a feature of SDS was additionally observed in three patients. Sanger sequencing verified that the variant was inherited from the diseased mother in SCN-3pa and is caused by *de novo* mutation events in the patients SCN-[4-7]pa. The variant was also confirmed as a *de novo* mutation in the patient described by Carapito *et al.* [120] and in seven of the 14 patients reported by Bellanne-Chantelont *et al.* [121]. This data demonstrates that the affected site exhibits increased mutability most likely caused by the trinucleotide repeat ACAACAACA between the positions 35,476,576 and 35,476,584 coding for three tryptophan residues at the residues 115 to 117. Instability of trinucleotide repeats as a consequence of deletions and expansions is a known feature in the genomes of various organisms [159] and might such explain the identification of recurrent *de novo* mutations in multiple patients.

Another rare and deleterious missense variant that changes alanine at position 144 (ALA-144) to proline (ENST00000216774:c.430G>C, p.Ala144Pro) was identified in patient SCN-8pa as a *de novo* mutation. As the affected alanine is located in a helix that is part of the GTP binding pocket, the mutation might also cause structural rearrangements causing impaired GTP binding as assumed for the mutation of THR-117. Even though the SCN phenotype of the patient would fit well to the mutation, its role is unclear, since a homozygous missense mutation was also found in the gene *VPS45*. This gene belongs to the set of prominent genes known to cause SCN [32], so it is possible that both mutations contribute equally to the pathophenotype, one overlaps the effect of the other or the effects are mutually reinforcing.

SRP19 The analysis of WES data of two sibling pairs affected by SCN and growth deficiency that belong to a larger family revealed a splice region variant in *SRP19* as the only overlapping candidate after variant prioritization (ENST00000505459, c.189+5G>A). Sanger sequencing confirmed that the variant is homozygous in the four patients SCN-9pa1 to SCN-9pa4 and a brother of the first sibling pair and that it is heterozygous in their consanguineous parents. The Human Splicing Finder predicts that the variant causes an alteration of the splice donor site, which has an effect on splicing. Experimental studies of transcripts of *SRP19* in patient cells could confirm aberrant expression of SRP and showed that the mutation introduces a new isoform of *SRP19* by skipping exon 3 (Y. Mizoguchi, S. Hesse, M. I. Linder, *et al.*, manuscript in preparation). Analogous to the *SRPRA* mutation in SCN-1pa, the introduction of the mutation in iPS cells results in reduced capacity to differentiate into neutrophil granulocytes, developed cells are more susceptible to apoptosis and unfolded protein response seems to be increased [124].

3.5.5 Functional context of the affected genes

To gain insight into the functional context of the genes listed in Table 3.4 and their relationships to each other, Figure 3.11 shows a protein-protein interaction (PPI) network of the encoded proteins and their closest interaction partners.

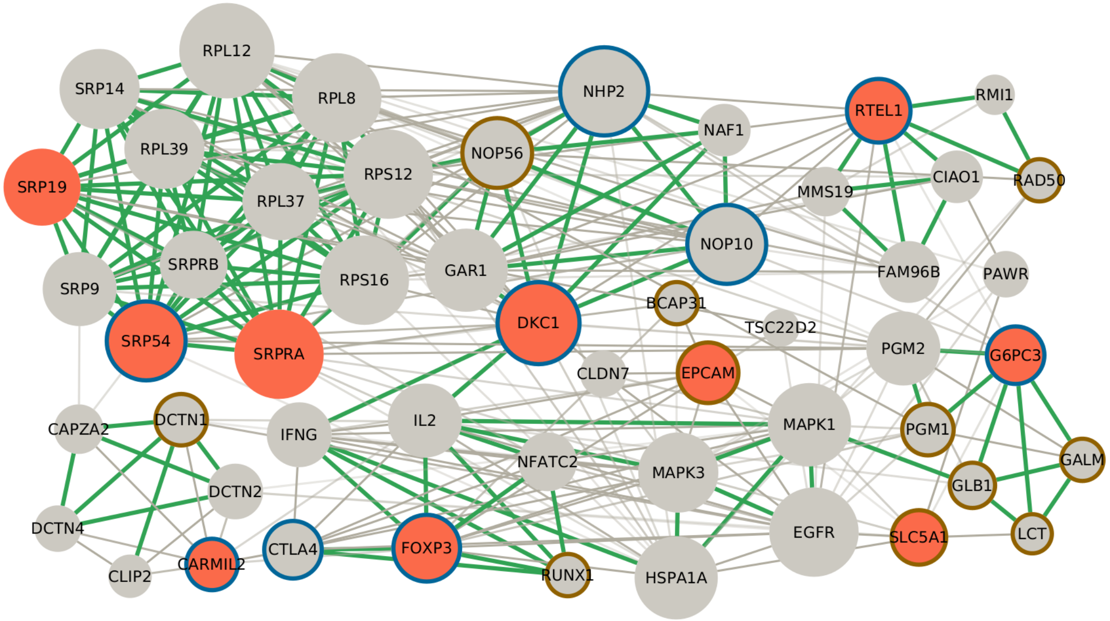


Figure 3.11: Interaction network of proteins affected by the identified pathogenic variants. The red nodes represent the ten genes harboring the 26 pathogenic mutations described in the preceding sections. Genes associated with inborn errors of immunity as reported by the International Union of Immunological Societies are highlighted by blue borders [32]. Brown borders indicate genes associated with other Mendelian diseases according to the Online Mendelian Inheritance in Man database [22]. The size of the nodes reflects the number of interaction partners. Interactions were extracted from the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database [160] in two steps. First, for each of the ten genes I selected the five most reliable protein interactions that have a combined interaction score of at least 0.15 when restricting to the interaction sources "experiments" and "databases". Then, I queried all interactions between the 52 resulting proteins that have a minimum interaction score of 0.15 and visualized the resulting network using Cytoscape [161]. Green edges highlight high confidence interactions (interaction score $\geq$ 0.7) based on the interaction sources "experiments" or "databases". The intensity of the color of the edges indicates the reliability of the interaction.

The network was extracted from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [160] as described in the caption of Figure 3.11. It comprises 52 proteins including nine proteins associated with IEI according to the latest IUIS report [32] and ten proteins associated with other diseases as reported in the OMIM database (see Table 3.5). The proteins are connected by 296 PPIs that have at least low confidence based on all provided interaction sources. When restricting to "databases" and "experiments" as interaction sources and taking only high confidence interactions into account, 129 interactions remain. Three of the proteins affected by the discussed pathogenic variants share no high confidence PPIs to any other protein, namely EPCAM, SLC5A1, and CARMIL2. All others are part of clusters of proteins internally strongly connected by high confidence PPI. Some of these cluster also share high confidence interactions to other clusters.

Gene	Disease associations	Source
BCAP31	Deafness, dystonia, and cerebral hypomyelination (XLR)	OMIM
CARMIL2	Severe combined immunodeficiency due to *RLTPR* deficiency (AR)	IUIS, OMIM
CTLA4	Autoimmune lymphoproliferative syndrome, type V (AD)	IUIS, OMIM
DCTN1	Perry syndrome (AD); Distal hereditary motor neuronopathy type VIIB (AD)	OMIM
DKC1	Dyskeratosis congenita (XLR)	IUIS, OMIM
EPCAM	Diarrhea 5, with tufting enteropathy, congenital (AR)	OMIM
FOXP3	Immunodysregulation, polyendocrinopathy, and enteropathy, X-linked (XLR)	IUIS, OMIM
G6PC3	Severe congenital neutropenia (AR)	IUIS, OMIM
GALM	Galactosemia IV (AR)	OMIM
GLB1	GM1-gangliosidosis, type I - III (AR); Mucopolysaccharidosis type IVB (Morquio) (AR)	OMIM
LCT	Congenital lactase deficiency (AR)	OMIM
NHP2	Dyskeratosis congenita (AR)	IUIS, OMIM
NOP10	Dyskeratosis congenita (AR)	IUIS, OMIM
NOP56	Spinocerebellar ataxia 36 (AD)	OMIM
PGM1	Congenital disorder of glycosylation, type It (AR)	OMIM
RAD50	Microcephaly and chromosomal instability without immunodeficiency (AR)	OMIM
RTEL1	Dyskeratosis congenita (AD, AR); Pulmonary fibrosis and/or bone marrow failure (AD)	IUIS, OMIM
RUNX1	Platelet disorder, familial, with associated myeloid malignancy (AD); Acute myeloid leukemia (AD)	OMIM
SLC5A1	Glucose/galactose malabsorption (AR)	OMIM
SRP19	Severe congenital neutropenia (AR)	novel
SRP54	Severe congenital neutropenia (AD)	IUIS, OMIM
SRPRA	Severe congenital neutropenia (AD)	novel

Table 3.5: Gene-disease associations. The table lists disease associations of the 22 genes highlighted in Figure 3.11 and specifies their source. Associations are taken from the International Union of Immunological Societies (IUIS) publication [32], from the Online Mendelian Inheritance in Man (OMIM) database [22], or were identified for the first time during the analysis of the sequencing data underlying this thesis (novel).

Interactions of EPCAM, SLC5A1 & G6PC3 Experimental evidence suggests that EPCAM expression is controlled by BCAP31 during the regulation of human embryonic stem cell adhesion, stemness, and survival [162]. While mutations in *EPCAM* cause DIAR5 [130], mutations in *BCAP31* cause the rare disease "Deafness, dystonia, and cerebral hypomyelination", which mainly affects the central nervous system but is not associated with gastrointestinal features [163]. Although there is a functional link between EPCAM and BCAP31, the caused disease phenotypes are not related, in contrast to SLC5A1 and LCT. STRING reports an association of SLC5A1 and LCT or lactase primarily because mutant forms of both result in a similar pathophenotype characterized by neonatal-onset watery diarrhea and failure to thrive. Both diseases, glucose-galactose malabsorption due to mutations in *SLC5A1* [132] and congenital lactase deficiency due to mutations in *LCT* are rare while the latter is primarily reported in Finland [164]. LCT is part of a cluster of six proteins responsible for galactose metabolism. Five of them are associated with Mendelian diseases (see Table 3.5) including G6PC3. Although defects in this protein are predominantly associated with neutropenia, we recently reported a patient having a mutation in *G6PC3* and suffering primarily from IBD symptoms [144]. This observation demonstrates that not only defects in functionally linked genes can produce different pathophenotypes, but even defects in the same gene can result in varying disease phenotypes.

Interaction of RTEL1 RTEL1 links two small protein clusters each fully connected by high confidence PPIs. Together with CIAO1, FAM96B and MMS19 it takes part in cytosolic iron-sulfur cluster assembly. Iron-sulfur clusters are essential for various biological processes, such as mitochondrial respiratory chain activity. Defects in their biogenesis have been associated with multiple human diseases [165]. Additionally, RTEL1 is involved in DNA replication processes together with RAD50 and RMI. Both, defects in RTEL1 and RAD50 cause genome instability but the resulting phenotypes differ. While mutations in *RTEL1* cause Dyskeratosis congenita characterized by a severe immune phenotype [149, 150], RAD50 deficiency is expressed trough anatomical anomalies without immunodeficiency [166].

Interactions of CARMIL2 CARMIL2 has no high confidence PPI to any of the other 51 proteins in the network depicted in Figure 3.11, but it has multiple weak interactions to a cluster of five cytoskeleton proteins including DCTN1 that is associated with distal hereditary motor neuronopathy type VIIB with variable onset, [167] and the neurodegenerative

Perry syndrome with adult onset [168]. Both diseases are completely different from the immune and IBD pathophenotype that we have reported for the mutation in *CARMIL2* [139]. A similarity in terms of associated diseases is rather reflected by the relation to CTLA4. Both proteins are involved in T cell regulation and their mutant forms cause diseases of immune dysregulation with overlapping symptoms, such as diarrhea [169].

Interactions of FOXP3 In addition to the association with CARMIL2, CTLA4 is part of a cluster of six proteins having high confidence PPIs with FOXP3. All of the interacting proteins are involved in the regulation of regulatory T cell differentiation. Mutations in both *CTLA4* and *FOXP3* have been described to cause regulatory T cell defects, which are phenotypically similar to defects in CARMIL2 but result in more severe pathophenotypes [170]. Among the proteins in the cluster, RUNX1 is another gene associated with Mendelian disease. While defects in CARMIL2, CTLA4 and FOXP3 are all forms of IEI, mutations in *RUNX1* are related to myeloid malignancy and acute myeloid leukemia (see Table 3.5) despite a regulatory relationship between FOXP3 and RUNX1 [171]. Through high-confidence PPIs between IL2 and MAPK1 and MAPK3 as well as IFNG and HSPA1A, the FOXP3 cluster is connected to a cluster of proteins involved in cellular response to oxidative stress consisting of EGFR, HSPA1A, MAPK1 and MAPK3. A relationship between regulatory T cells and oxidative stress was observed in tumors, for example, where oxidative stress regulates apoptosis and suppressor activity of regulatory T cells [172].

Interactions of DKC1 The FOXP3 cluster connects to a cluster of proteins interacting with DKC1 via two high confidence interactions between IFNG and DKC1, and IL2 and DKC1. According to the Pathway Interaction Database [173], both proteins regulate the activity of human telomerase, a complex consisting of telomerase reverse transcriptase, telomerase RNA, and dyskerin encoded by *DKC1* [174]. The proteins in the DKC1 cluster are all part of the small nucleolar ribonucleoprotein complex that is involved in ribosomal RNA modification. Mutations in *DKC1* trigger Dyskeratosis congenita [155], analogous to mutations in the genes *NHP2* and *NOP10* (see Table 3.5), both encoding ribonucleoproteins that are part of the interaction cluster of DKC1. Also the interacting ribonucleoprotein NOP56 is related to a Mendelian disease, namely Spinocerebellar ataxia 36, which is characterized by adult onset and neurological features without immune system anomalies [175].

Interactions of SRP19, SRP54 & SRPRA Through high confidence interactions with RPS12 and RPS16, NOP56 links the ribonucleoprotein cluster to the most pronounced cluster in the left upper part of Figure 3.11. The cluster includes SRP19, SRP54 and SRPRA, the three proteins of the SRP complex and its receptor that are mutated in overall twelve SCN patients of the analyzed WES cohort. The interaction partners of these three proteins are overlapping, thereby forming a cluster of proteins responsible for SRP-dependent co-translational protein targeting to the ER membrane. The similar pathophenotype of the twelve patients together with the observed proteome aberrations including the decreased abundance of granule proteins suggests that especially neutrophil granulocytes are highly dependent on balanced protein-synthesis, -trafficking, and -homeostasis (Y. Mizoguchi, S. Hesse, M. I. Linder, *et al.*, manuscript in preparation).

Summary Taken together, the proteins affected by the pathogenic mutations described in Section 3.5 and their interaction partners shown in Figure 3.11 are involved in multiple biological functions ranging from processes directly related to the immune system, such as T cell regulation, to ubiquitous processes such as DNA replication, ribosomal RNA modification, and protein targeting to the ER membrane. This diversity of biological processes is also reflected in the diversity of diseases associated with the underlying genes. Even directly interacting proteins are related to diseases with differing phenotypes, such as FOXP3 and RUNX1, or DKC1 and NOP56. On the other hand, however, there are genes associated with the same disease whose proteins are not connected by high confidence interactions in the STRING database when restricting to the interaction sources “databases” and “experiments”, for example DKC1 and RTEL1, both causing Dyskeratosis congenita, or G6PC3 and the three SRP proteins that are all causative for SCN. Two conclusions can be drawn from these observations. First, proteins interacting with each other do not necessarily cause similar pathophenotypes. Second, immune system processes are intertwined with and depend on multiple biological processes. As a consequence, mutations can lead to severe immunological diseases, even though the function of the affected gene is not specific to cell types or processes of the immune system.

Discussion 4

The analysis of the WGS data set and the WES data collection resulted in multiple genetic diagnoses and novel insights in the genetic etiology of IEI as presented in the last chapter. Besides the data itself, the chosen steps of variant prioritization including the development of SmartPhase, the computation of population-based allele frequencies, and the implementation of the CIP were crucial for these achievements. This chapter discusses the opportunities and limitations of the methodology and evaluates the biological and clinical importance of the discovered pathogenic variants and genes.

4.1 Challenges of variant prioritization and resulting practical approaches

While the identification of small genetic variants is a standardized step in the analysis of NGS data, the selection of those that are clinically relevant is a challenging task due to the large number of genetic variants in each individual. In the exomes of the presented cohort of 2,312 individuals, approximately 3.5 million variants were identified using the human reference genome build GRCh37. This number is still rather small when considering growing clinical cohorts and increasing usage of WGS [176]. Especially, the transition to WGS will raise the number of variants as illustrated by the analysis of the WGS data of family SCN-1 based on the same genome assembly. Together, the seven sequenced family members carry 6.0 million variants, almost double the amount of variants in the whole WES cohort. Two recent studies also relying on GRCh37 show how much the number increases when WGS is applied to larger cohorts. A rare disease study identified 172 million short variants in 13,037 individuals [176] and the gnomAD WGS data set based on 15,708 individuals reports 230 million variants [65]. Therefore, variant annotation and filter criteria for variant prioritization must become more and more sophisticated to reduce the set of all identified genetic variants to short lists of variants of interest. The following sections present the challenges of variant annotation and filtering, and discuss how problems were methodically approached and solutions implemented.

4.1.1 Variant annotation

In the context of rare disease etiology, rare and deleterious genetic variants are in the focus of diagnostic and research projects. To select such variants as candidates for the cause of disease, impact, frequency, and segregation criteria are first annotated for all identified variants and subsequently used to define filter criteria for variant prioritization.

Impact annotation The prediction of the impact of a variant on gene-level depends on the proper annotation of genes in the human genome, which is still an ongoing project [177]. Although the application of widely-used tools, such as the VEP, ensures a certain degree of standardization, it is important to keep in mind that the annotation of a genetic variant can change when new releases of gene annotation become available. In addition of being able to differentiate, e.g. synonymous from missense variants, the estimation of the harmfulness of a genetic variant is critical for its interpretation. The ACMG/AMP variant classification rules provide a valuable framework for standardized variant interpretation, but their application requires manual annotation of several predefined criteria. As it is unfeasible to review all variants of a patient manually, algorithms have been developed to automatize the assessment of variants. Because almost all of them have certain limitations, multiple tools should be combined to optimize the estimation of variant deleteriousness [108]. Accordingly, I've used the VEP together with its plugins SpliceAI, UTR annotator and LOFTEE to classify variants by their functional impact, as well as a combination of CADD scores and ACMG/AMP criteria annotated by InterVar to assess the deleteriousness of variants.

Frequency annotation The genome of each human contains approximately 20 genes that are completely inactivated by LoF variants. The majority of these seemingly deleterious but mostly benign variants has allele frequencies of 1% or greater [40]. Consequently, the interpretation of a genetic variant must consider its population-wide frequency to exclude variants that appear harmful, but are too common to be the sole cause of a rare disease. Although filtering by allele frequencies and genotype counts is powerful to reduce the number of candidate variants, there are two major limitations. First, variants that have not been observed in any reference data set will be enriched for sequencing artifacts specific to the laboratory where the variants were found [97]. Second, genetic variation in populations not represented in reference data sets can seem rare although it is frequent in the corresponding population. In the context of the analysis of the WES cohort, the first issue is tackled by including allele frequencies within the cohort in the filtering procedure. Due to the size of the cohort, recurrent sequencing artifacts can properly be detected and removed. The underrepresentation of Middle Eastern and South Asian populations in the gnomAD database is compensated by the stratification of the cohort according to the ethnic origin of the probands, and the subsequent use of population-specific allele frequencies during variant prioritization. This approach revealed that the WES cohort consists of six distinguishable populations that could be assigned to six geographic ethnicities using the ethnic origins reported by the individuals included. Based on this stratification, 59,921 population-specific variants were detected and excluded from the variant set corresponding to 19.15% of all frequency-filtered variants.

Segregation filtering In addition to the annotation of the functional impact and the frequency of a variant, the comparison of its segregation pattern with the observed inheritance pattern of the disease helps to evaluate its pathogenicity. While variant calling enables the differentiation between hetero- and homozygous genetic variants as potential mono- or biallelic causes of disease, the assessment of compound heterozygous variants is more challenging. First, it is necessary to enumerate all combinations of potential compound heterozygous variant pairs for each gene. Second, phasing algorithms are required to compute for each pair whether both variants were inherited from the same parent or one was inherited from the mother and the other from the father. To meet these challenges, we have developed the tool SmartPhase to enable fast and accurate phasing of single genes or specific variant pairs. For this purpose, SmartPhase creates haplotypes by means of trio phasing, read-based phasing and the integration of existing phasing information. Additionally, it is able to identify variant combinations as benign through logical rules. After we have shown the usability and accuracy of SmartPhase in the original publication [111], SmartPhase recently enabled the diagnosis of a patient suffering from a chronic interstitial lung disease by identifying a compound heterozygous defect in *FARSA* [178]. The application of SmartPhase to the WES cohort resulted in the exclusion of 58,741 variant pairs or 59% of the initial set of 99,300 potential compound heterozygous variant pairs. The remaining set of 40,559 variant pairs most likely still contains a considerable proportion of pairs located on the same allele as a consequence of the limitations of the applied phasing strategies. First, trio phasing requires parental sequencing data, which was available only for 15.49% of all 1,678 patients. Second, the efficiency of read-based phasing strongly depends on the length of the generated sequencing reads. As the length of the reads of NGS technologies is limited to a few hundred base pairs, the length of initially reconstructed haplotypes is in a similar range. Even though initial haplotypes can be extended by using paired-end reads and combining overlapping haplotypes, the read length is the most limiting factor of read-based phasing. However, the effect of both limitations will weaken in the future when more parents can be sequenced due to declining sequencing costs and longer reads become available due to the use of long-read sequencing technologies.

4.1.2 Variant filtering

When filtering variants there is always a risk to loose the variant one is searching for. This applies to variant prioritization based on annotated criteria, such as impact or frequency, but also to the preceding removal of supposedly false-positive variants from the initial variant call set. Therefore, it is necessary to check thoroughly whether the chosen filter criteria are appropriately defined. The validity of the filter criteria used for the analysis of the WES cohort was evaluated in three ways. First, the Ti/Tv ratio was assessed after the removal of low quality variants and the exclusion of samples with low sequencing coverage. The increase of the ratio from 1.93 to 2.09 in the set of kept variants and the decline to 1.00 in the set of discarded variants confirms that the quality control measures successfully removed false-positive variants. Second, all pathogenic variants in the WES cohort were used to evaluate the efficiency of quality and frequency filtering, which are part of the search for novel candidate genes. Of the 328 pathogenic variants, 14 or 4.27% variants were not found due to a lack of sequencing reads or because they failed criteria for variant quality or frequency. The percentage is minor in comparison to the achieved reduction of the initial variant set from $4,011,777$ variants by 63.69% to $1,457,066$ variants. This demonstrates that the selected quality and frequency filter criteria reflect the focus on the identification of fewer but promising candidate genes, while accepting that not all patients can be diagnosed in the cohort analysis. Third, criteria based on the impact annotation were evaluated for their individual effect on filtering variants and their combined effect on identifying candidate genes. The individual analysis of the criteria based on the annotations of SpliceAI, LOFTEE, InterVar and CADD showed that all of them achieve a minimum recall of 83.76%. The combination of all criteria results in a recall of 67.24% of the known disease-associated genes in the cohort. This leads to the estimate that approximately 30% of all genes that have not previously been associated with IEI and that have a causal defect in at least two unrelated patients in the cohort would not be reported as candidate genes by the CIP. A higher recall could be achieved by relaxing the filter criteria in the search for novel gene-disease associations, but this would inevitably increase the number of reported candidate genes. As the analysis of the WES cohort has already identified 205 candidate genes, this step will only be necessary when disease associations could be verified or excluded for all of the reported genes.

4.1.3 Implementation

Increasing the diagnostic rate of NGS studies on rare disease patients is a major challenge [125]. A study based on WES data of $1,133$ children with severe developmental disorders achieved an increase of the diagnostic rate from 27% to 40% through a reanalysis of all exome data three years after the first analysis [118]. The authors found that the majority of the new diagnoses resulted from newly discovered gene-disease associations. Similarly, the recognition of the variant in *SRPRA* in the index patient of family SCN-1 as a top candidate arose from the new insight that the interaction partner SRP54 is involved in neutropenia rather than the use of WGS after initial WES. Both examples demonstrate that the iterative reanalysis of WES data can be highly valuable to increase the diagnostic rate of exome sequencing studies. Recently, Appelbaum *et al.* even proposed an ethical duty to reinterpret genetic data when considering the likelihood of substantial patient benefit [179]. The need for iterative reanalysis reinforces the general requirements of proper and thorough data analysis. All steps of data analysis pipelines must be reproducible and well documented. Further, pipelines should be stable and easy to use to reduce the effort of starting and performing the desired analysis. The latter requirement is especially relevant when data should be reanalyzed regularly.

Both the published KNIME extension KNIME4NGS and the developed KNIME-based CIP fulfill these criteria. In general, KNIME workflows are implicitly documented and the data processing can easily be reproduced by re-executing the corresponding workflow. Through the graphical user interface, workflows can easily be created and altered, and the state of execution is always visible. The HTE of KNIME4NGS compensates for instabilities arising from the consecutive execution of integrated tools. The stability of the CIP results from the use of the provided stable KNIME nodes. While KNIME4NGS enables the creation of various workflows, the CIP represents a configured workflow for the identification of candidate genes. The division of variant prioritization in multiple steps makes it highly flexible as each filter criterion can be altered independently and the effects can be evaluated step by step. Consequently, the CIP can be considered as a template workflow that can be adapted quickly and easily to individual requirements. The generated candidate lists contain all required information helping to find diagnoses that were missed in previous analyses, and supporting the review of criteria annotated for potentially novel gene-disease associations. Taken together, KNIME4NGS is highly valuable for stable processing of sequencing data and the CIP is an important resource to support the iterative analysis of the growing WES data cohort at the Dr. von Hauner Children's Hospital.

4.2 Opportunities and limitations of the WES cohort

This thesis presents the first systematic analysis of the entire WES data set collected at the Dr. von Hauner Children's Hospital. Consisting of 1,678 patients and 634 healthy relatives, the cohort seems rather small compared to the currently largest published exome sequencing data set gnomAD with 125,748 individuals, but compared to other clinical exome sequencing studies, it is in the upper range. A meta-analysis of clinical WES lists 21 studies having between 31 and 3,040 probands enrolled, corresponding to an average of 477 probands [4]. Thus, the examined cohort is considerably larger than others and enables analyses that would not be possible or less reliable in cohorts of a few dozen or only a few hundred patients. First, the granularity of population stratification and the power of subsequent filtering of population-specific variants increases with the size of the cohort. A previous analysis of the cohort based on a total of 1,562 individuals resulted in a stratification into only four instead of six populations, for example. Second, the estimation of the impact and efficiency of the individual variant prioritization steps is the more reliable the more patients are included in a cohort and the more diagnoses are known. The set of 328 pathogenic small genetic variants allowed to determine filter criteria ensuring high sensitivity while generating concise lists of new candidate genes as discussed in Section 4.1.2. Third, the size of the cohort and the presence of overlapping pathophenotypes makes the identification of new candidate genes promising. Identifying defects in the same gene that are harbored only by a few patients worldwide requires a comprehensive collection of patients with a certain phenotype. Even though the cohort can be separated into patients suffering primarily from IBD, SCN or other immune defects, phenotypes are often overlapping and there are numerous examples of the same genetic defect resulting in differing pathophenotypes. For this reason, the cohort represents a unique opportunity to search for new candidate genes that trigger IEI. Despite the many opportunities afforded by the size and composition of the cohort, there are several limitations. Some of them arise from technical constraints of WES or generally NGS, while others are specific to clinical WES cohorts.

4.2.1 Technical limitations of WES and next-generation sequencing

Conceptually, WES is not able to identify the majority of genetic variation in non-coding regions of the human genome. Nevertheless, the library preparation kits used for the analyzed WES cohort also capture UTRs and flanking intron regions of exons. While the latter enabled the detection of two pathogenic intron variants, the analysis of potentially damaging UTR variants did not result in any plausible candidates. There might be no pathogenic UTR variation in the cohort or, more likely, the pathogenicity of identified variants was not recognized, because the understanding of the functional impact of UTR variants is still limited.

Library preparation of WES includes capturing of protein-coding DNA fragments and their amplification by polymerase chain reaction before sequencing. Both steps introduce bias causing an uneven coverage distribution across the exons of the human exome. Furthermore, the exact coverage distribution varies for each batch of chemicals used. Consequently, the detection of copy number variation is challenging because unusual high or low sequencing depth does not necessarily reflect a duplication or a deletion, but can result from the use of another batch of chemicals or other deviations during library preparation. Especially cohorts of patients that were sequenced over several years using the most recent library preparation kit at the time of sequencing are affected by variation during library preparation. As this applies to the analyzed WES cohort, an analysis of copy number variants on a larger scale was not performed.

Further limitations arise from the shortness of sequencing reads generated by NGS technologies. First, submicroscopic structural variants, such as insertions, translocations, or inversions, are hard to find because single reads do not cover the entire variant site. As a consequence, an analysis of structural variants was not performed for the WES cohort. Second, short reads offer only limited ability to recreate haplotypes to assign heterozygous variants to one of both paternal alleles. As discussed in Section 4.1.1, SmartPhase and other tools using read-based phasing will therefore perform more efficient when new sequencing technologies are established that generate reads longer than a few hundred base pairs. Third, short reads make it impossible to reliably cover regions of high sequence homology in the human genome. Depending on the percentage of identity, mapping of sequencing reads generated from homologous regions is completely impossible or results in misalignments and subsequently in unreliable variant calls. In the context of clinical WES, this is especially a problem when homologous regions overlap with protein-coding regions because there is always a chance that undiagnosed patients carry an unidentifiable pathogenic variant in a highly homologous exon or exon segment. Overall, there are $7,691$ exons of $1,168$ genes that have at least 98% sequence homology to other regions in the human genome at the DNA level [180]. These genes include 18 of the 408 genes reported to be causative for IEI [32], such as *NCF1*, where recombination events with highly homologous pseudogenes cause chronic granulomatous disease [181]. Consequently, whenever the pathophenotype of a patient is highly similar to the ones described for the 18 genes, other sequencing approaches should be considered to exclude pathogenic variants in the homologous regions of these genes.

4.2.2 Other limitations of clinical WES cohorts

Clinical WES faces several challenges that have an impact on the ability to analyze the generated data. The extent to which specific issues are encountered depends primarily on the exact composition of the respective cohort. Consequently, the points discussed below cannot be fully generalized, but are also not unique to the analyzed cohort.

Only parents of 15.49% of all patients have been sequenced in order to maximize the number of sequenced patients at the same cost. While this strategy gives as many children as possible the chance to get a genetic diagnosis, it reduces the chance of diagnosing *de novo* variants, as these may be overlooked in the set of rare heterozygous variants. This issue is illustrated by the fact that screening variants in known IEI genes found almost three times more monoallelic than biallelic variants. Following the proposal of pooled parent exome sequencing [182], parents are now sequenced at low coverage at the Dr. von Hauner Children's Hospital when patient only sequencing did not result in a good candidate. However, these sequencing data sets were excluded from the cohort analysis to avoid the unintended exclusion of candidate genes for novel disease associations as a consequence of supposedly deleterious variants in healthy individuals that are actually false-positives introduced by low coverage parental sequencing data.

The fact that at least one genetic variant in one of the 408 IEI genes was identified for 94.46% of the patients in the cohort when applying the frequency filtering criteria defined in Section 2.3.6, suggests that the increasing number of discovered disease genes will result in an increasing number of patients for whom at least one variant in a disease-associated gene will be found. Despite the specification of additional stringent filter criteria for candidate gene selection, the CIP still identified 205 candidate genes. This is a large amount, when considering that these genes have to be checked manually to evaluate whether they might be relevant. Both issues increase the risk that potentially important variants or genes are ignored due to multiple types of bias during manual review. Depending on individual experience and training, the focus may be on variants in known genes or on genes with an intuitive link to the disease of interest [42]. The effect of bias grows with the length of the resulting candidate lists, as the probability increases that the list contains a candidate that seems appropriate. Besides the percentage of trios, the length of candidate lists is primarily influenced by the number of individuals in a cohort and the strictness of the used filtering criteria. Enlarging patient cohorts without being able to stratify patients increases the number of genes affected by deleterious mutations in multiple patients. In order to restrict the length of candidate lists, more stringent filter criteria are applied, which in turn increases the probability of missing disease-causing genes. To tackle this issue, while still enabling the growth of clinical WES cohorts, it is necessary to characterize the phenotype of patients systematically to be able to identify groups of similar patients as a measure to reduce the search space and shorten candidate lists.

For this purpose, the Human Phenotype Ontology (HPO) provides a standardized vocabulary of phenotypic abnormalities observed in human disease [183]. The HPO enables the computation of ontology-based semantic similarity between any entities annotated with HPO terms, such as patients, genes, or diseases [184]. The tool Exomiser [185], for example, makes use of this feature to compute the similarity of a patient's phenotype to the known phenotypes of disease genes to integrate phenotypic information in variant prioritization. Despite the advantages of the HPO, there are several reasons why the patients of the analyzed WES cohort are only roughly characterized as IBD, SCN or other syndromic immune defect patients. First, the HPO does not yet capture all phenotypic features observed in different clinical disciplines. In addition to many other groups, some colleagues and I contributed new terms to the HPO recently [186]. While the vocabulary of the HPO will become increasingly complete in the future, other challenges regarding the application and the use of the HPO are still open. HPO terms can be extracted automatically from patient reports or annotated manually. The former approach ensures a uniform and fast annotation of patients, but has to tackle the various challenges of text mining, such as different languages, negations, or the use of individual abbreviations. Manual annotation requires much more time and rules to harmonize the annotation across all responsible persons. For both approaches it has to be defined whether a predefined set of features should be queried for every patient, all phenotypic features should be annotated, or only diagnosis-relevant features should be collected. Even though there are efforts to standardize the storage of phenotypic information as so-called Phenopackets [187], there are still only few tools available that make use of them. Notwithstanding these challenges, thorough phenotyping of patients seems highly promising to draw further insights from the collected WES cohort.

4.3 Importance of the discovered pathogenic variants

Most of the 293 distinct pathogenic genetic variants reported as diagnosis in the WES cohort are part of biallelic genotypes, alter the protein-coding sequence, and have evidence to be harmful through ACMG/AMP criteria and/or CADD scores higher than 15 (see Section 3.3.5). On the one hand, this observation confirms that the chosen criteria for variant prioritization are suitable for the search for pathogenic variants as discussed in Section 4.1. On the other hand, these properties result in part from technical limitations and an incomplete understanding of cellular processes. For example, there is a bias towards biallelic variants because the detection of *de novo* variants is limited through the low amount of trio sequencing. The pathogenic variants are almost exclusively located in protein-coding regions, since pathogenic variants in non-coding regions can only be comprehensively searched for by means of performing WGS on a regular basis along with a better understanding of the functional mechanisms of regulatory elements.

Although WES is designed to cover only exons, the used library preparation kits enable the identification of variants in the flanking intron regions of the captured exons. This resulted in the discovery of an intronic variant that introduces a new splice donor site in the gene *DOCK8* (c.3027+76A>G) in addition to a branch point mutation in *DKC1* (c.85-29T>G). Both variants demonstrate that WES analysis should be extended to flanking intronic regions if these are covered by the sequencing kits used. Additionally, the variant in *DKC1* exemplifies how splicing defects can be caused by variants other than ones that disrupt or generate splice acceptor or donor sites.

As its greatest advantage, WES enables an exome-wide search for pathogenic variants in all covered genes without restricting to already known disease-associated genes. Consequently, it can uncover the genetic cause, even if the initial phenotypic diagnosis was misleading, and beyond that, identify new gene defects and associated pathways. The first point is illustrated by the identified pathogenic variants in *EPCAM* and *SLC5A1*. The diagnosis of the 13 affected patients was only possible because the analysis was not restricted to IBD genes, although the pathophenotype suggested IBD as underlying disease. Especially for patients with IBD-like symptoms it is critical to consider all potentially pathogenic variants. The patients with defects in *CARMIL2*, *FOXP3*, *G6PC3*, *SRP54* and *RTEL1* all showed a combination of VEOIBD and other phenotypic features that extend the known phenotypic spectrum of defects in these genes. Focusing on genes related to the most prominent symptoms, or excluding genes because the observed phenotypes are not perfectly fitting to the already described pathophenotypes of defects in these genes, might have missed the causative variants. Further, these examples show that VEOIBD is often a manifestation of immune dysregulation rather than a disease with its own specific genetic etiology [188]. Prioritization of the variants in the WGS data of family SCN-1 and the WES data of family SCN-9 without restricting the search space to known disease-associated genes pointed to defects in two SRP genes as novel causes for SCN. The identification of novel and reported pathogenic variants in *SRP54* in the WES cohort gave additional evidence that SRP-dependent co-translational protein targeting to the ER membrane is a vulnerable process especially for the generation of neutrophils. This finding is further substantiated by the report of a patient suffering from SCN caused by a biallelic pathogenic variant in the gene *SRP68* [189].

Schürch *et al.* recently investigated *SRP54* deficient zebrafish in order to understand the functional mechanism of *SRP54* defects [190]. The authors report that homozygous knockouts are lethal while heterozygous knockouts show only mild neutropenia. However, they were able to aggravate neutropenia in heterozygous fish, and induce pancreatic defects by the injection of mRNAs carrying the mutations described by Carapito *et al.* [120]. Similarly, overexpression of mutated *SRP54* in wild-type fish induced neutropenia. Additionally, they show that impaired unconventional splicing of *XBP1*, one of the key transcription factors involved in unfolded protein response, drives the SDS-like phenotype by rescuing neutropenia through the injection of spliced XBP1 into zebrafish embryos. Three important conclusions can be drawn from these results. First, the pathogenicity of monoallelic mutations in *SRP54* is due to a mutation-specific dominant-negative effect as mutated *SRP54* caused neutropenia in wild-type fish and the severe SDS-like phenotype in heterozygous knockout fish. Therefore, most probably a dominant-negative effect also underlies the pathogenicity of the *de novo* mutation in *SRPRA* in patient SCN-1pa (p.Gln464Glu). Second, the mild neutropenia phenotype of heterozygous knockout fish without pancreatic involvement indicates that variants in *SRP54*, other than LoF variants, could exist that cause a less pronounced or even no pathophenotype when they are monoallelic, but cause yet unknown disease phenotypes in case they are biallelic. This hypothesis supports the assumed pathogenicity of the homozygous missense variant in *SRP54* in patient VEOIBD-6pa with weakly affected heterozygous carries in the family. Third, a more detailed hypothesis of the pathomechanism of the *SRPRA*, *SRP54* and *SRP19* mutations can be derived. As discussed in Section 3.2 and 3.5.4, all observed mutations impair the functionality of the SRP complex. As a consequence, unconventional splicing of XBP1 is hampered because the process requires SRP-dependent transport to the ER [191]. Due the lack of spliced XBP1, unfolded protein response cannot be initiated, leading to unresolved ER stress and ultimately to myeloid maturation arrest.

Taken together, the variant prioritization approach in my work allowed the discovery of pathogenic variants that revealed a new pathomechanism of SCN and demonstrate that pediatric IBD can indicate underlying IEI but also be the consequence of other gene defects. Moreover, the variants reflect the general observation that defects in the same gene may cause differing pathophenotypes just as defects in different genes can result in highly similar pathophenotypes. The effects of the described pathogenic variants are, as expected, very diverse and range from biallelic inactivation of genes by LoF variants, to disturbed gene function by homozygous missense variants in functional domains, to splicing defects by homozygous splice site alterations. Furthermore, they include a hemizygous branch point mutation, as well as dominant-negative effects through *de novo* variants. The analysis of the functional context of the individual genes further shows that there is no single biological function or cellular pathway that underlies all IEI. Rather, multiple pathways and functions are necessary to ensure proper functionality of all components of the immune system.

Conclusion and Outlook 5

This chapter summarizes the major achievements of my work and the novel contributions to the fields of NGS data analysis, IEI in particular and rare disease genetics in general. Further, it discusses directions for further analysis of the data to overcome shortcomings of the existing approaches and to make use of the opportunities offered by new technologies.

5.1 Contributions to NGS data analysis and the etiology of IEI

This work was motivated by the idea of using the entire WES data collection at the Dr. von Hauner Children's Hospital to extend the current knowledge on genetic causes of IEI. The development of new bioinformatic tools and routines played an important role in achieving this goal and resulted in three major contributions to the field of NGS data analysis. First, I was part of the team developing the KNIME extension KNIME4NGS that makes it possible to compile NGS workflows easily in a modular way [116]. Second, I was leading the design and implementation of SmartPhase enabling efficient and accurate phasing of variants of clinical interest [111]. Third, I established a novel pipeline for the joint and iterative analysis of the growing amount of WES data at the Dr. von Hauner Children's Hospital. Besides quality control at variant and sample level, variant annotation based on publicly available data and derived population-specific allele frequencies, the KNIME workflow for variant prioritization, termed CIP, is the key element to identify novel genetic defects underlying IEI.

The application of the developed tools and pipelines to WES data from $1,746$ patients and 705 healthy relatives, as well as to WGS data of one family, enabled definitive diagnoses in a substantial number of children with life-threatening diseases and generated several new insights into the pathogenesis of IEI. The identified pathogenic variants in the patients DIAR5-1pa to DIAR5-7pa, GGM-1pa to GGM-6pa, and VEOIBD-1pa to VEOIBD-7pa demonstrate that IBD-like symptoms can be the manifestation of an inborn error of immunity but also the consequence of other genetic defects. The published variants in *CARMIL2* [139], *FOXP3* [143], *G6PC3* [144], *SRP54* [144] and *RTEL1* [151] show that

IBD-like symptoms can be the first manifesting or most dominant phenotype of defects in these genes, while the described mutations in *EPCAM* and *SLC5A1* show that not only variants in IEI genes should be considered in pediatric IBD. Independent of the underlying etiology, an early genetic diagnosis is crucial because diseases like GGM can be treated efficiently through dietary adjustments, and IEI can be cured by allogeneic hematopoietic stem cell transplantation.

With regard to SCN, my work contributed considerably to the finding that the SRP and its receptor play a critical role for the proper development of neutrophils. The analysis of the WGS data revealed a *de novo* mutation in *SRPRA* as the cause of SCN with SDS-like phenotype. In addition to the identification of a reported pathogenic *SRP54* variant in five patients, the analysis of the exome data detected two novel pathogenic variants in *SRP54* in two unrelated patients and the same homozygous pathogenic mutation in *SRP19* in four closely related patients. Besides the first description of two novel human genetic defects in *SRPRA* and *SRP19*, we identified the disturbance of proteostasis and subsequent apoptosis of neutrophils as the main underlying pathomechanism for SRP-related mutations and as the main reason for the observed SCN phenotypes of the investigated patients (Y. Mizoguchi, S. Hesse, M. I. Linder, *et al.*, manuscript in preparation).

The discussed intronic *DKC1* mutation is the first pathogenic branch point variant that was detected in the WES cohort at the Dr. von Hauner Children's Hospital. Because branch point sequences are highly variable and extremely degenerate, their identification is challenging and it is hard to predict the possible effect of specific variants [157, 192]. As a consequence, only few examples of pathogenic branch point mutations are known. Therefore, the identified variant and functional experiments represent an important contribution to the awareness of the relevance of branch point mutations in rare disease genetics.

In addition to these findings, my work will also have a major influence on the future analysis of the growing WES data set at the Dr. von Hauner Children's Hospital. First, the implemented quality steps ensure that anomalies like duplicate, mislabeled, or contaminated samples can easily be identified and resolved, or removed without disrupting downstream analysis. Second, the CIP can be used to re-iterate the presented analysis when new gene-disease associations will have been discovered in order to identify new genetic diagnosis, to screen for new candidate genes after extending the WES cohort, or to analyze sub-cohorts with less stringent filter criteria when thorough phenotyping enables fine-grained stratification of patients into groups of similar pathophenotypes. Finally, the gained insights have direct impact on routine analysis by drawing more attention to variants in genes not associated with IEI in VEOIBD patients and to variants in all genes in the SRP complex and its receptor as potential causative in SCN patients.

In conclusion, I successfully implemented new tools and pipelines for the analysis of the collected WES data at the Dr. von Hauner Children's Hospital and gained new insights into the pathogenesis of IEI. Along the way, two points have emerged to be particularly critical for the search for genetic causes of rare disease in sequencing data of cohorts larger than a few dozen patients. First, there is a need for elaborate variant prioritization using multiple filter criteria to keep the resulting lists of candidate variants and genes manageable. Second, each filter criterion should be as independent as possible from already known gene-disease relationships to avoid bias towards functionally comprehensively described genes and to enable the discovery of novel genetic defects and underlying pathomechanisms.

5.2 Future directions

In spite of the progress made, the majority of the patients in the WES cohort at the Dr. von Hauner Children's Hospital is still lacking a genetic diagnosis. In general, there are a variety of reasons why WES analysis is not able to find the underlying genetic defect. The disease of some patients might not be caused by a monogenic defect, but results from some complex interplay of genetic variants and environmental factors. If there is a single causative genetic defect, WES might not be able to identify it, because the causative variant is located in a non-coding region or it is not discovered by short-read sequencing technologies. If the causal defect is identified, we might not recognize it, because it fails filter criteria, or is observed the first time without any evidence linking the affected gene to the pathophenotype. Finding a second patient with a defect in the same gene would be the most solid evidence of a potential gene-disease association. However, so far unidentified genetic defects will be less frequent than known defects, otherwise they would have been found already. Consequently, the chance of finding more than one patient with pathogenic variants in the same gene, if such patients exist at all, is unlikely when the size of a cohort is not increased. However, larger cohorts will also generate more random hits making it more challenging to differentiate real signals from noise. Additionally, cohorts of patients affected by a particular rare disease are often not easily extendable, because there might just be a few dozen or only a few hundred patients world-wide.

Because these challenges apply to most rare disease cohorts, opportunities for improvement are discussed in multiple recent publications [125, 193, 194]. Structured phenotyping of patients is a promising way to find new gene-disease associations in phenotypically highly similar patients as discussed in Section 4.2.2. Phenotyping also makes it possible to expand sequencing data sets without introducing noise into the analyses, because the type and severity of the underlying pathophenotype is not determined through the selection of patients to be sequenced, but can be taken into account during analysis. Besides the expansion of sequencing data sets, the role of non-coding variation in rare disease will be better understood as the use of WGS becomes more widespread. Additionally, multi-omics assays will help to prioritize and understand the molecular functional consequence of genetic variants.

5.2.1 Enlargement of cohorts of sequenced individuals

Following the aim of expanding our knowledge on human genetic variation and driven by the decreasing costs of NGS, increasingly larger cohorts of healthy and diseased individuals are getting sequenced or existing data is combined into larger cohorts. While the publication of the 1000 Genomes data set in 2015 included sequencing data of $2,504$ individuals [17], the gnomAD data set comprised already the exomes and gnomes of $141,456$ humans in 2020 [65]. A similar growth can be observed for cohorts of patients with rare diseases. While a review of genome and exome sequencing studies in children with suspected genetic diseases was based on studies between 31 and $3,040$ enrolled participants in 2018 [4], a study published in 2020 performed WGS of already $13,037$ participants [176]. Noticeably, the latter study carried out by the National Health Service in the United Kingdom plans to further perform WGS for $30,000$ individuals per month across multiple clinical genomics laboratories. Obviously, the volume of sequencing by the NGS facility at the Dr. von Hauner Children's Hospital as a single site is limited, but shall nevertheless comprise at least $1,000$ patients per year in the near future.

Both, the growth of reference and clinical data sets will improve rare disease diagnosis in several aspects. First, more comprehensive collections of genetic variants in individuals without severe disease will increase the number of variants that can be excluded as monogenic cause of disease in rare disease patients. Second, the identification of new LoF variants inactivating certain genes or a statistically significant lack of LoF variants will refine estimates of disease-related relevance of individual human genes [65]. Third, each new patient added to a cohort increases the chance to detect a genetic defect in more than one patient as a promising indication of pathogenicity, if the phenotypes are sufficiently similar. Fourth, larger patient cohorts may help to decipher genetic causes more complex than monogenic defects, such as digenic defects. The latter refer to a combination of two genetic defects that are both required to induce a pathophenotype [195, 196]. In comparison to the thousands of monogenic defects reported in OMIM [22], only 90 of such true digenic combinations are reported in the Digenic Disease Database in the most recent update of July 2017 [197]. Although it is unknown how frequent digenic or oligogenic defects underlie rare diseases, these numbers reflect the low chance of finding more than one patient with rare defects in the same genes as each incrementation of the number of potentially involved genes adds an equal number of dimensions to the search space. Nevertheless, increasing the size of patient cohorts may also increase the ability to identify new digenic or oligogenic inheritance modes, thereby furthering the understanding of their role in rare diseases.

5.2.2 Replacement of WES by WGS

Although current evidence does not support a considerably higher diagnostic utility of WGS in comparison to WES [4, 5], WGS will replace WES in the near future due to an increasingly overlapping range of cost [198] and the prospect of greater long-term benefit through improved analysis and interpretation of the generated data. This development is evident in the large-scale study of the National Health Service in the United Kingdom [176], and also the Dr. von Hauner Children's Hospital wants to rely more on WGS in the future.

In contrast to WES, library preparation of WGS does not include any selection and amplification steps of DNA fragments. Therefore, WGS offers a more comprehensive coverage of the protein-coding regions of the human genome and a uniform sequencing depth over the whole genome. Consequently, the chance of missing pathogenic protein-coding variants is reduced and structural variation can be detected more accurately as deviations in the sequencing depth must result from duplication or deletion events and the identification of breakpoints is not limited to exons. However, the primary advantage of WGS over WES is the ability to find variants in the non-coding regions of the human genome. Yet the interpretation of non-coding variation is not as intuitive as for variants in protein-coding exons. There are many more genetic variants in a genome than in an exome and filtering for rare variants is less powerful due to the currently several times smaller sizes of reference cohorts based on WGS in contrast to those based on WES. Results of genome-wide association studies suggest to focus the analysis of non-coding variants on regulatory regions as these are enriched for associated SNVs [199, 200], but predicting the impact of non-coding variation in regulatory regions and validating derived hypothesis remains challenging. Although WGS is currently primarily useful to detect protein-coding variants more comprehensively, and to identify structural variants, new findings on non-coding variants, and algorithms supporting their interpretation will continue to make WGS even more valuable in the future.

5.2.3 Generation and integration of multi-omics assays

Multiple different omics technologies enable the study of molecular details of cellular processes in humans in health and disease on various levels [201]. Based on NGS techniques, genomics refers to the identification of genetic variation, epigenomics describes the study of DNA methylation and chromatin accessibility, transcriptomics measures the abundance and sequence of RNAs and microbiomics comprises the characterization of microorganisms populating the skin, mucosal surfaces, and the gut. Proteomics and metabolomics measure the abundance of proteins and metabolites by mass spectrometry. Many of these technologies can also be applied at single-cell level enabling even more detailed insights into cellular processes [202].

Finding methods to integrate the wealth of multi-omics data is crucial to improve the diagnosis of diseases, to understand the details of pathomechanisms, and to use gained insights to optimize treatment [203]. Especially RNA sequencing has been proven to be valuable to improve the genetic diagnosis of Mendelian diseases by enabling variant calling in exons not covered by WES and linking variants without clear functional impact to aberrant expression, aberrant splicing or allelic expression imbalances. Multiple studies showed that transcriptome profiling of patients that could not be diagnosed through WES yields an additional diagnostic rate between 7.5% and 25% [204, 205, 206, 207].

To understand the interplay of multi-omics layers in children, the Dr. von Hauner Children's Hospital recently launched a large-scale project aiming to collect retina images and blood samples of $5,000$ children to perform multi-omics assays and use the generated data pool to identify and characterize biomarkers of pediatric disease.

Taken together, the future directions outlined will all play an important role in improving our ability to discover genetic causes of rare diseases. They will also help us to comprehensively describe and understand the functional mechanisms linking genetic variation to observed pathophenotypes. Finally, this is the prerequisite for developing targeted therapies and getting closer to the ultimate goal of offering each patient the optimal treatment based on their individual genetic makeup.

Appendix 6

6.1 Detailed view of the Candidate Identification Pipeline

This section presents the individual parts of the implemented Candidate Identification Pipeline described in Section 3.1.2 and abbreviated as CIP. Figure 6.1 shows the whole KNIME workflow as a thumbnail with an identifier for each functional group of the CIP. Parts that import annotations on patients, variants and genes are highlighted by yellow (A) boxes, which are individually depicted in Figure 6.2 to 6.5. The generation of candidate variant lists are shown in Figure 6.6 to 6.11 for dominant and in Figure 6.12 to 6.20 for recessive effects, in green (B) and brown (C) boxes respectively. The compilation of lists summarizing data on patients and pathogenic variants is highlighted by red (D) boxes and depicted in Figure 6.21 to 6.23.

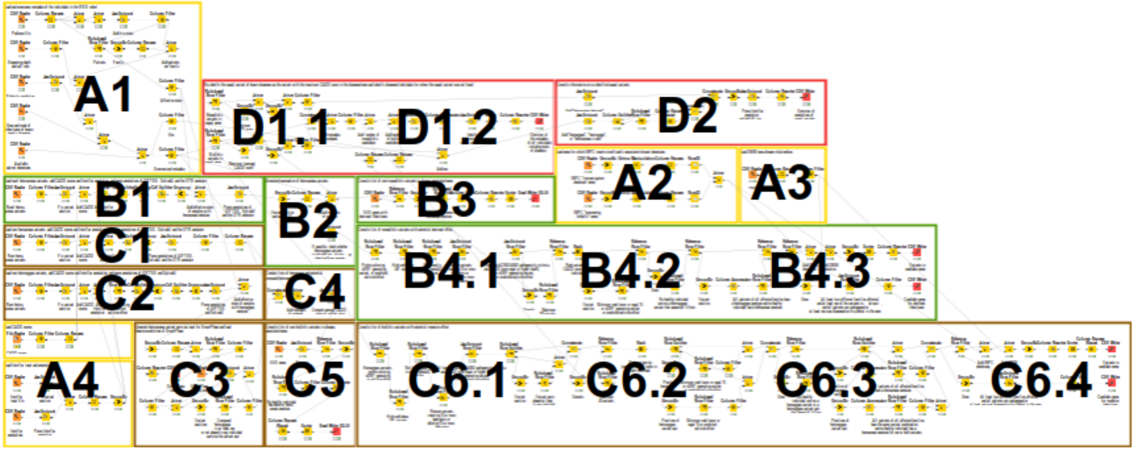


Figure 6.1: Identifiers for the individual parts of the Candidate Identification Pipeline (CIP). The figure shows the overall workflow consisting of 238 nodes. The identifiers were added to mark the positions of the individual parts shown in Figure 6.2 to 6.23.

A. Importing annotations

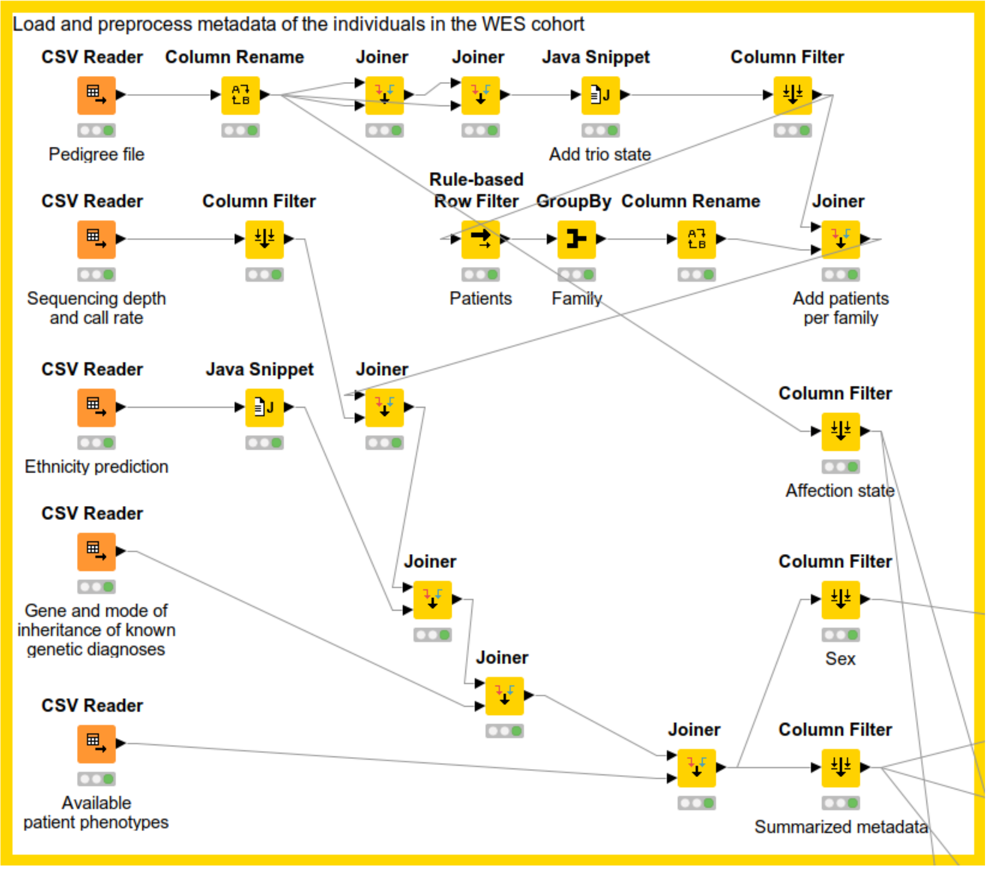


Figure 6.2: Part A1 of the CIP. The nodes import the pedigree file of the cohort, quality metrics of the exome sequencing (WES), the predicted ethnicity of each individual, known genetic diagnoses, and available patient phenotypes.

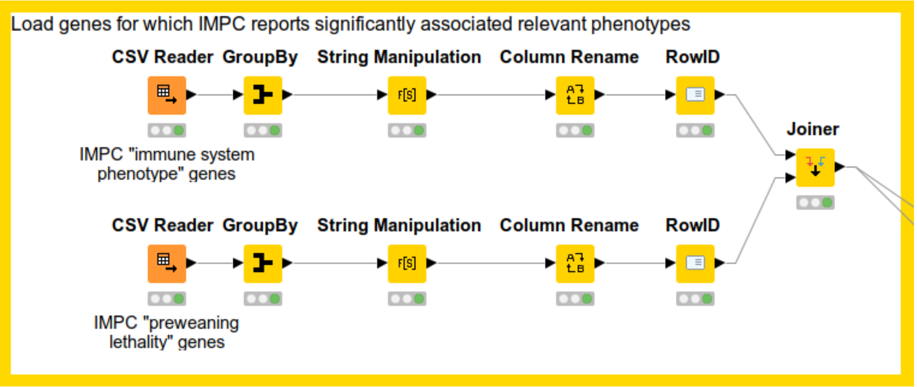

Load genes for which IMPC reports significantly associated relevant phenotypes
CSV Reader
GroupBy
String Manipulation
Column Rename
RowID
IMPC "immune system phenotype" genes
Joiner
CSV Reader
GroupBy
String Manipulation
Column Rename
RowID
IMPC "preweaning lethality" genes

Figure 6.3: Part A2 of the CIP. The nodes import all genes that are significantly associated with immune system phenotypes or preweaning lethality according to the International Mouse Phenotyping Consortium (IMPC) database.

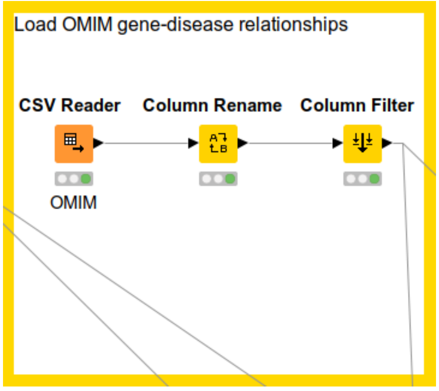


Figure 6.4: Part A3 of the CIP. The nodes import known gene-disease relationships provided by the Online Mendelian Inheritance in Man (OMIM) database.

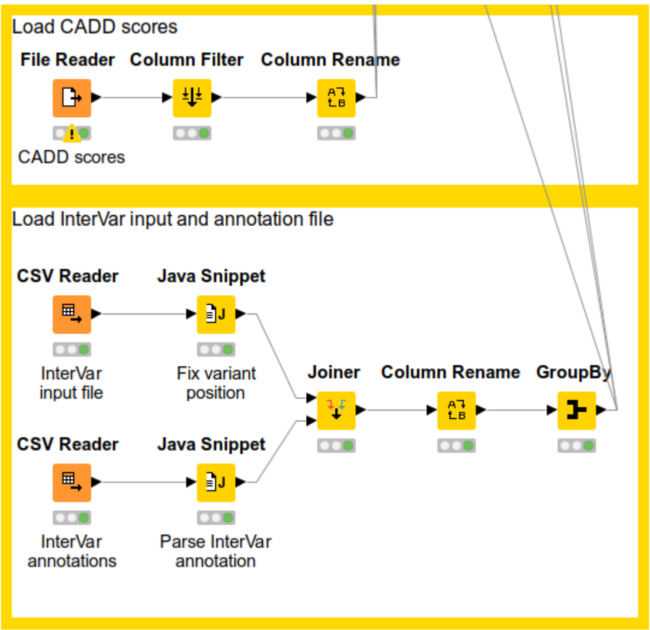


Figure 6.5: Part A4 of the CIP. The nodes in the upper part import the computed Combined Annotation Dependent Depletion (CADD) scores of all identified genetic variants. The nodes in the lower part read variant annotations of InterVar.

B. Dominant candidates.

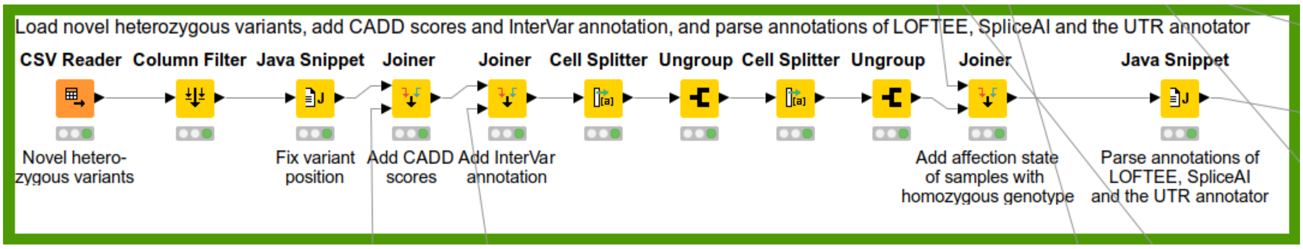

Load novel heterozygous variants, add CADD scores and InterVar annotation, and parse annotations of LOFTEE, SpliceAI and the UTR annotator
CSV Reader
Column Filter
Java Snippet
Joiner
Joiner
Cell Splitter
Ungroup
Cell Splitter
Ungroup
Joiner
Java Snippet
Novel heterozygous variants
Fix variant position
Add CADD scores
Add InterVar annotation
Add affection state of samples with homozygous genotype
Parse annotations of LOFTEE, SpliceAI and the UTR annotator

Figure 6.6: Part B1 of the CIP. The nodes import all genetic variants that are heterozygous in at least one patient of the cohort and that have not been reported in the used reference data sets. Variant annotations imported in part A4 (Figure 6.5) are added and prepared for further variant prioritization. The affection states of homozygous carriers of the imported variants are attached based on the metadata prepared in part A1 (Figure 6.2). CADD abbreviates Combined Annotation Dependent Depletion, LOFTEE refers to the Loss-Of-Function Transcript Effect Estimator and UTR stands for untranslated region.

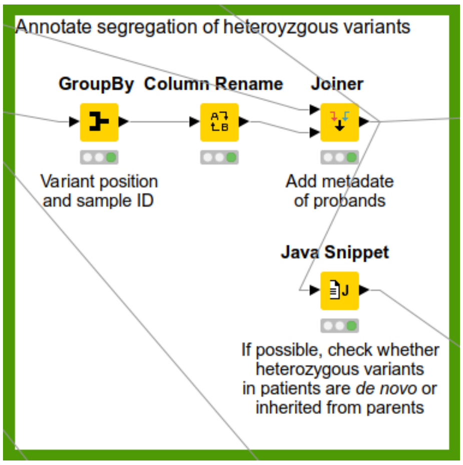

Annotate segregation of heteroyzgous variants
GroupBy
Column Rename
Joiner
Variant position and sample ID
Add metadate of probands
Java Snippet
If possible, check whether heterozygous variants in patients are de novo or inherited from parents

Figure 6.7: Part B2 of the CIP. If sequencing data of the parents of a patient is available, the nodes annotate whether potential dominant effects were inherited from diseased parents or are the result of a *de novo* event.

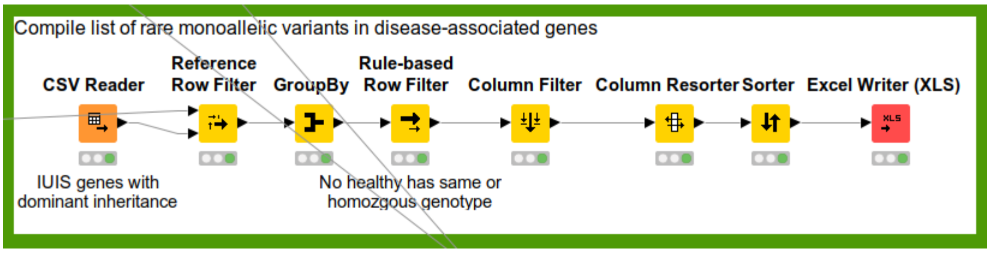

Compile list of rare monoallelic variants in disease-associated genes
CSV Reader
Reference Row Filter
GroupBy
Rule-based Row Filter
Column Filter
Column Resorter
Sorter
Excel Writer (XLS)
IUIS genes with dominant inheritance
No healthy has same or homozgous genotype

Figure 6.8: Part B3 of the CIP. The nodes compile a list of rare heterozygous variants only found in the patients of the cohort and in genes already associated with dominant inheritance of inborn errors of immunity as reported by the International Union of Immunological Societies (IUIS).

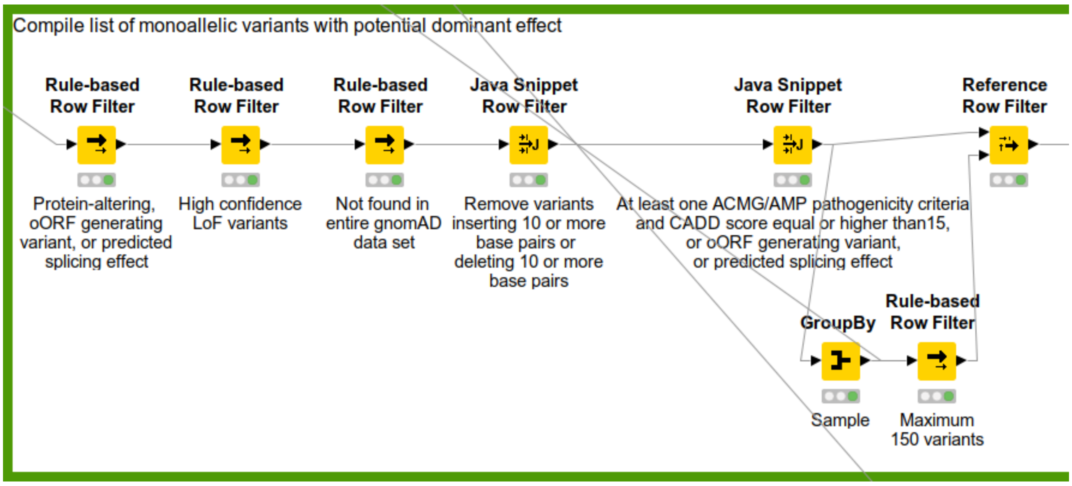

Compile list of monoallelic variants with potential dominant effect
Rule-based Row Filter
Protein-altering, oORF generating variant, or predicted splicing effect
Rule-based Row Filter
High confidence LoF variants
Rule-based Row Filter
Not found in entire gnomAD data set
Java Snippet Row Filter
Remove variants inserting 10 or more base pairs or deleting 10 or more base pairs
Java Snippet Row Filter
At least one ACMG/AMP pathogenicity criteria and CADD score equal or higher than15, or oORF generating variant, or predicted splicing effect
Reference Row Filter
GroupBy
Sample
Rule-based Row Filter
Maximum 150 variants

Figure 6.9: Part B4.1 of the CIP. The nodes filter the list of all heterozygous variants loaded in part B1 (Figure 6.6) according to the criteria defined in Section 2.3.6. Overlapping open reading frame is abbreviated as oORF, Loss-of-Function as LoF, Genome Aggregation Database as gnomAD, American College of Medical Genetics and Genomics and the Association for Molecular Pathology as ACMG/AMP, and Combined Annotation Dependent Depletion as CADD.

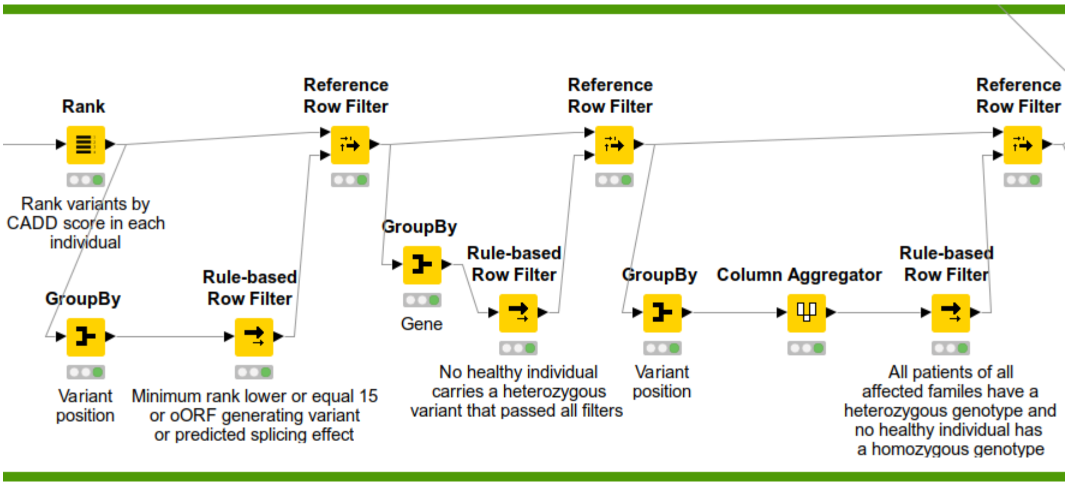

Rank
Rank variants by CADD score in each individual
GroupBy
Variant position
Rule-based Row Filter
Minimum rank lower or equal 15 or oORF generating variant or predicted splicing effect
Reference Row Filter
GroupBy
Gene
Rule-based Row Filter
No healthy individual carries a heterozygous variant that passed all filters
Reference Row Filter
GroupBy
Variant position
Column Aggregator
Rule-based Row Filter
All patients of all affected familes have a heterozygous genotype and no healthy individual has a homozygous genotype
Reference Row Filter

Figure 6.10: Part B4.2 of the CIP. The nodes continue the variant filtering shown in Figure 6.9 (part B4.1). Overlapping open reading frame is abbreviated as oORF and Combined Annotation Dependent Depletion as CADD.

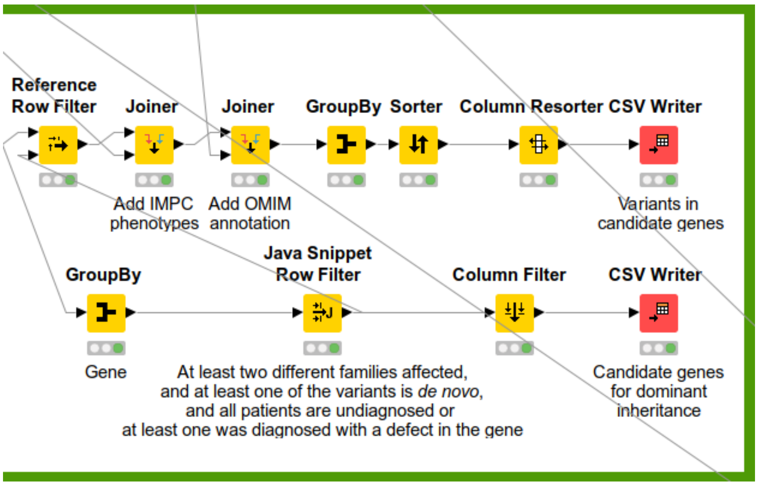

Reference Row Filter
Joiner
Add IMPC phenotypes
Joiner
Add OMIM annotation
GroupBy
Sorter
Column Resorter
CSV Writer
Variants in candidate genes
GroupBy
Gene
Java Snippet Row Filter
At least two different families affected, and at least one of the variants is de novo, and all patients are undiagnosed or at least one was diagnosed with a defect in the gene
Column Filter
CSV Writer
Candidate genes for dominant inheritance

Figure 6.11: Part B4.3 of the CIP. The nodes select candidate genes for dominant inheritance from the set of variants that result from the prioritization steps in part B4.1 (Figure 6.9) and B4.2 (Figure 6.10). Annotations of the International Mouse Phenotyping Consortium (IMPC) and the Online Mendelian Inheritance in Man (OMIM) database imported in part A2 (Figure 6.3) and A3 (Figure 6.4) are added.

C. Recessive candidates.

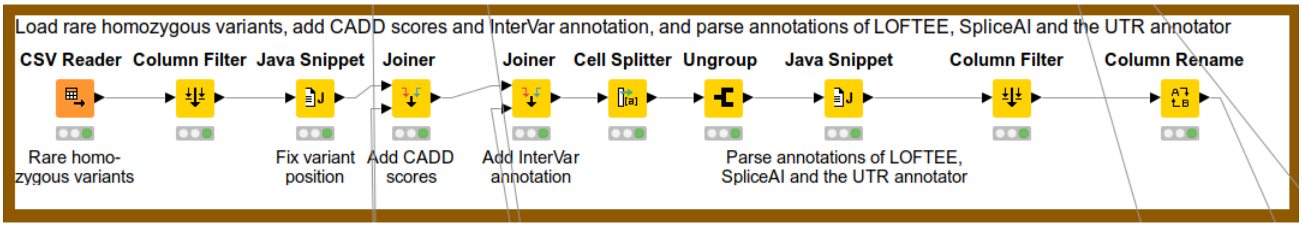

Load rare homozygous variants, add CADD scores and InterVar annotation, and parse annotations of LOFTEE, SpliceAI and the UTR annotator
CSV Reader
Column Filter
Java Snippet
Joiner
Joiner
Cell Splitter
Ungroup
Java Snippet
Column Filter
Column Rename
Rare homozygous variants
Fix variant position
Add CADD scores
Add InterVar annotation
Parse annotations of LOFTEE, SpliceAI and the UTR annotator

Figure 6.12: Part C1 of the CIP. The nodes import all genetic variants that are homozygous in at least one patient of the cohort and are rare according to the used reference data sets. Variant annotations imported in part A4 (Figure 6.5) are added and prepared for further variant prioritization. CADD abbreviates Combined Annotation Dependent Depletion, LOFTEE refers to the Loss-Of-Function Transcript Effect Estimator and UTR stands for untranslated region.

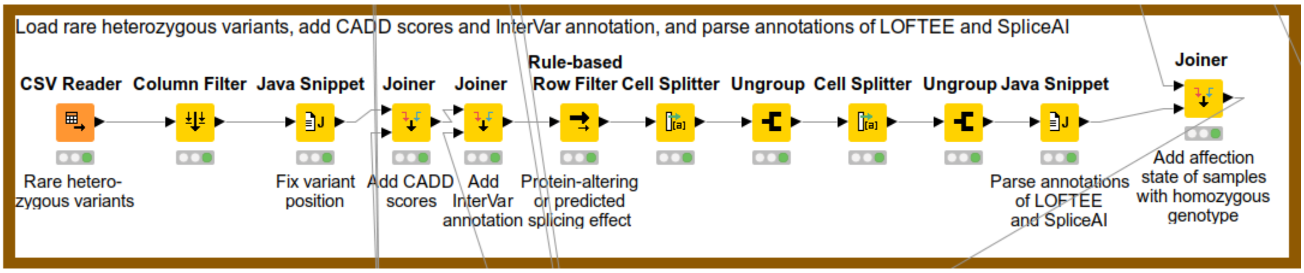
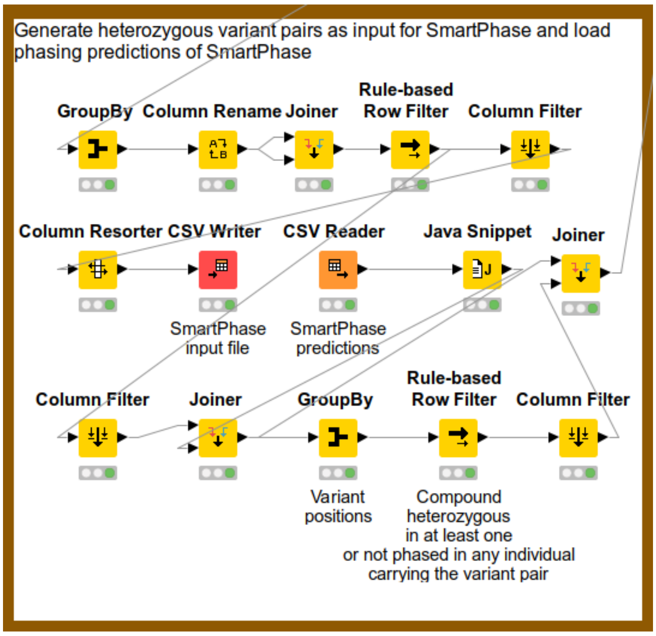
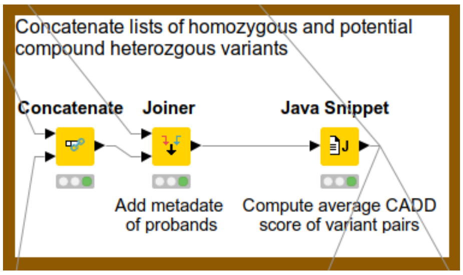

Load rare heterozygous variants, add CADD scores and InterVar annotation, and parse annotations of LOFTEE and SpliceAI
CSV Reader
Column Filter
Java Snippet
Joiner
Joiner
Rule-based Row Filter
Cell Splitter
Ungroup
Cell Splitter
Ungroup
Java Snippet
Joiner
Rare heterozygous variants
Fix variant position
Add CADD scores
Add InterVar annotation
Protein-altering or predicted splicing effect
Parse annotations of LOFTEE and SpliceAI
Add affection state of samples with homozygous genotype

Figure 6.13: Part C2 of the CIP. The nodes import all genetic variants that are heterozygous in at least one patient of the cohort and are rare according to the used reference data sets. Variant annotations imported in part A4 (Figure 6.5) are added and prepared for further variant prioritization. The affection states of homozygous carriers of the imported variants are attached based on the metadata prepared in part A1 (Figure 6.2). CADD abbreviates Combined Annotation Dependent Depletion, LOFTEE refers to the Loss-Of-Function Transcript Effect Estimator and UTR stands for untranslated region.



Generate heterozygous variant pairs as input for SmartPhase and load phasing predictions of SmartPhase

Figure 6.14: Part C3 of the CIP. Based on the rare heterozygous variants loaded by part C2 (Figure 6.13), the shown nodes generate pairs of heterozygous variants as input for SmartPhase. The results of SmartPhase are imported and variant pairs that are on the same allele or are considered benign based on their segregation pattern are discarded.



Concatenate lists of homozygous and potential compound heterozgous variants

Figure 6.15: Part C4 of the CIP. The list of heterozygous variant pairs generated in part C3 (Figure 6.14) is concatenated with the list of homozygous variants imported in part C1 (Figure 6.12). CADD abbreviates Combined Annotation Dependent Depletion.

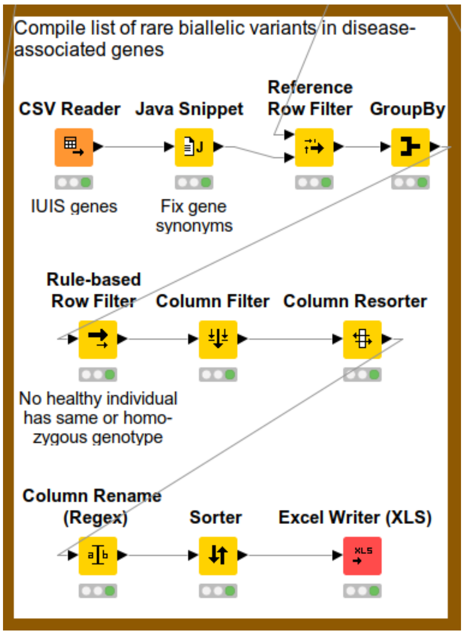


Figure 6.16: Part C5 of the CIP. The nodes compile a list of rare homozygous variants only found in the patients of the cohort and in genes already associated with inborn errors of immunity as reported by the International Union of Immunological Societies (IUIS).

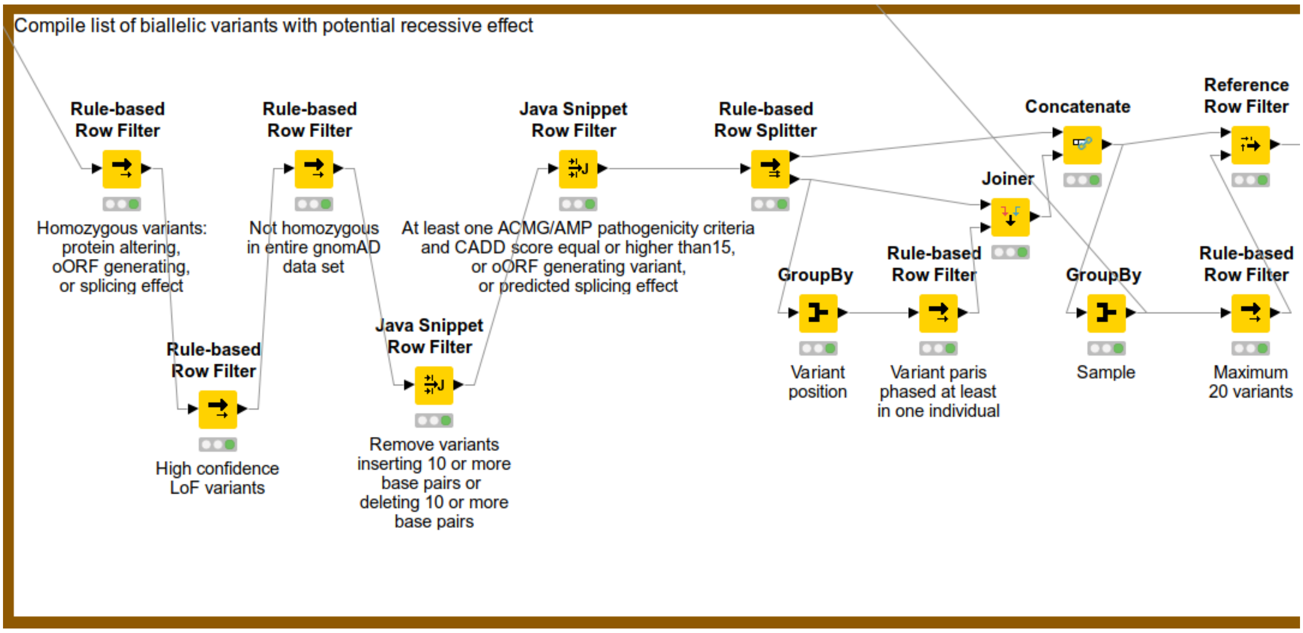

Compile list of biallelic variants with potential recessive effect
Rule-based Row Filter
Homozygous variants: protein altering, oORF generating, or splicing effect
Rule-based Row Filter
High confidence LoF variants
Rule-based Row Filter
Not homozygous in entire gnomAD data set
Java Snippet Row Filter
Remove variants inserting 10 or more base pairs or deleting 10 or more base pairs
Java Snippet Row Filter
At least one ACMG/AMP pathogenicity criteria and CADD score equal or higher than15, or oORF generating variant, or predicted splicing effect
Rule-based Row Splitter
GroupBy
Variant position
Rule-based Row Filter
Variant paris phased at least in one individual
Joiner
Concatenate
GroupBy
Sample
Reference Row Filter
Rule-based Row Filter
Maximum 20 variants

Figure 6.17: Part C6.1 of the CIP. The nodes filter the concatenated list of homozygous variants and heterozygous variant pairs generated in part C4 (Figure 6.15) according to the criteria defined in Section 2.3.6. Overlapping open reading frame is abbreviated as oORF, Loss-of-Function as LoF, Genome Aggregation Database as gnomAD, American College of Medical Genetics and Genomics and the Association for Molecular Pathology as ACMG/AMP, and Combined Annotation Dependent Depletion as CADD.

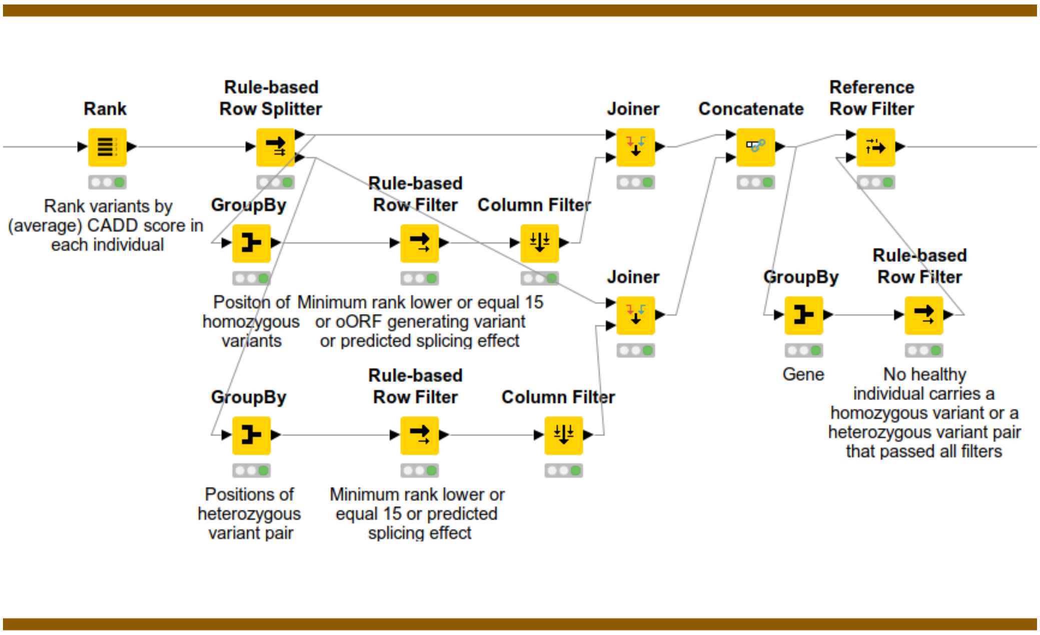

Rank
Rank variants by (average) CADD score in each individual
Rule-based Row Splitter
GroupBy
Positon of homozygous variants
Rule-based Row Filter
Minimum rank lower or equal 15 or oORF generating variant or predicted splicing effect
Column Filter
GroupBy
Positions of heterozygous variant pair
Rule-based Row Filter
Minimum rank lower or equal 15 or predicted splicing effect
Column Filter
Joiner
Joiner
Concatenate
Reference Row Filter
GroupBy
Gene
Rule-based Row Filter
No healthy individual carries a homozygous variant or a heterozygous variant pair that passed all filters

Figure 6.18: Part C6.2 of the CIP. The nodes continue the variant filtering shown in Figure 6.17 (part C6.1). Overlapping open reading frame is abbreviated as oORF and Combined Annotation Dependent Depletion as CADD.

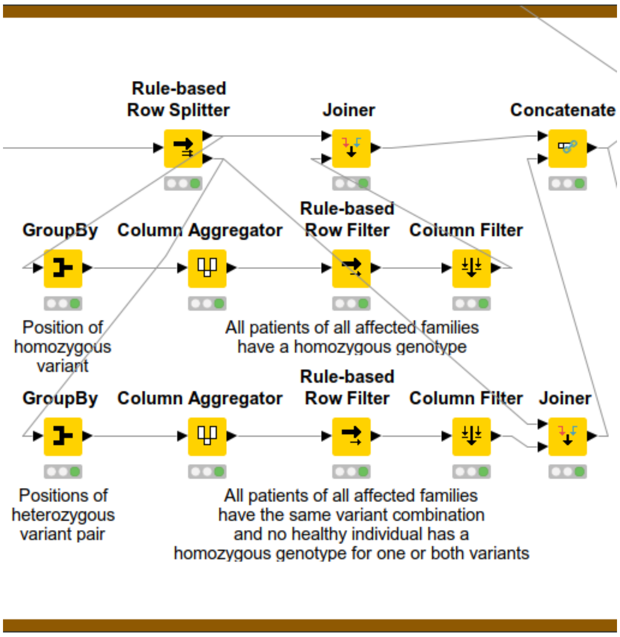

Rule-based Row Splitter
Joiner
Concatenate
GroupBy
Position of homozygous variant
Column Aggregator
Rule-based Row Filter
Column Filter
All patients of all affected families have a homozygous genotype
GroupBy
Positions of heterozygous variant pair
Column Aggregator
Rule-based Row Filter
Column Filter
Joiner
All patients of all affected families have the same variant combination and no healthy individual has a homozygous genotype for one or both variants

Figure 6.19: Part C6.3 of the CIP. The nodes continue the variant prioritization after part C6.1 (Figure 6.17) and C6.2 (Figure 6.18).

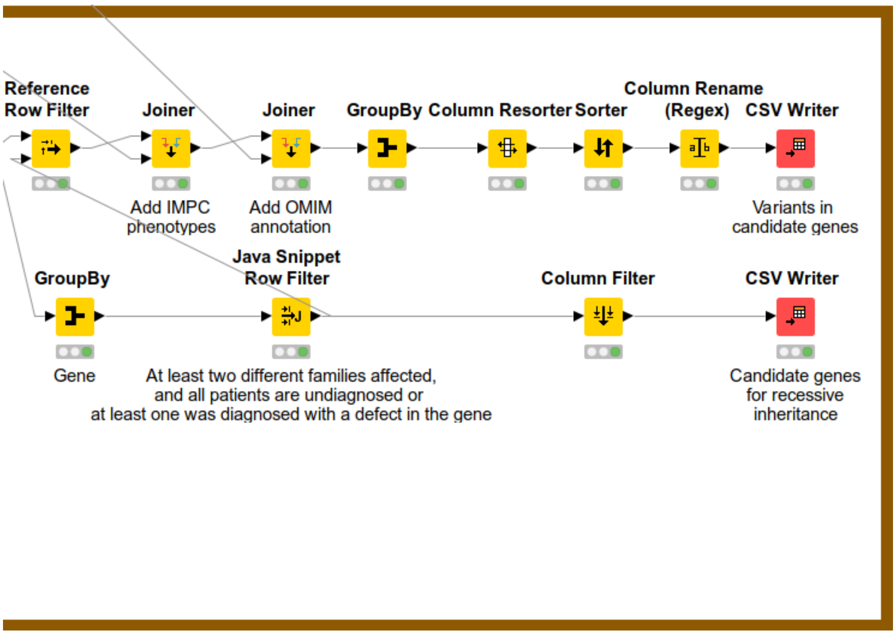


Figure 6.20: Part C6.4 of the CIP. The nodes select candidate genes for recessive inheritance from the set of variants that result from the prioritization steps shown in Figure 6.17 (part C6.1), 6.18 (part C6.2) and 6.19 (part C6.3). Annotations of the International Mouse Phenotyping Consortium (IMPC) and the Online Mendelian Inheritance in Man (OMIM) database imported in part A2 (Figure 6.3) and A3 (Figure 6.4) are added.

D. Overview lists.

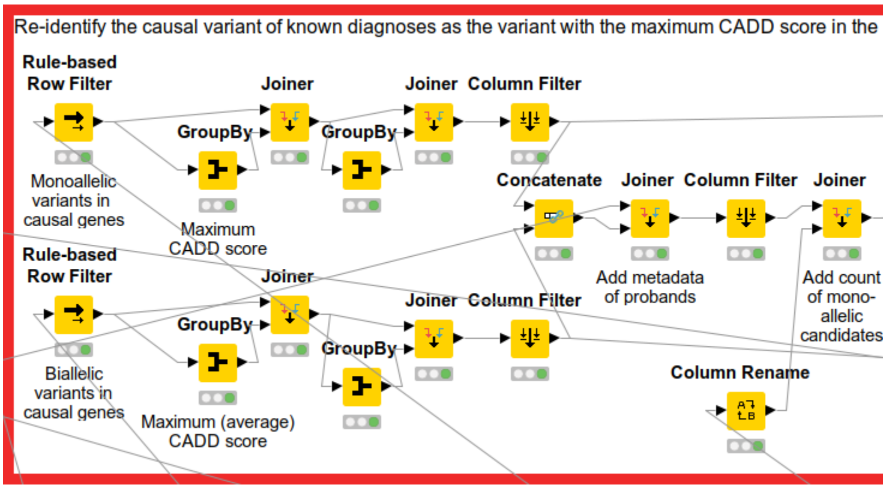

Re-identify the causal variant of known diagnoses as the variant with the maximum CADD score in the
Rule-based Row Filter
Joiner
Joiner
Column Filter
GroupBy
GroupBy
Monoallelic variants in causal genes
Maximum CADD score
Concatenate
Joiner
Column Filter
Joiner
Rule-based Row Filter
Joiner
Add metadata of probands
Add count of mono-allelic candidates
GroupBy
Joiner
Column Filter
GroupBy
Biallelic variants in causal genes
Maximum (average) CADD score
Column Rename

Figure 6.21: Part D1.1 of the CIP. Using the knowledge on known causative genes for a subset of the patients in the cohort, the nodes identify the causal variant by searching for the variant with the maximum Combined Annotation Dependent Depletion (CADD) score in the diagnosed gene using the results of part B2 (Figure 6.7) and C4 (Figure 6.15) as input.

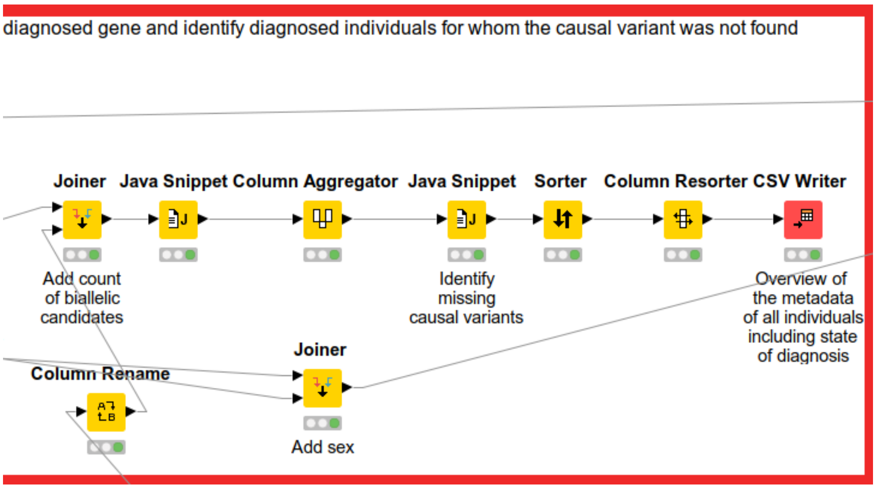

diagnosed gene and identify diagnosed individuals for whom the causal variant was not found
Joiner
Java Snippet
Column Aggregator
Java Snippet
Sorter
Column Resorter
CSV Writer
Add count of biallelic candidates
Identify missing causal variants
Overview of the metadata of all individuals including state of diagnosis
Joiner
Column Rename
Add sex

Figure 6.22: Part D1.2 of the CIP. In continuation of the nodes shown in Figure 6.21 (part D1.1), diagnosed individuals are identified for whom the causal variant is not part of the variant sets generated in Figure 6.7 (part B2) and 6.15 (part C4). These missing causal variants are reported as part of a summary of the metadata of all individuals.

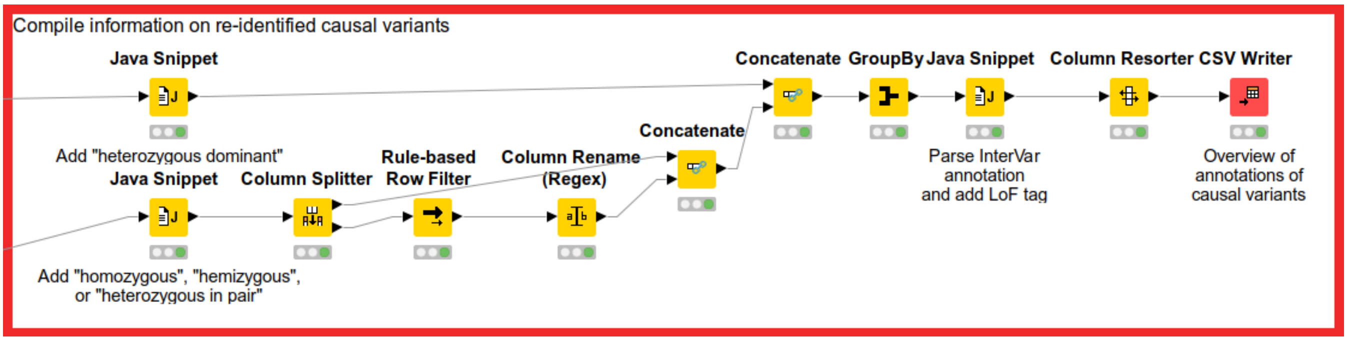

Compile information on re-identified causal variants
Java Snippet
Add "heterozygous dominant"
Java Snippet
Add "homozygous", "hemizygous", or "heterozygous in pair"
Column Splitter
Rule-based Row Filter
Column Rename (Regex)
Concatenate
Concatenate
GroupBy
Java Snippet
Parse InterVar annotation and add LoF tag
Column Resorter
CSV Writer
Overview of annotations of causal variants

Figure 6.23: Part D2 of the CIP. The nodes compile a summary of all causal variants identified in part D1.1 (Figure 6.21).

6.2 List of identified candidate genes

This section lists all candidate genes resulting from the application of the Candidate Identification Pipeline to the WES cohort. Because most candidate genes are still under review to identify promising genes for further research projects on their pathogenicity, only names of genes are shown for which an association with IEI has been published [32], or for which variants have been discussed in Chapter 3. Details on the individual columns and the used abbreviations can be found at the end of the table.

Chr	Gene	Variants in patients	CADD	Effect type	IMPC	OMIM
1	***CD55***	2 in 2	15.9 - 22.1	AR (hm=2)	-	AR
1	***CSF3R***	4 in 4	18.4 - 29.8	AR (hm=5)	lethal	AR
1	***HAX1***	5 in 11	23.1 - 33.0	AR (hm=11)	-	AR
1	***MYSM1***	3 in 3	25.3 - 37.0	AD (ht=2, dn=1)	immune	AR
1	***TNFRSF9***	2 in 2	24.9 - 32.0	AR (hm=2)	lethal	-
1	-	3 in 3	9.5 - 25.7	AD (ht=2, dn=1)	-	AD
1	-	2 in 2	32.0 - 33.0	AR (hm=2)	-	AR
1	-	3 in 3	22.9 - 27.5	AR (cpht=3)	lethal	-
1	-	2 in 2	23.9 - 34.0	AD (ht=1, dn=1)	-	-
1	-	1 in 2	28.1	AD (ht=1, dn=1)	-	-
1	-	4 in 4	25.9 - 29.3	AD (ht=3, dn=1)	-	-
1	-	2 in 2	25.3 - 27.4	AR (hm=1, cpht=1)	-	-
1	-	2 in 2	21.1 - 25.0	AR (hm=1, cpht=1)	-	-
1	-	2 in 2	17.5 - 23.8	AR (hm=1, cpht=1)	-	-
1	-	3 in 3	16.1 - 16.4	AR (hm=1, cpht=2)	-	-
1	-	2 in 2	21.2 - 21.4	AR (hm=2)	-	-
1	-	2 in 2	22.2 - 25.1	AR (hm=2)	-	-
2	***CTLA4***	4 in 4	24.5 - 34.0	AD (ht=3, dn=1)	-	AD
2	*EPCAM*	6 in 6	14.5 - 58.0	AR (hm=6)	-	AR
2	***NBAS***	2 in 2	15.1 - 23.3	AR (hm=2)	lethal	AR
2	***STAT1***	3 in 3	25.0 - 33.0	AR (hm=3)	immune	AD/AR
2	***ZAP70***	5 in 5	23.8 - 34.0	AR (hm=5)	-	AR
2	-	2 in 2	22.7 - 24.2	AR (hm=2)	-	AR
2	-	3 in 3	15.4 - 22.5	AR (hm=3)	-	AR
2	-	2 in 2	23.2 - 24.2	AR (hm=2)	-	AR
2	-	2 in 2	24.1 - 28.2	AR (hm=2)	-	AR
2	-	2 in 2	16.6 - 24.5	AR (hm=2)	lethal	-
2	-	1 in 2	31.0	AR (hm=2)	lethal	-
2	-	2 in 2	27.4 - 31.0	AD (ht=1, dn=1)	-	-
2	-	2 in 2	31.0 - 32.0	AD (ht=1, dn=1)	-	-

2	-	1 in 2	9.8	AR (hm=2)	-	-
3	***JAGN1***	2 in 2	23.8 - 31.0	AR (hm=2)	-	AR
3	-	2 in 2	25.9 - 34.0	AR (hm=2)	-	AD/AR
3	-	3 in 4	29.2 - 34.0	AD (ht=3, dn=1)	lethal	AR
3	-	2 in 2	23.0 - 29.3	AD (ht=1, dn=1)	-	AD
3	-	2 in 2	25.5 - 32.0	AD (ht=1, dn=1)	-	-
3	-	2 in 2	28.0 - 44.0	AD (ht=1, dn=1)	-	-
3	-	3 in 3	0.2 - 24.2	AR (hm=2, cpht=1)	-	-
4	***LRBA***	2 in 2	23.4 - 38.0	AR (hm=2)	-	AR
4	***NFKB1***	2 in 3	25.8 - 32.0	AD (ht=2, dn=1)	immune	AD
4	-	4 in 4	18.7 - 27.8	AD (ht=3, dn=1)	-	AD
4	-	2 in 3	16.4 - 28.0	AR (hm=2, cpht=1)	-	AR
4	-	2 in 2	19.1 - 22.7	AR (hm=2)	-	-
5	***IL7R***	2 in 2	21.5 - 29.9	AR (hm=2)	-	AR
5	***TTC37***	2 in 2	25.3 - 33.0	AR (hm=2)	-	AR
5	-	2 in 2	28.4 - 33.0	AD (ht=1, dn=1)	lethal	AR
5	-	2 in 2	27.6 - 35.0	AD (ht=1, dn=1)	lethal	AR
5	-	3 in 3	16.5 - 24.9	AR (hm=3)	immune	AD
5	-	3 in 3	27.2 - 32.0	AD (ht=2, dn=1)	-	AR
5	-	2 in 2	22.3 - 33.0	AR (hm=1, cpht=1)	-	AR
5	-	2 in 2	24.2 - 31.0	AR (hm=1, cpht=1)	-	-
5	-	2 in 2	24.9 - 25.3	AR (hm=2)	-	-
5	-	2 in 2	18.5 - 25.9	AR (hm=1, cpht=1)	-	-
5	-	2 in 2	19.2 - 28.0	AR (hm=2)	-	-
5	-	2 in 2	22.2 - 27.3	AR (hm=2)	-	-
6	***RIPK1***	4 in 4	22.8 - 27.4	AR (hm=4)	-	AD/AR
6	***SKIV2L***	3 in 3	24.3 - 43.0	AR (hm=3)	lethal	AR
6	-	2 in 2	17.0 - 19.8	AR (hm=2)	lethal	AR
6	-	2 in 2	21.6 - 25.7	AR (hm=2)	-	AD/AR
6	-	2 in 2	22.6 - 25.2	AR (hm=2)	-	AR
6	-	2 in 2	25.1 - 33.0	AD (ht=1, dn=1)	lethal	-
6	-	2 in 2	24.1 - 25.9	AR (hm=2)	lethal	-
6	-	2 in 2	16.5 - 25.8	AR (hm=1, cpht=1)	-	-
6	-	3 in 4	24.2 - 26.1	AR (hm=4)	-	-
6	-	2 in 2	23.5 - 26.9	AR (hm=2)	-	-
7	***ARPC1B***	3 in 3	31.0 - 38.0	AR (hm=3)	immune	AR
7	***NCF1***	1 in 2	35.0	AR (hm=2)	-	AR
7	***SBDS***	2 in 2	28.3 - 44.0	AR (hm=1, cpht=1)	-	AR
7	-	2 in 2	31.0 - 32.0	AR (hm=2)	-	AR
7	-	3 in 4	25.6 - 27.8	AD (ht=3, inherited=1)	-	-

7	-	1 in 2	16.1	AR (hm=2)	-	-
7	-	2 in 2	26.7 - 44.0	AR (hm=2)	-	-
7	-	2 in 2	18.5 - 18.6	AR (hm=2)	-	-
7	-	2 in 2	16.8 - 17.1	AR (hm=2)	-	-
7	-	3 in 2	21.9 - 23.6	AR (hm=3)	-	-
7	-	2 in 2	17.4 - 17.9	AR (hm=2)	-	-
7	-	3 in 3	24.8 - 46.0	AR (hm=2, cpht=1)	-	-
8	***VPS13B***	2 in 2	34.0 - 38.0	AR (hm=2)	-	AR
8	-	2 in 2	23.1 - 27.9	AR (hm=2)	lethal	AR
8	-	2 in 3	28.4 - 34.0	AD (ht=2, inherited=1)	-	AD
8	-	2 in 2	26.2 - 27.5	AR (hm=2)	-	AR
8	-	1 in 2	32.0	AD (ht=1, dn=1)	-	-
8	-	2 in 2	24.4 - 25.9	AR (hm=2)	-	-
8	-	2 in 2	24.2 - 26.4	AR (hm=2)	-	-
8	-	2 in 2	15.1 - 23.5	AR (hm=2)	-	-
9	-	1 in 2	27.5	AR (hm=2)	lethal	AR
9	-	1 in 2	33.0	AD (ht=1, dn=1)	-	-
9	-	1 in 2	25.3	AD (ht=1, dn=1)	-	-
9	-	2 in 2	22.3 - 27.0	AR (hm=1, cpht=1)	-	-
9	-	2 in 2	22.1 - 23.6	AR (hm=1, cpht=1)	-	-
10	-	4 in 4	26.1 - 33.0	AD (ht=3, dn=1)	lethal	AR
10	-	2 in 2	25.9 - 28.6	AD (ht=1, dn=1)	-	AD/AR
10	-	2 in 2	19.3 - 22.7	AR (hm=2)	-	AD
10	-	2 in 2	15.1 - 28.5	AD (ht=1, dn=1)	lethal	-
10	-	2 in 2	22.1 - 23.4	AR (hm=2)	-	-
11	*C_GENE_1*	2 in 8	23.3 - 25.0	AD (ht=7, dn=1)	-	AD/AR
11	***IL10RA***	2 in 2	24.3 - 26.1	AR (hm=2)	immune	AR
11	***RAG1***	3 in 3	24.7 - 28.9	AR (hm=3)	-	AR
11	***RAG2***	2 in 2	27.4 - 29.3	AR (hm=2)	-	AR
11	-	2 in 2	16.6 - 24.7	AD (ht=1, dn=1)	-	AD
11	-	1 in 3	33.0	AD (ht=2, dn=1)	-	-
11	-	5 in 4	22.5 - 32.0	AD (ht=4, dn=1)	-	-
11	-	2 in 2	22.5 - 22.9	AR (hm=2)	-	-
11	-	3 in 3	15.4 - 23.1	AR (hm=2, cpht=1)	-	-
11	-	2 in 2	20.9 - 23.2	AR (hm=2)	-	-
11	-	2 in 2	22.3 - 23.5	AR (hm=2)	-	-
12	***KRAS***	1 in 2	26.1	AD (ht=1, dn=1)	-	AD
12	***MVK***	2 in 2	24.6 - 31.0	AR (hm=2)	lethal	AD/AR
12	-	2 in 4	23.2 - 24.4	AD (ht=2, dn=2)	-	-

12	-	2 in 2	24.0 - 37.0	AD (ht=1, dn=1)	-	-
12	-	2 in 2	16.1 - 22.4	AR (hm=1, cpht=1)	-	-
12	-	2 in 2	15.8 - 22.3	AR (hm=2)	-	-
12	-	2 in 2	17.3 - 27.4	AR (hm=2)	-	-
12	-	2 in 2	17.8 - 23.6	AR (hm=2)	-	-
12	-	2 in 2	28.8 - 33.0	AR (hm=2)	-	-
12	-	2 in 2	22.4 - 29.9	AR (hm=1, cpht=1)	-	-
13	***LIG4***	1 in 2	26.3	AR (hm=2)	-	AR
13	-	2 in 2	19.2 - 32.0	AR (hm=2)	immune	-
13	-	1 in 2	33.0	AD (ht=1, dn=1)	-	-
14	***SRP54***	4 in 10	22.5 - 31.0	AD (ht=8, dn=2)	-	AD
14	-	2 in 2	23.8 - 25.1	AR (hm=1, cpht=1)	immune	-
14	-	2 in 2	25.0 - 31.0	AR (hm=2)	-	-
15	***RAB27A***	2 in 2	31.0 - 33.0	AR (hm=2)	-	AR
15	***RASGRP1***	3 in 2	26.9 - 28.1	AR (hm=3)	-	AR
15	-	2 in 2	24.1 - 29.5	AR (hm=2)	-	AR
15	-	3 in 3	18.6 - 28.9	AR (hm=3)	lethal	-
15	-	1 in 2	27.4	AD (ht=1, dn=1)	-	-
15	-	4 in 3	26.5 - 35.0	AD (ht=4, dn=3)	-	-
15	-	2 in 2	23.6 - 25.9	AR (hm=2)	-	-
15	-	2 in 2	16.2 - 28.4	AR (hm=2)	-	-
15	-	1 in 2	21.6	AR (hm=2)	-	-
16	***RLTPR***	4 in 4	23.5 - 33.0	AR (hm=4)	-	-
16	-	2 in 2	25.6 - 32.0	AD (ht=1, dn=1)	-	-
16	-	2 in 2	19.1 - 24.9	AR (hm=2)	-	-
16	-	2 in 2	20.9 - 22.5	AR (hm=2)	-	-
16	-	2 in 2	17.9 - 24.1	AR (hm=2)	-	-
17	***G6PC3***	7 in 8	18.3 - 33.0	AR (hm=7, cpht=1)	-	AR
17	***STAT3***	4 in 4	7.9 - 32.0	AD (ht=3, dn=1)	-	AD
17	-	2 in 2	25.2 - 30.0	AD (ht=1, dn=1)	lethal	-
17	-	2 in 2	20.4 - 25.4	AR (hm=1, cpht=1)	lethal	-
17	-	2 in 2	16.5 - 25.7	AR (hm=1, cpht=1)	immune	-
17	-	2 in 2	13.8 - 28.6	AR (hm=2)	-	-
17	-	2 in 2	24.1 - 32.0	AR (hm=2)	-	-
17	-	2 in 2	23.0 - 38.0	AR (hm=2)	-	-
17	-	2 in 2	17.6 - 22.1	AR (hm=2)	-	-
18	-	2 in 2	23.2 - 32.0	AD (ht=1, dn=2)	-	AD
18	-	3 in 3	15.6 - 27.0	AR (hm=2, cpht=1)	-	AR
18	-	2 in 2	22.2 - 29.1	AR (hm=2)	-	-
19	***ELANE***	2 in 2	11.5 - 26.2	AD (dn=2)	-	AD

19	***FCHO1***	2 in 3	26.1 - 33.0	AR (hm=3)	-	-
19	***IL12RB1***	3 in 3	23.4 - 32.0	AR (hm=3)	-	AR
19	***JAK3***	2 in 2	29.2 - 29.7	AR (hm=2)	-	AR
19	***TGFB1***	2 in 2	24.5 - 28.9	AR (hm=1, cpht=1)	-	AD/AR
19	***TYK2***	2 in 2	21.7 - 25.0	AR (hm=1, cpht=1)	-	AR
19	-	3 in 3	32.0 - 39.0	AD (ht=2, dn=1)	lethal	AD
19	-	2 in 2	23.2 - 29.3	AR (hm=1, cpht=1)	-	AR
19	-	2 in 2	17.0 - 23.3	AR (hm=2)	lethal	-
19	-	2 in 3	25.7 - 28.8	AD (ht=2, dn=1)	-	-
19	-	3 in 4	25.2 - 27.7	AD (ht=2, dn=2)	-	-
19	-	2 in 2	15.0 - 17.1	AR (hm=2)	-	-
19	-	2 in 2	16.0 - 25.8	AR (hm=2)	-	-
19	-	2 in 3	19.0 - 26.6	AR (hm=3)	-	-
20	***DNMT3B***	4 in 4	29.2 - 32.0	AR (hm=3, cpht=1)	lethal	AR
20	***ZNF341***	2 in 2	27.8 - 29.1	AR (hm=2)	-	AR
20	-	1 in 2	23.5	AD (ht=1, dn=1)	-	AD
20	-	3 in 3	21.6 - 32.0	AD (ht=2, dn=1)	-	AR
20	-	2 in 2	19.8 - 32.0	AR (hm=1, cpht=1)	-	AR
20	-	3 in 3	19.4 - 29.9	AR (hm=2, cpht=1)	lethal	-
21	***ICOSLG***	2 in 2	22.3 - 36.0	AR (hm=2)	-	-
21	***IL10RB***	6 in 8	19.1 - 33.0	AR (hm=8)	immune	AR
21	***ITGB2***	3 in 3	24.8 - 33.0	AR (hm=3)	-	AR
21	-	2 in 2	15.9 - 22.2	AR (hm=2)	-	AD
21	-	1 in 2	15.3	AR (hm=2)	-	AD
22	***CECR1***	3 in 3	21.9 - 25.2	AR (hm=3)	-	-
22	*SLC5A1*	5 in 5	22.7 - 40.0	AR (hm=5)	-	AR
22	***USP18***	1 in 2	19.5	AR (hm=2)	-	AR
22	-	1 in 2	24.7	AD (ht=1, dn=1)	-	-
X	***BTK***	4 in 7	26.2 - 32.0	XLR (hemi=7)	immune	XLR
X	***CD40LG***	3 in 3	22.9 - 26.3	XLR (hemi=3)	-	XLR
X	*C_GENE_2*	6 in 6	18.9 - 28.3	XLR (hemi=6)	immune	-
X	***DKC1***	2 in 2	23.4 - 25.7	XLR (hemi=2)	-	XLR
X	***G6PD***	3 in 3	18.8 - 24.5	XLR (hemi=3)	-	XLD
X	***MSN***	3 in 3	22.6 - 31.0	XLR (hemi=3)	-	XLR
X	***WAS***	5 in 5	16.0 - 32.0	XLR (hemi=5)	-	XLR
X	***XIAP***	3 in 3	24.5 - 33.0	XLR (hemi=3)	-	XLR
X	-	2 in 2	17.3 - 19.4	XLR (hemi=2)	lethal	XLD
X	-	3 in 3	15.6 - 33.0	XLD (ht=1, dn=2)	-	XLR
X	-	2 in 2	22.8 - 25.9	XLR (hemi=2)	-	XLR
X	-	2 in 2	15.8 - 24.8	XLR (hemi=2)	-	XLR

X	-	1 in 2	27.4	XLR (hemi=2)	-	XLD
X	-	1 in 2	17.3	XLR (hemi=2)	-	XLR
X	-	2 in 2	16.9 - 19.8	XLR (hemi=2)	-	
X	-	2 in 2	23.9 - 25.8	XLR (hemi=2)	-	XLR
X	-	2 in 2	22.1 - 22.9	XLR (hemi=2)	-	XLR
X	-	2 in 2	15.2 - 23.2	XLR (hemi=2)	-	XLR
X	-	2 in 2	20.9 - 22.2	XLR (hemi=2)	-	-
X	-	4 in 4	18.9 - 28.4	XLR (hemi=4)	-	-
X	-	2 in 2	22.6 - 25.8	XLR (hemi=2)	-	-
X	-	2 in 2	20.2 - 22.3	XLR (hemi=2)	-	-
X	-	2 in 2	16.8 - 28.1	XLR (hemi=2)	-	-
X	-	2 in 2	16.9 - 22.5	XLR (hemi=2)	-	-
X	-	2 in 2	16.8 - 20.2	XLR (hemi=2)	-	-
X	-	2 in 2	24.3 - 24.4	XLR (hemi=2)	-	-
X	-	2 in 2	23.1 - 33.0	XLR (hemi=2)	-	-
Y	-	2 in 2	23.1 - 24.4	YL (2)	-	-

Table 6.1: List of identified candidate genes. For each chromosome specified in the column "Chr", named genes are listed in lexicographical order, while unnamed genes are ordered by chromosomal position. Names are in bold if genes are associated with inborn errors of immunity [32]. The column "Variants in patients" indicates the number of variants or variant pairs that were identified and how many patients are affected by these. The number of variants can be smaller than the number of patients if a variant or variant pair was found in multiple patients. If a patient carries multiple variants in the same gene, the number of variants can be larger than the number of patients. The range of the Combined Annotation Dependent Depletion scores of the variants is given in the column "CADD". The column "Effect type" specifies whether the gene was identified as a candidate when searching for dominant effects (autosomal dominant (AD) or X-linked dominant (XLD)) or recessive effects (autosomal recessive (AR) or X-linked recessive (XLR)) including Y-linked variants (YL). Counts of the genotypes carried by the affected patients are given in the same column in brackets where hm is homozygous, ht is heterozygous, dn is *de novo*, cpht is compound heterozygous and hemi is hemizygous. The column "IMPC" informs whether preweaning lethality (lethal) or any immune system phenotypes (immune) were significantly associated with knockouts of orthologous mouse genes as reported by the International Mouse Phenotyping Consortium based on data release 11.0 of February 2020. The column "OMIM" indicates the mode of inheritance of gene-disease phenotype relationships annotated in Online Mendelian Inheritance in Man by June 2020.

List of Abbreviations

ACMG	American College of Medical Genetics and Genomics
AMP	Association for Molecular Pathology
BAM	Binary Alignment/Map
BQSR	Base Quality Score Recalibration
BWA	Burrow-Wheeler Aligner
CADD	Combined Annotation Dependent Depletion
cDNA	complementary DNA
CIP	Candidate Identification Pipeline
DC	dyskeratosis congenita
DIAR5	diarrhea-5 with congenital tufting enteropathy
DNA	deoxyribonucleic acid
DP	depth of coverage
EEUR	Eastern Europe
ER	endoplasmatic reticulum
ExAC	Exome Aggregation Consortium
GATK	Genome Analysis Toolkit
G-CSF	granulocyte colony-stimulating factor
GGM	glucose-galactose malabsorption
GME	Greater Middle East
gnomAD	Genome Aggregation Database
GRCh37	Genome Reference Consortium Human Build 37
GQ	genotype quality
GTP	guanosine-5'-triphosphate
gVCF	genome variant call format
HHS	Hyeraal-Hreidarsson syondrome
HPO	Human Phenotype Ontology
HTE	High-Throughput Executor
IBD	inflammatory bowel disease
IBS0	zero identity-by-state

IEI	inborn errors of immunity
IMPC	International Mouse Phenotyping Consortium
InDel	insertion or deletion
IPEX	Immunodysregulation, Polyendocrinopathy, and Enteropathy, X-linked
iPS	induced pluripotent stem cell
ISC	Indian Subcontinent
IUIS	International Union of Immunological Societies
KNIME	Konstanz Information Miner
LD	linkage disequilibrium
LoF	Loss-of-Function
LOFTEE	Loss-Of-Function Transcript Effect Estimator
NAFR	Northern Africa
NGS	next-generation sequencing
NMD	nonsense-mediated mRNA decay
mRNA	messenger RNA
OMIM	Online Mendelian Inheritance in Man
PPI	protein-protein interaction
RNA	ribonucleic acid
SAM	Sequence Alignment/Map
SDS	Shwachman-Diamond syndrome
SEAS	South East Asia
SCN	severe congenital neutropenia
SNV	single nucleotide variant
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SRP	signal recognition particle
Ti/Tv	transition/transversion
uORF	upstream open reading frame
UTR	untranslated region
WES	whole-exome sequencing
WEUR	Western Europe
WGS	whole-genome sequencing
VCF	variant call format
VEOIBD	very early onset IBD
VEP	Variant Effect Predictor
VQSR	Variant Quality Score Recalibration

List of Figures

List of Tables

Bibliography

[1] S. Nguengang Wakap, D. M. Lambert, A. Olry, C. Rodwell, C. Gueydan, V. Lanneau, D. Murphy, Y. Le Cam, and A. Rath, "Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database," Eur. J. Hum. Genet., vol. 28, pp. 165–173, Feb. 2020.

[2] World Health Organization, "Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016," 2018. Available: https://www.who.int/healthinfo/global_burden_disease/estimates/en/.

[3] E. A. Worthey, A. N. Mayer, G. D. Syverson, D. Helbling, B. B. Bonacci, B. Decker, J. M. Serpe, T. Dasu, M. R. Tschannen, R. L. Veith, M. J. Basehore, U. Broeckel, A. Tomita-Mitchell, M. J. Arca, J. T. Casper, D. A. Margolis, D. P. Bick, M. J. Hessner, J. M. Routes, J. W. Verbsky, H. J. Jacob, and D. P. Dimmock, "Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease," Genet. Med., vol. 13, pp. 255–62, Mar. 2011.

[4] M. M. Clark, Z. Stark, L. Farnaes, T. Y. Tan, S. M. White, D. Dimmock, and S. F. Kingsmore, "Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases," npj Genomic Med., vol. 3, p. 16, July 2018.

[5] C. F. Wright, D. R. FitzPatrick, and H. V. Firth, "Paediatric genomics: diagnosing rare disease in children," Nat. Rev. Genet., vol. 19, pp. 253–268, May 2018.

[6] G. Mendel, "Versuche über Pflanzen-Hybriden," Verhandlungen des Naturforschenden Vereins zu Brünn, vol. 4, pp. 3–47, 1866.

[7] F. H. C. Crick, "On protein synthesis," Symp. Soc. Exp. Biol., vol. 12, pp. 138–63, Sept. 1958.

[8] F. Crick, "Central dogma of molecular biology," Nature, vol. 227, pp. 561–3, Aug. 1970.

[9] K. H. Miga, S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, V. A. Schneider, T. Potapova,

J. Wood, W. Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G. G. Bouffard, A. M. Chang, N. F. Hansen, A. B. Wilfert, F. Thibaud-Nissen, A. D. Schmitt, J.-M. Belton, S. Selvaraj, M. Y. Dennis, D. C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N. J. Loman, N. Holmes, M. Loose, U. Surti, R. A. Risques, T. A. Graves Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J. C. Mullikin, P. A. Pevzner, J. L. Gerton, B. A. Sullivan, E. E. Eichler, and A. M. Phillippy, "Telomere-to-telomere assembly of a complete human X chromosome," Nature, vol. 585, pp. 79–84, Sept. 2020.

[10] R. M. Sherman and S. L. Salzberg, "Pan-genomics in the human genome era," Nat. Rev. Genet., vol. 21, pp. 243–254, Apr. 2020.

[11] J. G. R. Cardoso, M. R. Andersen, M. J. Herrgård, and N. Sonnenschein, "Analysis of genetic variation and potential applications in genome-scale metabolic modeling," Front. Bioeng. Biotechnol., vol. 3, p. 13, Feb. 2015.

[12] P. Balachandran and C. R. Beck, "Structural variant identification and characterization," Chromosome Res., vol. 28, pp. 31–47, Mar. 2020.

[13] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, and Exome Aggregation Consortium, "Analysis of protein-coding genetic variation in 60,706 humans," Nature, vol. 536, pp. 285–91, Aug. 2016.

[14] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton,

A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, 1000 Genomes Project Consortium, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, and J. O. Korbel, "An integrated map of structural variation in 2,504 human genomes," Nature, vol. 526, pp. 75–81, Oct. 2015.

[15] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," Nat. Rev. Genet., vol. 7, pp. 85–97, Feb. 2006.

[16] A. J. Holland and D. W. Cleveland, "Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements," Nat. Med., vol. 18, pp. 1630–8, Nov. 2012.

[17] 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis, "A global reference for human genetic variation," Nature, vol. 526, pp. 68–74, Oct. 2015.

[18] ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," Nature, vol. 489, pp. 57–74, Sept. 2012.

[19] M. A. Garcia-Blanco, A. P. Baraniak, and E. L. Lasda, "Alternative splicing in disease and therapy," Nat. Biotechnol., vol. 22, pp. 535–46, May 2004.

[20] S. E. Calvo, D. J. Pagliarini, and V. K. Mootha, "Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans," Proc. Natl. Acad. Sci., vol. 106, pp. 7507–7512, May 2009.

[21] N. Whiffin, K. J. Karczewski, X. Zhang, S. Chothani, M. J. Smith, D. G. Evans, A. M. Roberts, N. M. Quaife, S. Schafer, O. Rackham, J. Alföldi, A. H. O'Donnell-Luria, L. C. Francioli, Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, S. A. Cook, P. J. R. Barton, D. G. MacArthur, and J. S. Ware, "Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals," Nat. Commun., vol. 11, p. 2523, May 2020.

[22] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, "OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders," Nucleic Acids Res., vol. 43, pp. D789–98, Jan. 2015.

[23] L. Echevarria, K. Benistan, A. Toussaint, O. Dubourg, A. A. Hagege, D. Eladari, F. Jabbour, C. Beldjord, P. De Mazancourt, and D. P. Germain, "X-chromosome

inactivation in female patients with Fabry disease," Clin. Genet., vol. 89, pp. 44–54, Jan. 2016.

[24] F. A. Kondrashov and E. V. Koonin, "A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications," Trends Genet., vol. 20, pp. 287–90, July 2004.

[25] H. Kacser and J. A. Burns, "The molecular basis of dominance," Genetics, vol. 97, pp. 639–66, Sept. 1980.

[26] H. S. Kuehn, B. Boisson, C. Cunningham-Rundles, J. Reichenbach, A. Stray-Pedersen, E. W. Gelfand, P. Maffucci, K. R. Pierce, J. K. Abbott, K. V. Voelkerding, S. T. South, N. H. Augustine, J. S. Bush, W. K. Dolen, B. B. Wray, Y. Itan, A. Cobat, H. S. Sorte, S. Ganesan, S. Prader, T. B. Martins, M. G. Lawrence, J. S. Orange, K. R. Calvo, J. E. Niemela, J.-L. Casanova, T. A. Fleisher, H. R. Hill, A. Kumánovics, M. E. Conley, and S. D. Rosenzweig, "Loss of B Cells in Patients with Heterozygous Mutations in IKAROS," N. Engl. J. Med., vol. 374, pp. 1032–1043, Mar. 2016.

[27] D. Boutboul, H. S. Kuehn, Z. Van de Wyngaert, J. E. Niemela, I. Callebaut, J. Stoddard, C. Lenoir, V. Barlogis, C. Farnarier, F. Vely, N. Yoshida, S. Kojima, H. Kanegane, A. Hoshino, F. Hauck, L. Lhermitte, V. Asnafi, P. Roehrs, S. Chen, J. W. Verbsky, K. R. Calvo, A. Husami, K. Zhang, J. Roberts, D. Amrol, J. Sleaseman, A. P. Hsu, S. M. Holland, R. Marsh, A. Fischer, T. A. Fleisher, C. Picard, S. Latour, and S. D. Rosenzweig, "Dominant-negative IKZF1 mutations cause a T, B, and myeloid cell combined immunodeficiency," J. Clin. Invest., vol. 128, pp. 3071–3087, July 2018.

[28] K. L. Del Bel, R. J. Ragotte, A. Saferali, S. Lee, S. M. Vercauteren, S. A. Mostafavi, R. A. Schreiber, J. S. Prendiville, M. S. Phang, J. Halparin, N. Au, J. M. Dean, J. J. Priatel, E. Jewels, A. K. Junker, P. C. Rogers, M. Seear, M. L. McKinnon, and S. E. Turvey, "JAK1 gain-of-function causes an autosomal dominant immune dysregulatory and hypereosinophilic syndrome," J. Allergy Clin. Immunol., vol. 139, pp. 2016–2020.e5, Jan. 2017.

[29] D. Raviv, A. A. Dror, and K. B. Avraham, "Hearing loss: A common disorder caused by many rare alleles," Ann. N. Y. Acad. Sci., vol. 1214, pp. 168–179, Dec. 2010.

[30] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An Expanded View of Complex Traits: From Polygenic to Omnigenic," Cell, vol. 169, pp. 1177–1186, June 2017.

[31] S. Sivakumaran, F. Agakov, E. Theodoratou, J. G. Prendergast, L. Zgaga, T. Manolio, I. Rudan, P. McKeigue, J. F. Wilson, and H. Campbell, "Abundant Pleiotropy in Human Complex Diseases and Traits," Am. J. Hum. Genet., vol. 89, pp. 607–618, Nov. 2011.

[32] S. G. Tangye, W. Al-Herz, A. Bousfiha, T. Chatila, C. Cunningham-Rundles, A. Etzioni, J. L. Franco, S. M. Holland, C. Klein, T. Morio, H. D. Ochs, E. Oksenhendler, C. Picard, J. Puck, T. R. Torgerson, J.-L. Casanova, and K. E. Sullivan, "Human Inborn Errors of Immunity: 2019 Update on the Classification from the International Union of Immunological Societies Expert Committee," J. Clin. Immunol., vol. 40, pp. 24–64, Jan. 2020.

[33] D. N. Cooper, M. Krawczak, C. Polychronakos, C. Tyler-Smith, and H. Kehrer-Sawatzki, "Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease," Hum. Genet., vol. 132, pp. 1077–1130, Oct. 2013.

[34] S. J. Chapman and A. V. S. Hill, "Human genetic susceptibility to infectious disease," Nat. Rev. Genet., vol. 13, pp. 175–88, Feb. 2012.

[35] R. Plomin, C. M. A. Haworth, and O. S. P. Davis, "Common disorders are quantitative traits," Nat. Rev. Genet., vol. 10, pp. 872–8, Dec. 2009.

[36] International HapMap Consortium, "A haplotype map of the human genome," Nature, vol. 437, pp. 1299–320, Oct. 2005.

[37] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," Nat. Rev. Genet., vol. 6, pp. 95–108, Feb. 2005.

[38] W. Y. S. Wang, B. J. Barratt, D. G. Clayton, and J. A. Todd, "Genome-wide association studies: theoretical and practical concerns," Nat. Rev. Genet., vol. 6, pp. 109–18, Feb. 2005.

[39] S. Lee, G. R. Abecasis, M. Boehnke, and X. Lin, "Rare-variant association analysis: study designs and statistical tests," Am. J. Hum. Genet., vol. 95, pp. 5–23, July 2014.

[40] D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M.-M. Suner, T. Hunt, I. H. A. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, and C. Tyler-Smith, "A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes," Science, vol. 335, pp. 823–828, Feb. 2012.

[41] "A milestone in human genetics highlights diversity gaps," Nature, vol. 581, pp. 356–356, May 2020.

[42] T. Stoeger, M. Gerlach, R. I. Morimoto, and L. A. Nunes Amaral, "Large-scale investigation of the reasons why potentially important genes are ignored," PLoS Biol., vol. 16, p. e2006643, Sept. 2018.

[43] "The portal for rare diseases and orphan drugs." `https://www.orpha.net`. Accessed: 2020-05-08.

[44] N. L. Chamberlain, E. D. Driver, and R. L. Miesfeld, "The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function," Nucleic Acids Res., vol. 22, pp. 3181–3186, Aug. 1994.

[45] A. D. Sperfeld, J. Karitzky, D. Brummer, H. Schreiber, J. Häussler, A. C. Ludolph, and C. O. Hanemann, "X-linked Bulbospinal Neuronopathy," Arch. Neurol., vol. 59, p. 1921, Dec. 2002.

[46] F. Hauck and C. Klein, "Pathogenic mechanisms and clinical implications of congenital neutropenia syndromes," Curr. Opin. Allergy Clin. Immunol., vol. 13, pp. 596–606, Dec. 2013.

[47] J. Skokowa, D. C. Dale, I. P. Touw, C. Zeidler, and K. Welte, "Severe congenital neutropenias," Nat. Rev. Dis. Prim., vol. 3, p. 17032, Dec. 2017.

[48] J. Spoor, H. Farajifard, and N. Rezaei, "Congenital neutropenia and primary immunodeficiency diseases," Crit. Rev. Oncol. Hematol., vol. 133, pp. 149–162, Jan. 2019.

[49] C. Klein, "Congenital neutropenia," in Stiehm's Immune Deficiencies, pp. 797–812, Academic Press, 2020.

[50] D. C. Dale, R. E. Person, A. A. Bolyard, A. G. Aprikyan, C. Bos, M. A. Bonilla, L. A. Boxer, G. Kannourakis, C. Zeidler, K. Welte, K. F. Benson, and M. Horwitz, "Mutations in the gene encoding neutrophil elastase in congenital and cyclic neutropenia," Blood, vol. 96, pp. 2317–2322, Oct. 2000.

[51] C. Klein, M. Grudzien, G. Appaswamy, M. Germeshausen, I. Sandrock, A. A. Schäffer, C. Rathinam, K. Boztug, B. Schwinzer, N. Rezaei, G. Bohn, M. Melin, G. Carlsson, B. Fadeel, N. Dahl, J. Palmblad, J.-I. Henter, C. Zeidler, B. Grimbacher, and K. Welte, "HAX1 deficiency causes autosomal recessive severe congenital neutropenia (Kostmann disease)," Nat. Genet., vol. 39, pp. 86–92, Jan. 2007.

[52] J. Ouahed, E. Spencer, D. Kotlarz, D. S. Shouval, M. Kowalik, K. Peng, M. Field, L. Grushkin-Lerner, S.-Y. Pai, A. Bousvaros, J. Cho, C. Argmann, E. Schadt, D. P. B. Mcgovern, M. Mokry, E. Nieuwenhuis, H. Clevers, F. Powrie, H. Uhlig, C. Klein, A. Muise, M. Dubinsky, and S. B. Snapper, "Very Early Onset Inflammatory Bowel Disease: A Clinical Approach With a Focus on the Role of Genetics and Underlying Immune Deficiencies," Inflamm. Bowel Dis., vol. 26, pp. 820–842, May 2020.

[53] J. Pazmandi, A. Kalinichenko, R. C. Ardy, and K. Boztug, "Early-onset inflammatory bowel disease as a model disease to identify key regulators of immune homeostasis mechanisms," Immunol. Rev., vol. 287, pp. 162–185, Jan. 2019.

[54] J. R. Kelsen, M. A. Conrad, N. Dawany, T. Patel, R. Shraim, A. Merz, K. Maurer, K. E. Sullivan, and M. Devoto, "The Unique Disease Course of Children with Very Early onset-Inflammatory Bowel Disease," Inflamm. Bowel Dis., vol. 26, pp. 909–918, May 2020.

[55] E. I. Benchimol, C. N. Bernstein, A. Bitton, M. W. Carroll, H. Singh, A. R. Otley, M. Vutcovici, W. El-Matary, G. C. Nguyen, A. M. Griffiths, D. R. Mack, K. Jacobson, N. Mojaverian, D. Tanyingoh, Y. Cui, Z. J. Nugent, J. Coulombe, L. E. Targownik, J. L. Jones, D. Leddin, S. K. Murthy, and G. G. Kaplan, "Trends in Epidemiology of Pediatric Inflammatory Bowel Disease in Canada: Distributed Network Analysis of Multiple Population-Based Provincial Health Administrative Databases," Am. J. Gastroenterol., vol. 112, pp. 1120–1134, July 2017.

[56] Y. Ye, S. Manne, W. R. Treem, and D. Bennett, "Prevalence of Inflammatory Bowel Disease in Pediatric and Adult Populations: Recent Estimates From Large National Databases in the United States, 2007-2016," Inflamm. Bowel Dis., vol. 26, pp. 619–625, Mar. 2020.

[57] F. Charbit-Henrion, M. Parlato, S. Hanein, R. Duclaux-Loras, J. Nowak, B. Begue, S. Rakotobe, J. Bruneau, C. Fourrage, O. Alibeu, F. Rieux-Laucat, E. Lévy, M.-C. Stolzenberg, F. Mazerolles, S. Latour, C. Lenoir, A. Fischer, C. Picard, M. Aloi, J. A. Dias, M. B. Hariz, A. Bourrier, C. Breuer, A. Breton, J. Bronski, S. Buderus, M. Cananzi, S. Coopman, C. Crémilleux, A. Dabadie, C. Dumant-Forest, O. E. Gurkan, A. Fabre, A. Fischer, M. G. Diaz, Y. Gonzalez-Lama, O. Goulet, G. Guariso, N. Gurcan, M. Homan, J.-P. Hugot, E. Jeziorski, E. Karanika, A. Lachaux, P. Lewindon, R. Lima, F. Magro, J. Major, G. Malamut, E. Mas, I. Mattyus, L. M. Mearin, J. Melek, V. M. Navas-Lopez, A. Paerregaard, C. Pelatan, B. Pigneur, I. P. Pais, J. Rebeuh, C. Romano, N. Siala, C. Strisciuglio, M. Tempia-Caliera, P. Tounian, D. Turner, V. Urbonas, S. Willot, F. M. Ruemmele, and N. Cerf-Bensussan, "Diagnostic Yield of Next-generation Sequencing in Very Early-onset Inflammatory Bowel Diseases: A Multicentre Study," J. Crohns. Colitis, vol. 12, pp. 1104–1112, Aug. 2018.

[58] E. Crowley, N. Warner, J. Pan, S. Khalouei, A. Elkadri, K. Fiedler, J. Foong, A. L. Turinsky, D. Bronte-Tinkew, S. Zhang, J. Hu, D. Tian, D. Li, J. Horowitz, I. Siddiqui, J. Upton, C. M. Roifman, P. C. Church, D. A. Wall, A. K. Ramani, D. Kotlarz, C. Klein, H. Uhlig, S. B. Snapper, C. Gonzaga-Jauregui, A. D. Paterson, D. P. B. McGovern, M. Brudno, T. D. Walters, A. M. Griffiths, and A. M. Muise, "Prevalence and Clinical Features of Inflammatory Bowel Diseases Associated With Monogenic

Variants, Identified by Whole-Exome Sequencing in 1000 Children at a Single Center," Gastroenterology, vol. 158, pp. 2208–2220, June 2020.

[59] M. Ditschkowski, H. Einsele, R. Schwerdtfeger, D. Bunjes, R. Trenschel, D. W. Beelen, and A. H. Elmaagacli, "Improvement of inflammatory bowel disease after allogeneic stem-cell transplantation," Transplantation, vol. 75, pp. 1745–7, May 2003.

[60] D. A. Marshall, K. V. MacDonald, S. Heidenreich, T. Hartley, F. P. Bernier, M. K. Gillespie, B. McInnes, A. M. Innes, C. M. Armour, and K. M. Boycott, "The value of diagnostic testing for parents of children with rare genetic diseases," Genet. Med., vol. 21, pp. 2798–2806, June 2019.

[61] J.-L. Casanova, M. E. Conley, S. J. Seligman, L. Abel, and L. D. Notarangelo, "Guidelines for genetic studies in single patients: lessons from primary immunodeficiencies," J. Exp. Med., vol. 211, pp. 2137–49, Oct. 2014.

[62] "Research for Rare diseases and Personalised Medicine." `https://www.ten-for-rare.com`. Accessed: 2020-02-06.

[63] "Care-for-Rare Foundation." `https://www.care-for-rare.org`. Accessed: 2020-02-06.

[64] B. M. Henn, L. L. Cavalli-Sforza, and M. W. Feldman, "The great human expansion," Proc. Natl. Acad. Sci. U. S. A., vol. 109, pp. 17758–64, Oct. 2012.

[65] K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, and D. G. MacArthur, "The mutational constraint spectrum quantified from variation in 141,456 humans," Nature, vol. 581, pp. 434–443, May 2020.

[66] M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman, and D. R. Maglott, "ClinVar: improving access to variant interpretations and supporting evidence," Nucleic Acids Res., vol. 46, pp. D1062–D1067, Jan. 2018.

[67] A. E. Kazdin, “The meanings and measurement of clinical significance,” J. Consult. Clin. Psychol., vol. 67, pp. 332–339, June 1999.

[68] P. Ranganathan, C. Pramesh, and M. Buyse, “Common pitfalls in statistical analysis: Clinical versus statistical significance,” Perspect. Clin. Res., vol. 6, p. 169, July 2015.

[69] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H. L. Rehm, and ACMG Laboratory Quality Assurance Committee, “Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology,” Genet. Med., vol. 17, pp. 405–24, May 2015.

[70] S. M. Harrison, E. R. Riggs, D. R. Maglott, J. M. Lee, D. R. Azzariti, A. Niehaus, E. M. Ramos, C. L. Martin, M. J. Landrum, and H. L. Rehm, “Using ClinVar as a Resource to Support Variant Interpretation,” Curr. Protoc. Hum. Genet., vol. 89, pp. 8.16.1–8.16.23, Apr. 2016.

[71] S. D. M. Brown and M. W. Moore, “The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping,” Mamm. Genome, vol. 23, pp. 632–40, Oct. 2012.

[72] P. D. Stenson, E. V. Ball, M. Mort, A. D. Phillips, J. A. Shiel, N. S. T. Thomas, S. Abeysinghe, M. Krawczak, and D. N. Cooper, “Human Gene Mutation Database (HGMD): 2003 update,” Hum. Mutat., vol. 21, pp. 577–81, June 2003.

[73] M. E. Dickinson, A. M. Flenniken, X. Ji, L. Teboul, M. D. Wong, J. K. White, T. F. Meehan, W. J. Weninger, H. Westerberg, H. Adissu, C. N. Baker, L. Bower, J. M. Brown, L. B. Caddle, F. Chiani, D. Clary, J. Cleak, M. J. Daly, J. M. Denegre, B. Doe, M. E. Dolan, S. M. Edie, H. Fuchs, V. Gailus-Durner, A. Galli, A. Gambadoro, J. Gallegos, S. Guo, N. R. Horner, C.-W. Hsu, S. J. Johnson, S. Kalaga, L. C. Keith, L. Lanoue, T. N. Lawson, M. Lek, M. Mark, S. Marschall, J. Mason, M. L. McElwee, S. Newbigging, L. M. J. Nutter, K. A. Peterson, R. Ramirez-Solis, D. J. Rowland, E. Ryder, K. E. Samocha, J. R. Seavitt, M. Selloum, Z. Szoke-Kovacs, M. Tamura, A. G. Trainor, I. Tudose, S. Wakana, J. Warren, O. Wendling, D. B. West, L. Wong, A. Yoshiki, International Mouse Phenotyping Consortium, D. G. MacArthur, G. P. Tocchini-Valentini, X. Gao, P. Flicek, A. Bradley, W. C. Skarnes, M. J. Justice, H. E. Parkinson, M. Moore, S. Wells, R. E. Braun, K. L. Svenson, M. H. de Angelis, Y. Herault, T. Mohun, A.-M. Mallon, R. M. Henkelman, S. D. M. Brown, D. J. Adams, K. C. K. Lloyd, C. McKerlie, A. L. Beaudet, M. Bućan, and S. A. Murray, “High-throughput discovery of novel developmental phenotypes,” Nature, vol. 537, pp. 508–514, Sept. 2016.

[74] A. Raj, S. A. Rifkin, E. Andersen, and A. van Oudenaarden, “Variability in gene expression underlies incomplete penetrance,” Nature, vol. 463, pp. 913–8, Feb. 2010.

[75] N. A. Karp, J. Mason, A. L. Beaudet, Y. Benjamini, L. Bower, R. E. Braun, S. D. M. Brown, E. J. Chesler, M. E. Dickinson, A. M. Flenniken, H. Fuchs, M. H. de Angelis, X. Gao, S. Guo, S. Greenaway, R. Heller, Y. Herault, M. J. Justice, N. Kurbatova, C. J. Lelliott, K. C. K. Lloyd, A.-M. Mallon, J. E. Mank, H. Masuya, C. McKerlie, T. F. Meehan, R. F. Mott, S. A. Murray, H. Parkinson, R. Ramirez-Solis, L. Santos, J. R. Seavitt, D. Smedley, T. Sorg, A. O. Speak, K. P. Steel, K. L. Svenson, International Mouse Phenotyping Consortium, S. Wakana, D. West, S. Wells, H. Westerberg, S. Yaacoby, and J. K. White, "Prevalence of sexual dimorphism in mammalian phenotypic traits," Nat. Commun., vol. 8, p. 15475, June 2017.

[76] T. F. Meehan, N. Conte, D. B. West, J. O. Jacobsen, J. Mason, J. Warren, C.-K. Chen, I. Tudose, M. Relac, P. Matthews, N. Karp, L. Santos, T. Fiegel, N. Ring, H. Westerberg, S. Greenaway, D. Sneddon, H. Morgan, G. F. Codner, M. E. Stewart, J. Brown, N. Horner, International Mouse Phenotyping Consortium, M. Haendel, N. Washington, C. J. Mungall, C. L. Reynolds, J. Gallegos, V. Gailus-Durner, T. Sorg, G. Pavlovic, L. R. Bower, M. Moore, I. Morse, X. Gao, G. P. Tocchini-Valentini, Y. Obata, S. Y. Cho, J. K. Seong, J. Seavitt, A. L. Beaudet, M. E. Dickinson, Y. Herault, W. Wurst, M. H. de Angelis, K. C. K. Lloyd, A. M. Flenniken, L. M. J. Nutter, S. Newbigging, C. McKerlie, M. J. Justice, S. A. Murray, K. L. Svenson, R. E. Braun, J. K. White, A. Bradley, P. Flicek, S. Wells, W. C. Skarnes, D. J. Adams, H. Parkinson, A.-M. Mallon, S. D. M. Brown, and D. Smedley, "Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium," Nat. Genet., vol. 49, pp. 1231–1238, Aug. 2017.

[77] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," Genome Res., vol. 20, pp. 1297–303, Sept. 2010.

[78] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," Nat. Genet., vol. 43, pp. 491–8, May 2011.

[79] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo, "From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline," Curr. Protoc. Bioinformatics, vol. 43, pp. 11.10.1–11.10.33, Oct. 2013.

[80] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," arXiv, May 2013. Available: `http://arxiv.org/abs/1303.3997v2`.

[81] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform,” Bioinformatics, vol. 25, pp. 1754–60, July 2009.

[82] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, “The Sequence Alignment/Map format and SAMtools,” Bioinformatics, vol. 25, pp. 2078–9, Aug. 2009.

[83] Broad Institute, “Picard Toolkit,” 2019. GitHub Repository: `http://broadinstitute.github.io/picard/`.

[84] R. Poplin, V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. Van der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. MacArthur, and E. Banks, “Scaling accurate genetic variant discovery to tens of thousands of samples,” bioRxiv, July 2018. Available: `https://www.biorxiv.org/content/10.1101/201178v3`.

[85] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project Analysis Group, “The variant call format and VCFtools,” Bioinformatics, vol. 27, pp. 2156–8, Aug. 2011.

[86] A. R. Carson, E. N. Smith, H. Matsui, S. K. Brækkan, K. Jepsen, J.-B. Hansen, and K. A. Frazer, “Effective filtering strategies to improve data quality from population-based whole exome sequencing studies,” BMC Bioinformatics, vol. 15, p. 125, May 2014.

[87] B. S. Pedersen and A. R. Quinlan, “Who’s Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy,” Am. J. Hum. Genet., vol. 100, pp. 406–413, Mar. 2017.

[88] G. Jun, M. Flickinger, K. N. Hetrick, J. M. Romm, K. F. Doheny, G. R. Abecasis, M. Boehnke, and H. M. Kang, “Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data,” Am. J. Hum. Genet., vol. 91, pp. 839–48, Nov. 2012.

[89] D. H. Alexander, J. Novembre, and K. Lange, “Fast model-based estimation of ancestry in unrelated individuals,” Genome Res., vol. 19, pp. 1655–64, Sept. 2009.

[90] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, “PLINK: a tool set for whole-genome association and population-based linkage analyses,” Am. J. Hum. Genet., vol. 81, pp. 559–75, Sept. 2007.

[91] D. H. Alexander and K. Lange, "Enhancements to the ADMIXTURE algorithm for individual ancestry estimation," BMC Bioinformatics, vol. 12, p. 246, June 2011.

[92] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham, "The Ensembl Variant Effect Predictor," Genome Biol., vol. 17, p. 122, June 2016.

[93] U. Paila, B. A. Chapman, R. Kirchner, and A. R. Quinlan, "GEMINI: integrative exploration of genetic variation and genome annotations," PLoS Comput. Biol., vol. 9, p. e1003153, July 2013.

[94] Q. Li and K. Wang, "InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines," Am. J. Hum. Genet., vol. 100, pp. 267–280, Feb. 2017.

[95] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: predicting the deleteriousness of variants throughout the human genome," Nucleic Acids Res., vol. 47, pp. D886–D894, Jan. 2019.

[96] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham, "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor," Bioinformatics, vol. 26, pp. 2069–70, Aug. 2010.

[97] D. G. MacArthur and C. Tyler-Smith, "Loss-of-function variants in the genomes of healthy humans," Hum. Mol. Genet., vol. 19, pp. R125–R130, Oct. 2010.

[98] R. G. H. Lindeboom, M. Vermeulen, B. Lehner, and F. Supek, "The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy," Nat. Genet., vol. 51, pp. 1645–1651, Nov. 2019.

[99] K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, and K. K.-H. Farh, "Predicting Splicing from Primary Sequence with Deep Learning," Cell, vol. 176, pp. 535–548.e24, Jan. 2019.

[100] X. Liu, C. Wu, C. Li, and E. Boerwinkle, "dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs," Hum. Mutat., vol. 37, pp. 235–41, Mar. 2016.

[101] X. Liu, X. Jian, and E. Boerwinkle, "dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions," Hum. Mutat., vol. 32, pp. 894–9, Aug. 2011.

[102] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, E. de Castro, P. Coggill, M. Corbett,

U. Das, L. Daugherty, L. Duquenne, R. D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A. F. Quinn, C. Rivoire, A. Sangrador-Vegas, J. D. Selengut, C. J. A. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajanarthanan, P. D. Thomas, C. H. Wu, C. Yeats, and S.-Y. Yong, "InterPro in 2011: new developments in the family and domain prediction database," Nucleic Acids Res., vol. 40, pp. D306–12, Jan. 2012.

[103] A. F. A. Smit, R. Hubley, and P. Green, "RepeatMasker Open-3.0," 1996–2010. Available: `http://www.repeatmasker.org`.

[104] M. Haeussler, A. S. Zweig, C. Tyner, M. L. Speir, K. R. Rosenbloom, B. J. Raney, C. M. Lee, B. T. Lee, A. S. Hinrichs, J. N. Gonzalez, D. Gibson, M. Diekhans, H. Clawson, J. Casper, G. P. Barber, D. Haussler, R. M. Kuhn, and W. J. Kent, "The UCSC Genome Browser database: 2019 update," Nucleic Acids Res., vol. 47, pp. D853–D858, Jan. 2019.

[105] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants," Nat. Genet., vol. 46, pp. 310–5, Mar. 2014.

[106] C. A. Mather, S. D. Mooney, S. J. Salipante, S. Scroggins, D. Wu, C. C. Pritchard, and B. H. Shirts, "CADD score has limited clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel," Genet. Med., vol. 18, pp. 1269–1275, Dec. 2016.

[107] E. H. Baugh, R. Simmons-Edler, C. L. Müller, R. F. Alford, N. Volfovsky, A. E. Lash, and R. Bonneau, "Robust classification of protein variation using structural modelling and large-scale data integration," Nucleic Acids Res., vol. 44, pp. 2501–13, Apr. 2016.

[108] R. Ghosh, N. Oak, and S. E. Plon, "Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines," Genome Biol., vol. 18, p. 225, Nov. 2017.

[109] K. M. Boycott, M. R. Vanstone, D. E. Bulman, and A. E. MacKenzie, "Rare-disease genetics in the era of next-generation sequencing: discovery to translation," Nat. Rev. Genet., vol. 14, pp. 681–91, Oct. 2013.

[110] P. Hager, "Haplotype estimation methods and their application in clinics," Bachelor's Thesis, Technische Universität München/Ludwig-Maximilians-Universität München, Aug. 2017.

[111] P. Hager, H.-W. Mewes, M. Rohlfs, C. Klein, and T. Jeske, “SmartPhase: Accurate and fast phasing of heterozygous variant pairs for genetic diagnosis of rare diseases,” PLoS Comput. Biol., vol. 16, p. e1007613, Feb. 2020.

[112] M. Martin, M. Patterson, S. Garg, S. O Fischer, N. Pisanti, G. W. Klau, A. Schönhuth, and T. Marschall, “WhatsHap: fast and accurate read-based phasing,” bioRxiv, Nov. 2016. Available: `https://www.biorxiv.org/content/10.1101/085050v2`.

[113] M. Claussnitzer, J. H. Cho, R. Collins, N. J. Cox, E. T. Dermitzakis, M. E. Hurles, S. Kathiresan, E. E. Kenny, C. M. Lindgren, D. G. MacArthur, K. N. North, S. E. Plon, H. L. Rehm, N. Risch, C. N. Rotimi, J. Shendure, N. Soranzo, and M. I. McCarthy, “A brief history of human disease genetics,” Nature, vol. 577, pp. 179–189, Jan. 2020.

[114] M. Kozak, “An analysis of 5’-noncoding sequences from 699 vertebrate messenger RNAs,” Nucleic Acids Res., vol. 15, pp. 8125–48, Oct. 1987.

[115] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, “KNIME: The Konstanz Information Miner,” in Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer, 2007.

[116] M. Hastreiter, T. Jeske, J. Hoser, M. Kluge, K. Ahomaa, M.-S. Friedl, S. J. Kopetzky, J.-D. Quell, H.-W. Mewes, and R. Küffner, “KNIME4NGS: A comprehensive toolbox for next generation sequencing analysis,” Bioinformatics, vol. 33, pp. 1565–1567, May 2017.

[117] K. Ahomaa, “Detection of causal variants in rare children’s diseases from a combination of whole-genome and whole-exome sequencing,” Master’s Thesis, Technische Universität München, Nov. 2016.

[118] C. F. Wright, J. F. McRae, S. Clayton, G. Gallone, S. Aitken, T. W. FitzGerald, P. Jones, E. Prigmore, D. Rajan, J. Lord, A. Sifrim, R. Kelsell, M. J. Parker, J. C. Barrett, M. E. Hurles, D. R. FitzPatrick, H. V. Firth, and DDD Study, “Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders,” Genet. Med., vol. 20, pp. 1216–1223, Oct. 2018.

[119] A. Yates, W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann,

A. Vullo, S. P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G. Spudich, S. J. Trevanion, F. Cunningham, B. L. Aken, D. R. Zerbino, and P. Flicek, "Ensembl 2016," Nucleic Acids Res., vol. 44, pp. D710–6, Jan. 2016.

[120] R. Carapito, M. Konantz, C. Paillard, Z. Miao, A. Pichot, M. S. Leduc, Y. Yang, K. L. Bergstrom, D. H. Mahoney, D. L. Shardy, G. Alsaleh, L. Naegely, A. Kolmer, N. Paul, A. Hanauer, V. Rolli, J. S. Müller, E. Alghisi, L. Sauteur, C. Macquin, A. Morlon, C. S. Sancho, P. Amati-Bonneau, V. Procaccio, A.-L. Mosca-Boidron, N. Marle, N. Osmani, O. Lefebvre, J. G. Goetz, S. Unal, N. A. Akarsu, M. Radosavljevic, M.-P. Chenard, F. Rialland, A. Grain, M.-C. Béné, M. Eveillard, M. Vincent, J. Guy, L. Faivre, C. Thauvin-Robinet, J. Thevenon, K. Myers, M. D. Fleming, A. Shimamura, E. Bottollier-Lemallaz, E. Westhof, C. Lengerke, B. Isidor, and S. Bahram, "Mutations in signal recognition particle SRP54 cause syndromic neutropenia with Shwachman-Diamond-like features," J. Clin. Invest., vol. 127, pp. 4090–4103, Nov. 2017.

[121] C. Bellanné-Chantelot, B. Schmaltz-Panneau, C. Marty, O. Fenneteau, I. Callebaut, S. Clauin, A. Docet, G.-L. Damaj, T. Leblanc, I. Pellier, C. Stoven, S. Souquere, I. Antony-Debré, B. Beaupain, N. Aladjidi, V. Barlogis, F. Bauduer, P. Bensaid, O. Boespflug-Tanguy, C. Berger, Y. Bertrand, L. Carausu, C. Fieschi, C. Galambrun, A. Schmidt, H. Journel, F. Mazingue, B. Nelken, T. C. Quah, E. Oksenhendler, M. Ouachée, M. Pasquet, V. Saada, F. Suarez, G. Pierron, W. Vainchenker, I. Plo, and J. Donadieu, "Mutations in the SRP54 gene cause severe congenital neutropenia as well as Shwachman-Diamond-like syndrome," Blood, vol. 132, pp. 1318–1331, Sept. 2018.

[122] K. Wild, G. Bange, D. Motiejunas, J. Kribelbauer, A. Hendricks, B. Segnitz, R. C. Wade, and I. Sinning, "Structural Basis for Conserved Regulation and Adaptation of the Signal Recognition Particle Targeting Complex," J. Mol. Biol., vol. 428, pp. 2880–97, July 2016.

[123] D. Akopian, K. Shen, X. Zhang, and S.-o. Shan, "Signal recognition particle: an essential protein-targeting machine," Annu. Rev. Biochem., vol. 82, pp. 693–721, June 2013.

[124] Y. Mizoguchi, S. Hesse, M. Linder, N. Zietara, M. Lyszkiewicz, Y. Liu, M. Tatematsu, P. Grabowski, K. Ahomaa, T. Jeske, S. Hollizeck, E. Rusha, M. K. Saito, M. Kobayashi, Z. Alizadeh, Z. Pourpak, S. Iurian, N. Rezaei, E. Unal, M. Drukker, B. Walzog, F. Hauck, J. Rappsilber, and C. Klein, "Defects in Signal Recognition Particle (SRP) Components Reveal an Essential and Non-Redundant Role for Granule Biogenesis and Differentiation of Neutrophil Granulocytes," Blood, vol. 134, pp. 216–216, Nov. 2019.

[125] L. Frésard and S. B. Montgomery, “Diagnosing rare diseases after the exome,” Cold Spring Harb. Mol. case Stud., vol. 4, p. a003392, Dec. 2018.

[126] J. Wang, L. Raskin, D. C. Samuels, Y. Shyr, and Y. Guo, “Genome measures used for quality control are dependent on gene function and ancestry,” Bioinformatics, vol. 31, pp. 318–23, Feb. 2015.

[127] M. F. Lin, I. Jungreis, and M. Kellis, “PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions,” Bioinformatics, vol. 27, pp. i275–82, July 2011.

[128] E.-O. Glocker, D. Kotlarz, K. Boztug, E. M. Gertz, A. A. Schäffer, F. Noyan, M. Perro, J. Diestelhorst, A. Allroth, D. Murugan, N. Hätscher, D. Pfeifer, K.-W. Sykora, M. Sauer, H. Kreipe, M. Lacher, R. Nustede, C. Woellner, U. Baumann, U. Salzer, S. Koletzko, N. Shah, A. W. Segal, A. Sauerbrey, S. Buderus, S. B. Snapper, B. Grimbacher, and C. Klein, “Inflammatory bowel disease and mutations affecting the interleukin-10 receptor,” N. Engl. J. Med., vol. 361, pp. 2033–45, Nov. 2009.

[129] T. Nochi, Y. Yuki, K. Terahara, A. Hino, J. Kunisawa, M.-N. Kweon, T. Yamaguchi, and H. Kiyono, “Biological role of Ep-CAM in the physical interaction between epithelial cells and lymphocytes in intestinal epithelium,” Clin. Immunol., vol. 113, pp. 326–39, Dec. 2004.

[130] M. Sivagnanam, J. L. Mueller, H. Lee, Z. Chen, S. F. Nelson, D. Turner, S. H. Zlotkin, P. B. Pencharz, B.-Y. Ngan, O. Libiger, N. J. Schork, J. E. Lavine, S. Taylor, R. O. Newbury, R. D. Kolodner, and H. M. Hoffman, “Identification of EpCAM as the gene for congenital tufting enteropathy,” Gastroenterology, vol. 135, pp. 429–37, Aug. 2008.

[131] E. M. Wright, D. D. Loo, M. Panayotova-Heiermann, M. P. Lostao, B. H. Hirayama, B. Mackenzie, K. Boorer, and G. Zampighi, “'Active' sugar transport in eukaryotes,” J. Exp. Biol., vol. 196, pp. 197–212, Nov. 1994.

[132] E. Turk, B. Zabel, S. Mundlos, J. Dyer, and E. M. Wright, “Glucose/galactose malabsorption caused by a defect in the Na+/glucose cotransporter,” Nature, vol. 350, pp. 354–6, Mar. 1991.

[133] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto, and R. D. Finn, “The Pfam protein families database in 2019,” Nucleic Acids Res., vol. 47, pp. D427–D432, Jan. 2019.

[134] UniProt Consortium, “UniProt: a worldwide hub of protein knowledge,” Nucleic Acids Res., vol. 47, pp. D506–D515, Jan. 2019.

[135] M. Pavšič, G. Gunčar, K. Djinović-Carugo, and B. Lenarčič, "Crystal structure and its bearing towards an understanding of key biological functions of EpCAM," Nat. Commun., vol. 5, p. 4764, Aug. 2014.

[136] U. Schnell, J. Kuipers, J. L. Mueller, A. Veenstra-Algra, M. Sivagnanam, and B. N. G. Giepmans, "Absence of cell-surface EpCAM in congenital tufting enteropathy," Hum. Mol. Genet., vol. 22, pp. 2566–71, July 2013.

[137] R. Roncagalli, M. Cucchetti, N. Jarmuzynski, C. Grégoire, E. Bergot, S. Audebert, E. Baudelet, M. G. Menoita, A. Joachim, S. Durand, M. Suchanek, F. Fiore, L. Zhang, Y. Liang, L. Camoin, M. Malissen, and B. Malissen, "The scaffolding function of the RLTPR protein explains its essential role for CD28 co-stimulation in mouse and human T cells," J. Exp. Med., vol. 213, pp. 2437–2457, Oct. 2016.

[138] T. Schober, T. Magg, M. Laschinger, M. Rohlfs, N. D. Linhares, J. Puchalka, T. Weisser, K. Fehlner, J. Mautner, C. Walz, K. Hussein, G. Jaeger, B. Kammer, I. Schmid, M. Bahia, S. D. Pena, U. Behrends, B. H. Belohradsky, C. Klein, and F. Hauck, "A human immunodeficiency syndrome caused by mutations in CARMIL2," Nat. Commun., vol. 8, p. 14209, Jan. 2017.

[139] T. Magg, A. Shcherbina, D. Arslan, M. M. Desai, S. Wall, V. Mitsialis, R. Conca, E. Unal, N. Karacabey, A. Mukhina, Y. Rodina, P. D. Taur, D. Illig, B. Marquardt, S. Hollizeck, T. Jeske, F. Gothe, T. Schober, M. Rohlfs, S. Koletzko, E. Lurz, A. M. Muise, S. B. Snapper, F. Hauck, C. Klein, and D. Kotlarz, "CARMIL2 Deficiency Presenting as Very Early Onset Inflammatory Bowel Disease," Inflamm. Bowel Dis., vol. 25, pp. 1788–1795, Oct. 2019.

[140] J. D. Fontenot, M. A. Gavin, and A. Y. Rudensky, "Foxp3 programs the development and function of CD4+CD25+ regulatory T cells," Nat. Immunol., vol. 4, pp. 330–6, Apr. 2003.

[141] C. L. Bennett, J. Christie, F. Ramsdell, M. E. Brunkow, P. J. Ferguson, L. Whitesell, T. E. Kelly, F. T. Saulsbury, P. F. Chance, and H. D. Ochs, "The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3," Nat. Genet., vol. 27, pp. 20–1, Jan. 2001.

[142] B. Li, A. Samanta, X. Song, K. T. Iacono, P. Brennan, T. A. Chatila, G. Roncador, A. H. Banham, J. L. Riley, Q. Wang, Y. Shen, S. J. Saouaf, and M. I. Greene, "FOXP3 is a homo-oligomer and a component of a supramolecular regulatory complex disabled in the human XLAAD/IPEX autoimmune disease," Int. Immunol., vol. 19, pp. 825–835, July 2007.

[143] D. T. Duztas, L. Al-Shadfan, H. Ozturk, H. Yazan, E. Cakir, N. Unver, O. Ekinci, B. Dalgic, M. Rohlfs, T. Jeske, C. Klein, D. Kotlarz, and O. E. Gurkan, "New

Findings of Immunodysregulation, Polyendocrinopathy, and Enteropathy X-linked Syndrome (IPEX); Granulomas in Lung and Duodenum," Pediatr. Dev. Pathol., Mar. 2021.

[144] F. Rahmani, E. Rayzan, M. R. Rahmani, S. Shahkarami, S. Zoghi, A. Rezaei, Z. Aryan, M. Najafi, M. Rohlfs, T. Jeske, M. Aflatoonian, Z. Chavoshzadeh, F. Farahmand, F. Motamed, P. Rohani, H. Alimadadi, A. Mahdaviani, M. Mansouri, M. Tavakol, M. Vanderberg, D. Kotlarz, C. Klein, and N. Rezaei, "Clinical and Mutation Description of the First Iranian Cohort of Infantile Inflammatory Bowel Disease: The Iranian Primary Immunodeficiency Registry (IPIDR)," Immunol. Invest., vol. 50, pp. 445–459, May 2021.

[145] C. C. Martin, J. K. Oeser, C. A. Svitek, S. I. Hunter, J. C. Hutton, and R. M. O'Brien, "Identification and characterization of a human cDNA and gene encoding a ubiquitously expressed glucose-6-phosphatase catalytic subunit-related protein," J. Mol. Endocrinol., vol. 29, pp. 205–22, Oct. 2002.

[146] K. Boztug, G. Appaswamy, A. Ashikov, A. A. Schäffer, U. Salzer, J. Diestelhorst, M. Germeshausen, G. Brandes, J. Lee-Gossler, F. Noyan, A.-K. Gatzke, M. Minkov, J. Greil, C. Kratz, T. Petropoulou, I. Pellier, C. Bellanné-Chantelot, N. Rezaei, K. Mönkemöller, N. Irani-Hakimeh, H. Bakker, R. Gerardy-Schahn, C. Zeidler, B. Grimbacher, K. Welte, and C. Klein, "A syndrome with congenital neutropenia and mutations in G6PC3," N. Engl. J. Med., vol. 360, pp. 32–43, Jan. 2009.

[147] M. R. Pool, J. Stumm, T. A. Fulga, I. Sinning, and B. Dobberstein, "Distinct modes of signal recognition particle interaction with the ribosome," Science, vol. 297, pp. 1345–8, Aug. 2002.

[148] J.-B. Vannier, S. Sandhu, M. I. R. Petalcorin, X. Wu, Z. Nabi, H. Ding, and S. J. Boulton, "RTEL1 is a replisome-associated helicase that promotes telomere and genome-wide replication," Science, vol. 342, pp. 239–42, Oct. 2013.

[149] B. J. Ballew, V. Joseph, S. De, G. Sarek, J.-B. Vannier, T. Stracker, K. A. Schrader, T. N. Small, R. O'Reilly, C. Manschreck, M. M. Harlan Fleischut, L. Zhang, J. Sullivan, K. Stratton, M. Yeager, K. Jacobs, N. Giri, B. P. Alter, J. Boland, L. Burdett, K. Offit, S. J. Boulton, S. A. Savage, and J. H. J. Petrini, "A recessive founder mutation in regulator of telomere elongation helicase 1, RTEL1, underlies severe immunodeficiency and features of Hoyeraal Hreidarsson syndrome," PLoS Genet., vol. 9, p. e1003695, Aug. 2013.

[150] A. J. Walne, T. Vulliamy, M. Kirwan, V. Plagnol, and I. Dokal, "Constitutional mutations in RTEL1 cause severe dyskeratosis congenita," Am. J. Hum. Genet., vol. 92, pp. 448–53, Mar. 2013.

[151] A. Ziv, L. Werner, L. Konnikova, A. Awad, T. Jeske, M. Hastreiter, V. Mitsialis, T. Stauber, S. Wall, D. Kotlarz, C. Klein, S. B. Snapper, Y. Tzfati, B. Weiss, R. Somech, and D. S. Shouval, "An RTEL1 Mutation Links to Infantile-Onset Ulcerative Colitis and Severe Immunodeficiency," J. Clin. Immunol., vol. 40, pp. 1010–1019, Oct. 2020.

[152] A. M. Fedick, L. Shi, C. Jalas, N. R. Treff, J. Ekstein, R. Kornreich, L. Edelmann, L. Mehta, and S. A. Savage, "Carrier screening of RTEL1 mutations in the Ashkenazi Jewish population," Clin. Genet., vol. 88, pp. 177–81, Aug. 2015.

[153] A. S. Venteicher, E. B. Abreu, Z. Meng, K. E. McCann, R. M. Terns, T. D. Veenstra, M. P. Terns, and S. E. Artandi, "A human telomerase holoenzyme protein required for Cajal body localization and telomere synthesis," Science, vol. 323, pp. 644–8, Jan. 2009.

[154] S. Schwartz, D. A. Bernstein, M. R. Mumbach, M. Jovanovic, R. H. Herbst, B. X. León-Ricardo, J. M. Engreitz, M. Guttman, R. Satija, E. S. Lander, G. Fink, and A. Regev, "Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA," Cell, vol. 159, pp. 148–162, Sept. 2014.

[155] A. Marrone and I. Dokal, "Dyskeratosis congenita: molecular insights into telomerase function, ageing and cancer," Expert Rev. Mol. Med., vol. 6, pp. 1–23, Dec. 2004.

[156] F.-O. Desmet, D. Hamroun, M. Lalande, G. Collod-Béroud, M. Claustres, and C. Béroud, "Human Splicing Finder: an online bioinformatics tool to predict splicing signals," Nucleic Acids Res., vol. 37, p. e67, May 2009.

[157] Q. Zhang, X. Fan, Y. Wang, M.-A. Sun, J. Shao, and D. Guo, "BPP: a sequence-based algorithm for branch point prediction," Bioinformatics, vol. 33, pp. 3166–3172, Oct. 2017.

[158] M. Chen and J. L. Manley, "Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches," Nat. Rev. Mol. Cell Biol., vol. 10, pp. 741–54, Nov. 2009.

[159] I. V. Kovtun and C. T. McMurray, "Features of trinucleotide repeat instability in vivo," Cell Res., vol. 18, pp. 198–213, Jan. 2008.

[160] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. von Mering, "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," Nucleic Acids Res., vol. 47, pp. D607–D613, Jan. 2019.

[161] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," Genome Res., vol. 13, pp. 2498–504, Nov. 2003.

[162] W.-T. Kim, H. Seo Choi, H. Min Lee, Y.-J. Jang, and C. J. Ryu, "B-cell receptor-associated protein 31 regulates human embryonic stem cell adhesion, stemness, and survival via control of epithelial cell adhesion molecule," Stem Cells, vol. 32, pp. 2626–41, Oct. 2014.

[163] P. Cacciagli, J. Sutera-Sardo, A. Borges-Correia, J.-C. Roux, I. Dorboz, J.-P. Desvignes, C. Badens, M. Delepine, M. Lathrop, P. Cau, N. Lévy, N. Girard, P. Sarda, O. Boespflug-Tanguy, and L. Villard, "Mutations in BCAP31 cause a severe X-linked phenotype with deafness, dystonia, and central hypomyelination and disorganize the Golgi apparatus," Am. J. Hum. Genet., vol. 93, pp. 579–86, Sept. 2013.

[164] M. Kuokkanen, J. Kokkonen, N. S. Enattah, T. Ylisaukko-Oja, H. Komu, T. Varilo, L. Peltonen, E. Savilahti, and I. Jarvela, "Mutations in the translated region of the lactase gene (LCT) underlie congenital lactase deficiency," Am. J. Hum. Genet., vol. 78, pp. 339–44, Feb. 2006.

[165] T. A. Rouault and W. H. Tong, "Iron-sulfur cluster biogenesis and human disease," Trends Genet., vol. 24, pp. 398–407, Aug. 2008.

[166] R. Waltes, R. Kalb, M. Gatei, A. W. Kijas, M. Stumm, A. Sobeck, B. Wieland, R. Varon, Y. Lerenthal, M. F. Lavin, D. Schindler, and T. Dörk, "Human RAD50 deficiency in a Nijmegen breakage syndrome-like disorder," Am. J. Hum. Genet., vol. 84, pp. 605–16, May 2009.

[167] I. Puls, C. Jonnakuty, B. H. LaMonte, E. L. F. Holzbaur, M. Tokito, E. Mann, M. K. Floeter, K. Bidus, D. Drayna, S. J. Oh, R. H. Brown, C. L. Ludlow, and K. H. Fischbeck, "Mutant dynactin in motor neuron disease," Nat. Genet., vol. 33, pp. 455–6, Apr. 2003.

[168] M. J. Farrer, M. M. Hulihan, J. M. Kachergus, J. C. Dächsel, A. J. Stoessl, L. L. Grantier, S. Calne, D. B. Calne, B. Lechevalier, F. Chapon, Y. Tsuboi, T. Yamada, L. Gutmann, B. Elibol, K. P. Bhatia, C. Wider, C. Vilariño-Güell, O. A. Ross, L. A. Brown, M. Castanedes-Casey, D. W. Dickson, and Z. K. Wszolek, "DCTN1 mutations in Perry syndrome," Nat. Genet., vol. 41, pp. 163–5, Feb. 2009.

[169] D. Schubert, C. Bode, R. Kenefeck, T. Z. Hou, J. B. Wing, A. Kennedy, A. Bulashevska, B.-S. Petersen, A. A. Schäffer, B. A. Grüning, S. Unger, N. Frede, U. Baumann, T. Witte, R. E. Schmidt, G. Dueckers, T. Niehues, S. Seneviratne, M. Kanariou, C. Speckmann, S. Ehl, A. Rensing-Ehl, K. Warnatz, M. Rakhmanov,

R. Thimme, P. Hasselblatt, F. Emmerich, T. Cathomen, R. Backofen, P. Fisch, M. Seidl, A. May, A. Schmitt-Graeff, S. Ikemizu, U. Salzer, A. Franke, S. Sakaguchi, L. S. K. Walker, D. M. Sansom, and B. Grimbacher, “Autosomal dominant immune dysregulation syndrome in humans with CTLA4 mutations,” Nat. Med., vol. 20, pp. 1410–1416, Dec. 2014.

[170] Y. Wang, C. S. Ma, Y. Ling, A. Bousfiha, Y. Camcioglu, S. Jacquot, K. Payne, E. Crestani, R. Roncagalli, A. Belkadi, G. Kerner, L. Lorenzo, C. Deswarte, M. Chrabieh, E. Patin, Q. B. Vincent, I. Müller-Fleckenstein, B. Fleckenstein, F. Ailal, L. Quintana-Murci, S. Fraitag, M.-A. Alyanakian, M. Leruez-Ville, C. Picard, A. Puel, J. Bustamante, S. Boisson-Dupuis, M. Malissen, B. Malissen, L. Abel, A. Hovnanian, L. D. Notarangelo, E. Jouanguy, S. G. Tangye, V. Béziat, and J.-L. Casanova, “Dual T cell- and B cell-intrinsic deficiency in humans with biallelic RLTPR mutations,” J. Exp. Med., vol. 213, pp. 2413–2435, Oct. 2016.

[171] M. Ono, H. Yaguchi, N. Ohkura, I. Kitabayashi, Y. Nagamura, T. Nomura, Y. Miyachi, T. Tsukada, and S. Sakaguchi, “Foxp3 controls regulatory T-cell function by interacting with AML1/Runx1,” Nature, vol. 446, pp. 685–9, Apr. 2007.

[172] T. Maj, W. Wang, J. Crespo, H. Zhang, W. Wang, S. Wei, L. Zhao, L. Vatan, I. Shao, W. Szeliga, C. Lyssiotis, J. R. Liu, I. Kryczek, and W. Zou, “Oxidative stress controls regulatory T cell apoptosis and suppressor activity and PD-L1-blockade resistance in tumor,” Nat. Immunol., vol. 18, pp. 1332–1341, Dec. 2017.

[173] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow, “PID: the Pathway Interaction Database,” Nucleic Acids Res., vol. 37, pp. D674–9, Jan. 2009.

[174] S. B. Cohen, M. E. Graham, G. O. Lovrecz, N. Bache, P. J. Robinson, and R. R. Reddel, “Protein composition of catalytically active human telomerase from immortal cells,” Science, vol. 315, pp. 1850–3, Mar. 2007.

[175] H. Kobayashi, K. Abe, T. Matsuura, Y. Ikeda, T. Hitomi, Y. Akechi, T. Habu, W. Liu, H. Okuda, and A. Koizumi, “Expansion of intronic GGCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement,” Am. J. Hum. Genet., vol. 89, pp. 121–30, July 2011.

[176] E. Turro, W. J. Astle, K. Megy, S. Gräf, D. Greene, O. Shamardina, H. L. Allen, A. Sanchis-Juan, M. Frontini, C. Thys, J. Stephens, R. Mapeta, O. S. Burren, K. Downes, M. Haimel, S. Tuna, S. V. V. Deevi, T. J. Aitman, D. L. Bennett, P. Calleja, K. Carss, M. J. Caulfield, P. F. Chinnery, P. H. Dixon, D. P. Gale, R. James, A. Koziell, M. A. Laffan, A. P. Levine, E. R. Maher, H. S. Markus, J. Morales, N. W. Morrell, A. D. Mumford, E. Ormondroyd, S. Rankin, A. Rendon, S. Richardson, I. Roberts, N. B. A. Roy, M. A. Saleem, K. G. C. Smith, H. Stark,

R. Y. Y. Tan, A. C. Themistocleous, A. J. Thrasher, H. Watkins, A. R. Webster, M. R. Wilkins, C. Williamson, J. Whitworth, S. Humphray, D. R. Bentley, NIHR BioResource for the 100,000 Genomes Project, N. Kingston, N. Walker, J. R. Bradley, S. Ashford, C. J. Penkett, K. Freson, K. E. Stirrups, F. L. Raymond, and W. H. Ouwehand, "Whole-genome sequencing of patients with rare diseases in a national health system," Nature, vol. 583, pp. 96–102, July 2020.

[177] A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, and P. Flicek, "GENCODE reference annotation for the human and mouse genomes," Nucleic Acids Res., vol. 47, pp. D766–D773, Jan. 2019.

[178] L. A. Schuch, M. Forstner, C. K. Rapp, Y. Li, D. E. C. Smith, M. I. Mendes, F. Delhommel, M. Sattler, N. Emiralioğlu, E. Z. Taskiran, D. Orhan, N. Kiper, M. Rohlfs, T. Jeske, M. Hastreiter, M. Gerstlauer, A. Torrent-Vernetta, A. Moreno-Galdó, B. Kammer, F. Brasch, S. Reu-Hofer, and M. Griese, "FARS1-related disorders caused by biallelic mutations in cytosolic phenylalanyl-tRNA synthetase genes: Look beyond the lungs!," Clin. Genet., vol. 99, pp. 789–801, June 2021.

[179] P. S. Appelbaum, E. Parens, S. M. Berger, W. K. Chung, and W. Burke, "Is there a duty to reinterpret genetic data? The ethical dimensions," Genet. Med., vol. 22, pp. 633–639, Mar. 2020.

[180] D. Mandelker, R. J. Schmidt, A. Ankala, K. McDonald Gibson, M. Bowser, H. Sharma, E. Duffy, M. Hegde, A. Santani, M. Lebo, and B. Funke, "Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing," Genet. Med., vol. 18, pp. 1282–1289, Dec. 2016.

[181] J. Roesler, J. T. Curnutte, J. Rae, D. Barrett, P. Patino, S. J. Chanock, and A. Goerlach, "Recombination events between the p47-phox gene and its highly homologous pseudogenes are the main cause of autosomal recessive chronic granulomatous disease," Blood, vol. 95, pp. 2150–2156, Mar. 2000.

[182] H. Dashnow, K. M. Bell, Z. Stark, T. Y. Tan, S. M. White, and A. Oshlack, "Pooled-parent exome sequencing to prioritise de novo variants in genetic disease," bioRxiv, Apr. 2019. Available: https://www.biorxiv.org/content/early/2019/04/07/601740.

[183] S. Köhler, L. Carmody, N. Vasilevsky, J. O. B. Jacobsen, D. Danis, J.-P. Gourdine, M. Gargano, N. L. Harris, N. Matentzoglu, J. A. McMurry, D. Osumi-Sutherland, V. Cipriani, J. P. Balhoff, T. Conlin, H. Blau, G. Baynam, R. Palmer, D. Gratian, H. Dawkins, M. Segal, A. C. Jansen, A. Muaz, W. H. Chang, J. Bergerson, S. J. F. Laulederkind, Z. Yüksel, S. Beltran, A. F. Freeman, P. I. Sergouniotis, D. Durkin, A. L. Storm, M. Hanauer, M. Brudno, S. M. Bello, M. Sincan, K. Rageth, M. T. Wheeler, R. Oegema, H. Lourghi, M. G. Della Rocca, R. Thompson, F. Castellanos, J. Priest, C. Cunningham-Rundles, A. Hegde, R. C. Lovering, C. Hajek, A. Olry, L. Notarangelo, M. Similuk, X. A. Zhang, D. Gómez-Andrés, H. Lochmüller, H. Dollfus, S. Rosenzweig, S. Marwaha, A. Rath, K. Sullivan, C. Smith, J. D. Milner, D. Leroux, C. F. Boerkoel, A. Klion, M. C. Carter, T. Groza, D. Smedley, M. A. Haendel, C. Mungall, and P. N. Robinson, "Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources," Nucleic Acids Res., vol. 47, pp. D1018–D1027, Jan. 2019.

[184] D. Sánchez, M. Batet, D. Isern, and A. Valls, "Ontology-based semantic similarity: A new feature-based approach," Expert Syst. Appl., vol. 39, pp. 7718–7728, July 2012.

[185] D. Smedley, J. O. B. Jacobsen, M. Jäger, S. Köhler, M. Holtgrewe, M. Schubach, E. Siragusa, T. Zemojtel, O. J. Buske, N. L. Washington, W. P. Bone, M. A. Haendel, and P. N. Robinson, "Next-generation diagnostics and disease-gene discovery with the Exomiser," Nat. Protoc., vol. 10, pp. 2004–15, Dec. 2015.

[186] S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower, T. J. Callahan, C. G. Chute, J. L. Est, P. D. Galer, S. Ganesan, M. Griese, M. Haimel, J. Pazmandi, M. Hanauer, N. L. Harris, M. J. Hartnett, M. Hastreiter, F. Hauck, Y. He, T. Jeske, H. Kearney, G. Kindle, C. Klein, K. Knoflach, R. Krause, D. Lagorce, J. A. McMurry, J. A. Miller, M. C. Munoz-Torres, R. L. Peters, C. K. Rapp, A. M. Rath, S. A. Rind, A. Z. Rosenberg, M. M. Segal, M. G. Seidel, D. Smedley, T. Talmy, Y. Thomas, S. A. Wiafe, J. Xian, Z. Yüksel, I. Helbig, C. J. Mungall, M. A. Haendel, and P. N. Robinson, "The Human Phenotype Ontology in 2021," Nucleic Acids Res., vol. 49, pp. D1207–D1217, Jan. 2021.

[187] J. O. B. Jacobsen, P. N. Robinson, and C. J. Mungall, "Phenopackets Documentation," 2019. Available: `https://phenopackets-schema.readthedocs.io`.

[188] M. G. Seidel, G. Kindle, B. Gathmann, I. Quinti, M. Buckland, J. van Montfrans, R. Scheible, S. Rusch, L. M. Gasteiger, B. Grimbacher, N. Mahlaoui, S. Ehl, M. Abinun, M. Albert, S. B. Cohen, J. Bustamante, A. Cant, J.-L. Casanova, H. Chapel, G. de Saint Basile, E. de Vries, I. Dokal, J. Donadieu, A. Durandy, D. Edgar, T. Espanol, A. Etzioni, A. Fischer, B. Gaspar, R. Gatti, A. Gennery, S. Grigoriadou,

S. Holland, G. Janka, M. Kanariou, C. Klein, H. Lachmann, D. Lilic, A. Manson, N. Martinez, I. Meyts, N. Moes, D. Moshous, B. Neven, H. Ochs, C. Picard, E. Renner, F. Rieux-Laucat, R. Seger, A. Soresina, D. Stoppa-Lyonnet, V. Thon, A. Thrasher, F. van de Veerdonk, A. Villa, C. Weemaes, K. Warnatz, B. Wolska, and S.-Y. Zhang, "The European Society for Immunodeficiencies (ESID) Registry Working Definitions for the Clinical Diagnosis of Inborn Errors of Immunity," J. Allergy Clin. Immunol. Pract., vol. 7, pp. 1763–1770, July 2019.

[189] B. Schmaltz-Panneau, A. Pagnier, S. Clauin, J. Buratti, C. Marty, O. Fenneteau, K. Dieterich, B. Beaupain, J. Donadieu, I. Plo, and C. Bellanné-Chantelot, "Identification of biallelic germline variants of SRP68 in a sporadic case with severe congenital neutropenia," Haematologica, vol. 106, pp. 1216–1219, Apr. 2021.

[190] C. Schürch, T. Schaefer, J. S. Müller, P. Hanns, M. Arnone, A. Dumlin, J. Schärer, I. Sinning, K. Wild, J. Skokowa, K. Welte, R. Carapito, S. Bahram, M. Konantz, and C. Lengerke, "SRP54 mutations induce congenital neutropenia via dominant-negative effects on XBP1 splicing," Blood, vol. 137, pp. 1340–1352, Mar. 2021.

[191] S. Kanda, K. Yanagitani, Y. Yokota, Y. Esaki, and K. Kohno, "Autonomous translational pausing is required for XBP1u mRNA recruitment to the ER via the SRP pathway," Proc. Natl. Acad. Sci. U. S. A., vol. 113, pp. E5886–E5895, Oct. 2016.

[192] A. Anna and G. Monika, "Splicing mutations in human genetic disorders: examples, detection, and confirmation," J. Appl. Genet., vol. 59, pp. 253–268, Aug. 2018.

[193] K. M. Boycott, T. Hartley, L. G. Biesecker, R. A. Gibbs, A. M. Innes, O. Riess, J. Belmont, S. L. Dunwoodie, N. Jojic, T. Lassmann, D. Mackay, I. K. Temple, A. Visel, and G. Baynam, "A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers," Cell, vol. 177, pp. 32–37, Mar. 2019.

[194] T. Hartley, G. Lemire, K. D. Kernohan, H. E. Howley, D. R. Adams, and K. M. Boycott, "New Diagnostic Approaches for Undiagnosed Rare Genetic Diseases," Annu. Rev. Genomics Hum. Genet., vol. 21, pp. 351–372, Aug. 2020.

[195] A. A. Schäffer, "Digenic inheritance in medical genetics," J. Med. Genet., vol. 50, pp. 641–52, Oct. 2013.

[196] A. Gazzo, D. Raimondi, D. Daneels, Y. Moreau, G. Smits, S. Van Dooren, and T. Lenaerts, "Understanding mutational effects in digenic diseases," Nucleic Acids Res., vol. 45, p. e140, Sept. 2017.

[197] A. M. Gazzo, D. Daneels, E. Cilia, M. Bonduelle, M. Abramowicz, S. Van Dooren, G. Smits, and T. Lenaerts, "DIDA: A curated and annotated digenic diseases database," Nucleic Acids Res., vol. 44, pp. D900–7, Jan. 2016.

[198] K. Schwarze, J. Buchanan, J. C. Taylor, and S. Wordsworth, “Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature,” Genet. Med., vol. 20, pp. 1122–1130, Oct. 2018.

[199] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits,” Proc. Natl. Acad. Sci. U. S. A., vol. 106, pp. 9362–7, June 2009.

[200] A. J. Schork, W. K. Thompson, P. Pham, A. Torkamani, J. C. Roddey, P. F. Sullivan, J. R. Kelsoe, M. C. O’Donovan, H. Furberg, Tobacco and Genetics Consortium, Bipolar Disorder Psychiatric Genomics Consortium, Schizophrenia Psychiatric Genomics Consortium, N. J. Schork, O. A. Andreassen, and A. M. Dale, “All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs,” PLoS Genet., vol. 9, p. e1003449, Apr. 2013.

[201] J. Labory, M. Fierville, S. Ait-El-Mkadem, S. Bannwarth, V. Paquis-Flucklinger, and S. Bottini, “Multi-Omics Approaches to Improve Mitochondrial Disease Diagnosis: Challenges, Advances, and Perspectives,” Front. Mol. Biosci., vol. 7, p. 590842, Nov. 2020.

[202] J. Lee, D. Y. Hyeon, and D. Hwang, “Single-cell multiomics: technologies and data analysis methods,” Exp. Mol. Med., vol. 52, pp. 1428–1442, Sept. 2020.

[203] K. J. Karczewski and M. P. Snyder, “Integrative omics for health and disease,” Nat. Rev. Genet., vol. 19, pp. 299–310, May 2018.

[204] B. B. Cummings, J. L. Marshall, T. Tukiainen, M. Lek, S. Donkervoort, A. R. Foley, V. Bolduc, L. B. Waddell, S. A. Sandaradura, G. L. O’Grady, E. Estrella, H. M. Reddy, F. Zhao, B. Weisburd, K. J. Karczewski, A. H. O’Donnell-Luria, D. Birnbaum, A. Sarkozy, Y. Hu, H. Gonorazky, K. Claeys, H. Joshi, A. Bournazos, E. C. Oates, R. Ghaoui, M. R. Davis, N. G. Laing, A. Topf, Genotype-Tissue Expression Consortium, P. B. Kang, A. H. Beggs, K. N. North, V. Straub, J. J. Dowling, F. Muntoni, N. F. Clarke, S. T. Cooper, C. G. Bönnemann, and D. G. MacArthur, “Improving genetic diagnosis in Mendelian disease with transcriptome sequencing,” Sci. Transl. Med., vol. 9, p. eaal5209, Apr. 2017.

[205] L. S. Kremer, D. M. Bader, C. Mertes, R. Kopajtich, G. Pichler, A. Iuso, T. B. Haack, E. Graf, T. Schwarzmayr, C. Terrile, E. Koňaříková, B. Repp, G. Kastenmüller, J. Adamski, P. Lichtner, C. Leonhardt, B. Funalot, A. Donati, V. Tiranti, A. Lombes, C. Jardel, D. Gläser, R. W. Taylor, D. Ghezzi, J. A. Mayr, A. Rötig, P. Freisinger, F. Distelmaier, T. M. Strom, T. Meitinger, J. Gagneur, and H. Prokisch, “Genetic diagnosis of Mendelian disorders via RNA sequencing,” Nat. Commun., vol. 8, p. 15824, June 2017.

[206] L. Frésard, C. Smail, N. M. Ferraro, N. A. Teran, X. Li, K. S. Smith, D. Bonner, K. D. Kernohan, S. Marwaha, Z. Zappala, B. Balliu, J. R. Davis, B. Liu, C. J. Prybol, J. N. Kohler, D. B. Zastrow, C. M. Reuter, D. G. Fisk, M. E. Grove, J. M. Davidson, T. Hartley, R. Joshi, B. J. Strober, S. Utiramerur, Undiagnosed Diseases Network, Care4Rare Canada Consortium, L. Lind, E. Ingelsson, A. Battle, G. Bejerano, J. A. Bernstein, E. A. Ashley, K. M. Boycott, J. D. Merker, M. T. Wheeler, and S. B. Montgomery, "Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts," Nat. Med., vol. 25, pp. 911–919, June 2019.

[207] S. Maddirevula, H. Kuwahara, N. Ewida, H. E. Shamseldin, N. Patel, F. Alzahrani, T. AlSheddi, E. AlObeid, M. Alenazi, H. S. Alsaif, M. Alqahtani, M. AlAli, H. Al Ali, R. Helaby, N. Ibrahim, F. Abdulwahab, M. Hashem, N. Hanna, D. Monies, N. Derar, A. Alsagheir, A. Alhashem, B. Alsaleem, H. Alhebbi, S. Wali, R. Umarov, X. Gao, and F. S. Alkuraya, "Analysis of transcript-deleterious variants in Mendelian disorders: implications for RNA-based diagnostics," Genome Biol., vol. 21, p. 145, June 2020.