From Data to Diagnosis:

Innovative Bioinformatics Strategies for Diagnosing Rare Genetic Diseases

Wouter Steyaert

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

From Data to Diagnosis:

Innovative Bioinformatics Strategies for Diagnosing Rare Genetic Diseases

Wouter Andre Ragelinus Steyaert

Author: Wouter Andre Ragelinus Steyaert
Title: From Data to Diagnosis: Innovative Bioinformatics Strategies for Diagnosing Rare Genetic Diseases

Radboud Dissertations Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Proefschrift AIO | Guus Gijben
Cover: Proefschrift AIO | Guntra Laivacuma
Printing: DPN Rikken/Pumbo

ISBN: 9789465150031
DOI: 10.54195/9789465150031
Free download at: www.boekenbestellen.nl/radboud-university-press/dissertations

RADBOUD
UNIVERSITY
PRESS

From Data to Diagnosis: Innovative Bioinformatics Strategies for Diagnosing Rare Genetic Diseases

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

donderdag 23 januari 2025
om 14.30 uur precies

door

Wouter Andre Ragelinus Steyaert
geboren op 21 november 1982
te Gent (België)

Promotoren:
Prof. dr. H.G. Brunner
Prof. dr. C.F.H.A. Gilissen
Prof. dr. A. Hoischen

Manuscriptcommissie:
Prof. dr. M.A. Huijnen
Dr. M. Zamani Esteki (Maastricht UMC+)
Prof. dr. N.C. Voermans

From Data to Diagnosis: Innovative Bioinformatics Strategies for Diagnosing Rare Genetic Diseases

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Thursday, January 23, 2025
at 2.30 p.m.

by

Wouter Andre Ragelinus Steyaert
born on November 21, 1982
in Gent (Belgium)

PhD supervisors:
Prof. dr. H.G. Brunner
Prof. dr. C.F.H.A. Gilissen
Prof. dr. A. Hoischen

Manuscript Committee:
Prof. dr. M.A. Huijnen
Dr. M. Zamani Esteki (Maastricht UMC+)
Prof. dr. N.C. Voermans

Table of contents

Chapter 1

General introduction

Rare diseases, though individually rare, collectively affect approximately 6-8% of the global population, representing nearly 400-500 million people worldwide (Nguengang Wakap et al., 2020). A central challenge in rare disease research and diagnostics is to identify the genetic alteration(s) in a patient's genome that explain their condition. Successfully pinpointing the causative variant not only provides treatment or disease management options, but also concludes a costly diagnostic journey. Additionally, for couples at risk of transmitting a genetic disease, this identification allows for the estimation of recurrence risks and the possibility of prenatal screenings and preimplantation diagnostics (Biesecker & Green, 2014; Text box 1).

With current DNA sequencing technologies and data analysis strategies, the disease-causing variant is identified in approximately 20-70% of patients, depending on the disease and the patients' inclusion criteria (Smedley et al., 2021; Turro et al., 2020). Two primary factors contribute to the significant portion of patients who remain undiagnosed: first, despite decades of technological advances, a substantial fraction of genetic variants is still difficult or impossible to detect; second, the functional consequences of many genetic variants remain unknown, complicating clinical interpretation. In this thesis, we address both of these factors. Specifically, we contribute to the identification of previously undetected genetic variants using novel bioinformatics approaches and advanced sequencing technologies. Additionally, we clinically reassessed a large number of genetic variants, utilizing the latest knowledge databases.

The Identification of Genetic Variants

To identify genetic variants within a human genome, the order of nucleotides needs to be determined. This process is called DNA sequencing and was, until recently, only feasible for relatively short DNA fragments (typically 100 or 150 base pairs). Once the sequences of all of these short DNA fragments are determined, they can be computationally aligned onto a reference genome. In this step, the genetic origin of these DNA sequences is determined. By mapping all of the sequence reads that originate from a complete human genome onto the reference sequence, the DNA sequence of the individual that is sequenced becomes apparent. In the next step, the DNA sequence is compared to the reference sequence in a process called variant calling (Figure 1a).

The two most used DNA sequencing techniques in clinical settings are exome sequencing (ES) and (short-read) genome sequencing (GS; Figure 2; H. Lee et al., 2014; Stranneheim et al., 2021).

The importance of making a genetic diagnosis.

- End of diagnostic odyssey: Obtaining a genetic diagnosis can conclude a prolonged, expensive, and potentially invasive diagnostic process. This resolution not only eliminates the need for further unnecessary investigations but also confers substantial psychological benefits. By removing the uncertainty surrounding the patient's condition, it significantly enhances quality of life and provides a clearer path for subsequent medical management and care (Carmichael et al., 2015).
- Psychosocial support: For many patients and families, having a clear genetic diagnosis can provide relief by ending the uncertainty of not knowing the cause of symptoms. It can also connect them with support groups and resources specific to their genetic condition (Boycott et al., 2017; Robin, 2006).
- Prognosis: Knowing the genetic basis of a disease can provide critical information about the expected course of the condition. This can help in planning long-term care and managing symptoms more effectively (Robin, 2006).
- Family planning: The establishment of a genetic diagnosis allows for the determination of recurrence risk when the parents of a child with a rare condition are considering having more children. Additionally, reproductive options such as preimplantation genetic diagnosis can be offered (Aartsma-Rus et al., 2016; Robin, 2006).
- Treatment options: For an increasing number of rare genetic disorders, treatment is possible. In many cases, it is crucial to know the precise genetic cause (Tambuyzer et al., 2020).
- Scientific understanding: Genetic diagnoses contribute to the broader understanding of genetic disorders, facilitating research into new treatments and potential cures. This research can lead to advancements in medical science that benefit not only the individual but also society as a whole (Boycott et al., 2017).

Text box 1: The importance of making a genetic diagnosis

Exome sequencing

Exome sequencing (ES) is a genomic technique that focuses on sequencing the protein-coding regions of the genome. Since only about 1% of a complete human genome translates into protein and since most fully-penetrant disease-explanatory genetic variants are present within these regions or very close by, ES is a highly cost-effective diagnostic sequencing strategy (Teer & Mullikin, 2010). To capture an individual's exome, the DNA is first fragmented into smaller pieces. Then, RNA baits, which are complementary to the protein-coding regions of the genome, are used to selectively capture these regions of interest. The DNA fragments that hybridize onto the RNA baits are retained, while non-target DNA is washed away. Finally, the captured DNA fragments are sequenced to reveal the genomic sequence within the protein-coding regions (Warr et al., 2015).

While ES has been instrumental in discovering numerous gene-disease associations, the technique also has several limitations. First, although most large-effect genetic variants reside within the coding portions of the genome, a growing number of phenotypes is explained by variants in the non-coding space of the human genome, often alterations in promoters, enhancers, insulators or other regulatory elements (French & Edwards, 2020). To genetically diagnose patients with disease-causing variants in these regions, ES falls short (Figure 2). Second, exome-enrichment kits do not capture all coding regions effectively. Depending on the enrichment kit that is used, roughly 2 to 10% of coding sequences in the human genome is not or very poorly enriched (Lelieveld et al., 2015; Yaldiz et al., 2023). Third, the process of exome capture itself introduces bias in the sequencing experiment. Two alleles from a certain autosomal genomic region may hybridize with different efficiencies when genetic variation is present on one the alleles. This might ultimately result in the variant being undetected when the fraction of sequencing reads that support the variant allele is too small (Meynert et al., 2014). Furthermore, hybridization efficiencies of the various baits that together capture the entire exome are highly variable, causing the sequence coverage to be uneven. As a consequence, variant discovery may be challenged, especially for copy number variants (CNVs) since these variants are typically identified by the comparison and evaluation of coverage profiles (cf. Genome sequencing).

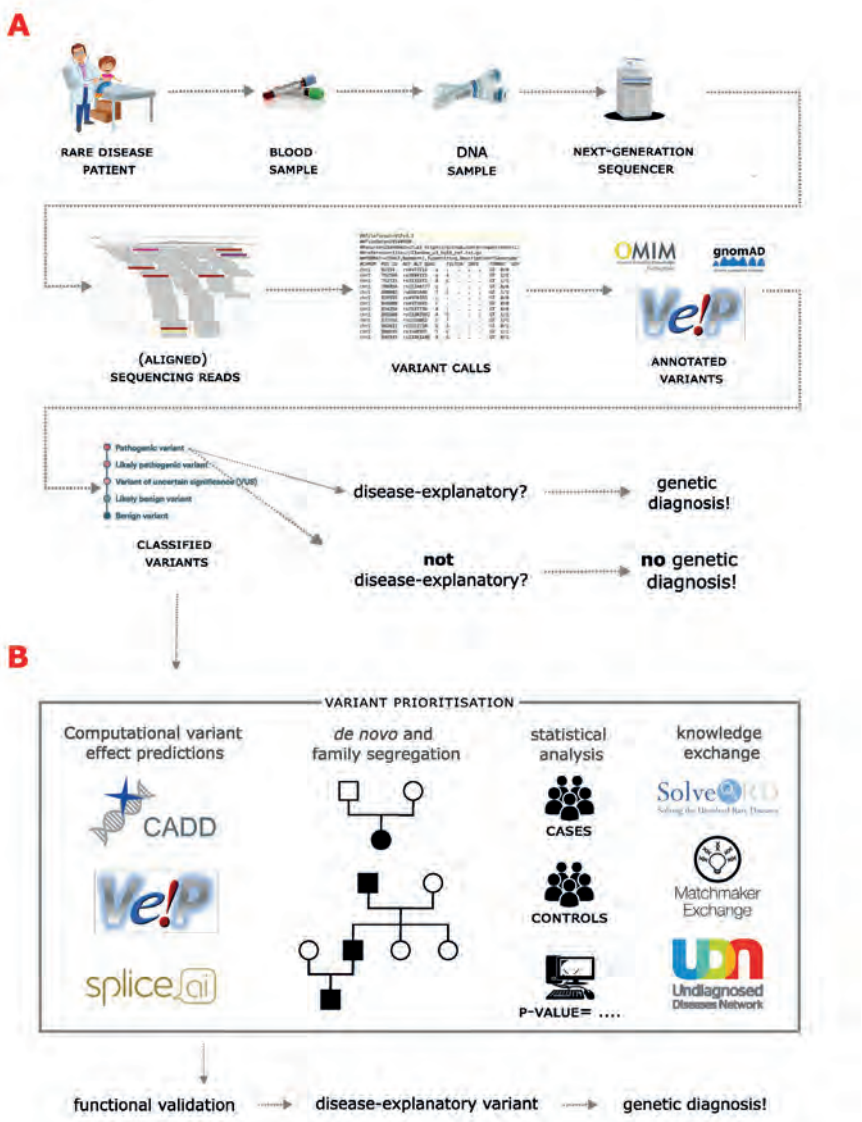


Figure 1: From rare disease patient to genetic diagnosis. Panel A: After blood is drawn from a rare disease patient, DNA is extracted from white blood cells. Next, the DNA is prepared for sequencing with an NGS instrument. The resulting NGS reads are aligned onto a reference genome allowing for variant calling. In order to be able to classify the genetic variants, variant annotation needs to be done. In this process biological and genetic information is added to the variant file to facilitate variant classification. If a pathogenic variant is identified that is fully disease-explanatory, the patient receives a genetic diagnosis. On the other hand, if no disease-explanatory genetic variant is identified, the patient remains undiagnosed. Panel B: Research laboratories might prioritize genetic variants and continue with functional work to confirm or falsify the disease-relation of a handful of top-candidate disease-causing variants. These efforts, often conducted jointly between different laboratories, to enhance study power, contribute to the expanding repertoire of known disease genes.

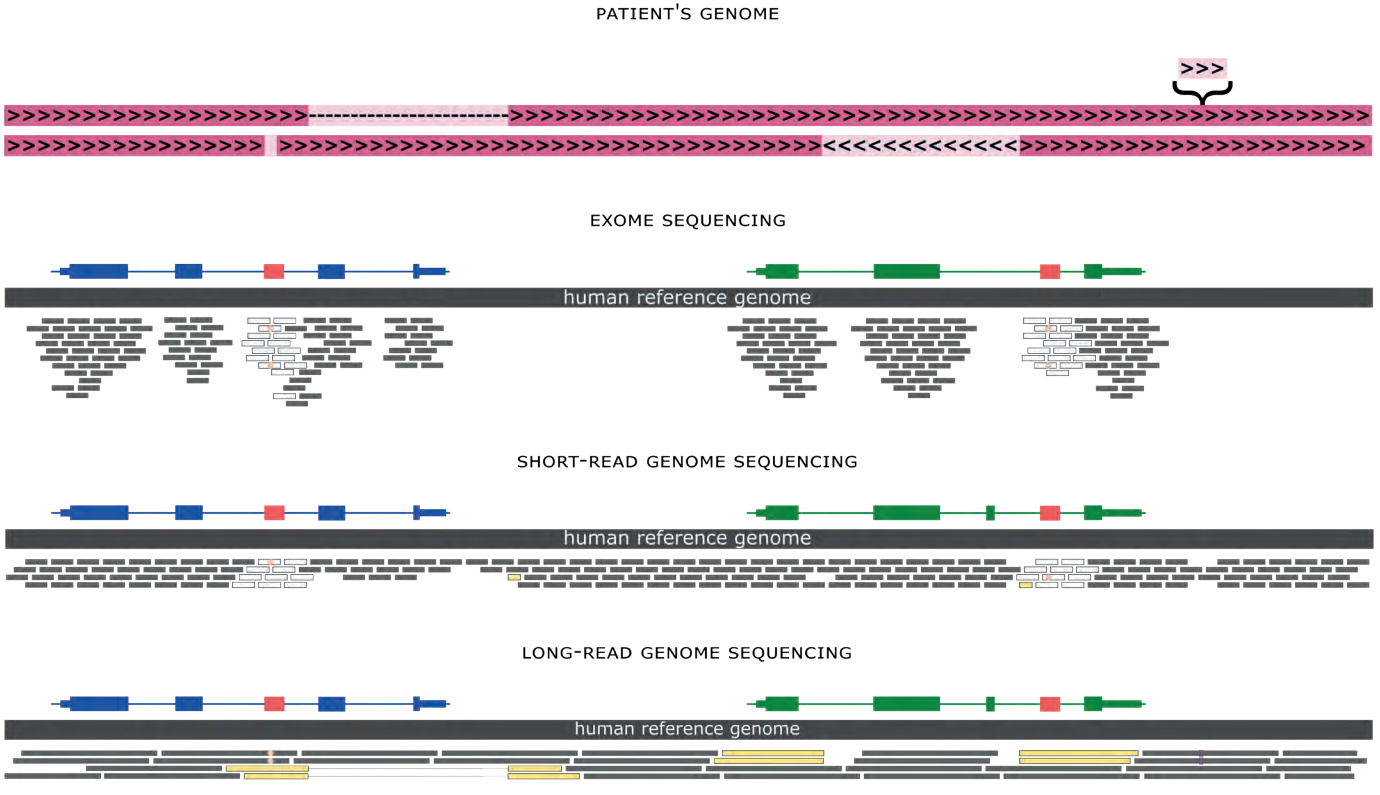

PATIENT'S GENOME
>>>
EXOME SEQUENCING
human reference genome
SHORT-READ GENOME SEQUENCING
human reference genome
LONG-READ GENOME SEQUENCING
human reference genome

Figure 2: Schematic depiction of the strengths of different sequencing technologies to comprehensively identify genome variation. The two alleles of the sequenced genome are illustrated in purple at the top of the figure, the arrows indicate the orientation of the DNA sequence relative to the reference genome. The patient whose DNA is sequenced in the illustration has four heterozygous genetic variants: a multi-exon deletion, an inversion, an insertion and a single nucleotide variant in a paralogous exon. All of these variants are displayed with a pink background in the DNA sequence of the patient. The human reference genome is illustrated with a dark grey rectangle in each of the three subfigures. Above this reference sequence, the location and structure of two multi-exonic genes is shown. A single exon of both of these genes is identical in sequence, indicated in red. Only in the long-read sequencing experiment the reads that align onto these red exons are coloured grey, indicating that they align uniquely onto the reference genome. As a result, the single nucleotide variant within the red exon of the blue coloured gene can be detected. In the exome experiment, only the deletion can possibly be detected, based on the reduced number of reads at that particular locus, although the large variability in read depth which is intrinsic to ES might hamper variant identification. The inversion, which also affects a coding exon is not identifiable in ES because it is a copy-neutral genetic event and its breakpoints are located in the non-coding space of the genome. In short-read genome sequencing, the inversion could theoretically be detected, but typically, only very few reads align onto the breakpoint regions. Even dedicated algorithms most often fail to identify these events. Also the insertion which is too small to be detectable with read depth methods and too large to be encapsulated by short sequence reads will most often remain unidentified with short-read sequencing technologies. The length of the reads in LRS ensure that the reads can be aligned accurately in a split fashion around the deletion and inversion, revealing the event. To indicate this, reads that are clipped, thus only partly aligned onto a specific locus in the reference genome, are coloured yellow. Overall, LRS offers the best guarantee to identify all difficult variant types.

Genome sequencing

Sequencing complete genomes has, until now, been conducted on a much smaller scale as compared to ES. However, the drastically fallen sequencing costs in parallel with technological advancements in computer hardware and software (such as the sharp decline in cost for storing information on disk, ameliorated capabilities of computer processing units and more optimized data compression techniques) are leading to an increasing number of human genomes being sequenced in research and clinical care (Manolio et al., 2019; H. C. Martin et al., 2014; Riess et al., 2024; Scocchia et al., 2019; Thiffault et al., 2019).

Because there is no DNA enrichment procedure in GS, the technique suffers less from allele imbalance and uneven sequence coverage which results in a higher quality variant list – even with a lower sequence depth (Belkadi et al., 2015; Yaldiz et al., 2023). In addition, sequencing complete genomes enables the identification of genetic variants in deep intronic and intergenic regions. The interpretation of these variants is still in its infancy, but it already has significantly expanded our understanding of the human genome and it also already resulted in a substantial number of exome-negative rare disease patients to receive a genetic diagnosis (Alfares et al., 2018). Despite the fact that GS is an excellent way for studying genetic variation in the entire human genome, not all genetic variants can be identified. Several technical and biological phenomena are responsible for this (Figure 2).

Sequence homology

As a result of small and large-scale duplication events in the evolution of the human species, around 5% of the human genome contains non-unique sequences (Mandelker et al., 2016). Since sequence reads are much smaller as compared to these replicated chunks of DNA, read alignment in a significant part of the human genome is ambiguous. Technically, when a short sequence read aligns equally well to multiple locations in the reference genome, it is assigned a mapping quality of zero. Such sequence reads are typically ignored by the variant calling algorithms and, as a result, the vast majority of genetic variants within these regions remains unidentified. Several known disease genes however partly or completely reside within these duplicated DNA regions and thus their incomplete analysis is probably one of the reasons why diagnostic success rates are relatively modest nowadays (Ebbert et al., 2019; Mandelker et al., 2016).

Structural variants

Through exome and genome sequencing, it is relatively straightforward to reliably identify single nucleotide variants (SNVs), multi-nucleotide variants (MNVs) and

small insertions and deletions (small indels) in the unique parts of the human genome (Poplin, Chang, et al., 2018; Poplin et al., 2018; Supernat et al., 2018). Both in research and clinical practice there is no more need for orthogonally validating high-quality variant calls (Beck et al., 2016). Genetic variants that alter the structure of the DNA with more that 50 base pairs, typically defined as structural variants (SVs), are much harder to identify. Given the larger size of these variants, next-generation sequencing reads cannot fully encompass them while at the same time align with proper mapping qualities. Consequently, the detection of such structural DNA changes necessitates reliance on alternative, less precise signals (for example aberrations of coverage profiles and the inspection of partly aligned reads or read pairs). Recent advances in bioinformatics software however enabled the identification of various types of SVs from genome sequencing data with relative accuracy (cf. Towards improved variant identification).

Towards improved variant identification

Both the fields of bioinformatics and sequencing technologies evolve rapidly. In recent years, several efforts have been undertaken to further improve and expand variant calling from exome and genome sequencing data.

Bioinformatics software

Bioinformatics tools for the analysis of next-generation sequencing data continue to improve. While single nucleotide variants (SNVs), multi-nucleotide variants (MNVs), and small indels are easy to identify with high accuracy, more complex variant types remain challenging to detect. Over the past few years, novel algorithms and software tools have been developed that have demonstrated that these difficult-to-detect variant types, such as *de novo* mutations, alterations of the mitochondrial DNA, and structural variants, can now be identified from exome and genome sequencing data with moderate accuracy.

Improved *de novo* mutation discovery

De novo mutations (DNMs) are genetic changes that are present in the offspring but absent from the parents. These new mutations of which each and every individual has around 75 can arise in a germ cell of one of the parents or during early embryogenesis (Conrad et al., 2011). Several Mendelian diseases are (predominantly) caused by DNMs and thus its accurate identification is of utmost importance in rare disease genetics (Veltman & Brunner, 2012). Identifying DNMs from sequencing data is a complex challenge that involves accurately genotyping a specific genetic site across three distinct samples. Commonly used tools for this task include GATK, DeNovoGear, and Platypus, but they all generate large numbers

of potential *de novo* variants requiring researchers to further refine these lists (Poplin et al., 2018; Ramu et al., 2013; Rimmer et al., 2014). To accommodate for this, novel *de novo* variant callers have been developed that take advantage of the latest advances in machine learning. These recent methods, such as DeNovoCNN and DeepTrio, demonstrate superior performance as compared to other methods (Hu et al., 2022; Khazeeva et al., 2022).

Mitochondrial DNA analysis

In the currently used exome enrichment kits there are no probes present to capture the mitochondrial genome. Because most genetic laboratories restrict variant discovery to regions that are targeted by the experiment, the mitochondrial genome is typically not analysed in data from ES experiments (Samuels et al., 2013). Nevertheless, based on epidemiological studies, 1/5,000 individuals have or are at risk for developing mitochondrial DNA (mtDNA) disease (Schaefer et al., 2008). Due to the high abundance of mtDNA copies in the cell, most often the mtDNA is decently covered in ES data as a result of off-target reads. Several studies have shown the value of ES for the analysis of mtDNA by comparing the resulting variants with a golden standard such as Sanger or targeted next-generation sequencing (Griffin et al., 2014; Wagner et al., 2019). In addition to that, dedicated software tools were developed to improve variant identification and estimation of heteroplasmy levels using exome data (Calabrese, Simone, Diroma, Santorsola, Guttà, et al., 2014; Picardi & Pesole, 2012).

Structural Variant Calling

In recent years, significant progress has been made in accurately identifying structural variations (SVs) in sequencing data. Dedicated tools have been developed for various subtypes, which have already proven their utility in medical genetics.

Copy number variants

Copy number variants (CNVs) are typically defined as deletions and duplications of DNA segments larger than 1 kb and up to several Mb (Freeman et al., 2006). The first bioinformatic tools to identify CNVs in ES and GS data were published roughly ten years ago (Boeva et al., 2012; Krumm et al., 2012; Plagnol et al., 2012; Roller et al., 2016). Most of these tools make use of the fact that the number of sequencing reads that align onto a certain genomic site is a proxy for the number of alleles in the sample this is sequenced. Since the total number of sequencing reads is variable among different samples, the sequence coverage needs to be normalized first. Subsequently, the coverage profile of the sample at study is statistically compared with a group of technically comparable samples (a reference cohort).

Although algorithms exist for both ES and GS, the significantly better coverage evenness in GS, as well as the continuity of the data points, make GS much more suitable for identifying CNVs. Additionally, the availability of coverage data for intronic and intergenic regions facilitates the estimation or determination of breakpoints, which may be essential for clinical interpretation (Hehir-Kwa et al., 2015). It is important to note that when only a small number of data points are responsible for an aberrant coverage profile, the uncertainty associated with the variant call is large. Particularly, small heterozygous CNVs (the order of a single exon), are the most challenging to detect in coverage profiles that intrinsically exhibit quite some variability. In contrast, very large CNVs, spanning several mega bases, can often be identified with great accuracy. By comparing different CNV callers and by orthogonally validating the calls, researchers gained insights into the capabilities of the different published tools (Gordeeva et al., 2021; Tan et al., 2014; Yao et al., 2017). By adopting the lessons learned from these comparison studies and applying stringent filter protocols to remove false positive CNV calls, multiple studies demonstrated an increased diagnostic yield by considering CNVs in clinical exome and genome cohorts (Gambin et al., 2017; Gross et al., 2019; Meynert et al., 2014; Retterer et al., 2015; Zhai et al., 2021).

Short tandem repeat expansions

Short tandem repeats are repeating motifs of 1-6 base pairs comprising roughly 3% of the human genome (Lander et al., 2001). The accurate estimation of the repeat length from ES or GS data is challenging because the repeat length exceeds the length of a typical sequence read. However, the expansion of part of these STRs lead to human, mostly neurodegenerative, disease (La Spada & Taylor, 2010). Recent bioinformatic developments however have successfully enabled the identification and prioritization of these pathogenic STR expansions (Casse et al., 2023; Dolzhenko et al., 2019, 2020; van der Sanden et al., 2021). Basically, these tools take advantage of the fact that the number of reads that align onto the repeat locus is a proxy for the length of the locus. Statistical modelling allows for the discrimination between different autosomal alleles.

Mobile element insertions

Mobile elements are DNA sequences that have the ability, if active, to insert new copies elsewhere in the genome via an RNA intermediate. Mobile element insertions (MEIs) can lead to disease if these sequences are for example inserted in the coding sequence of disease-relevant genes. Currently, more than 150 MEIs have been linked to human disease (Hancks & Kazazian, 2016; Qian et al., 2017). Because MEIs are SVs, dedicated bioinformatic tools are needed for their identification in ES

and GS data (Demidov et al., 2021; Gardner, Lam, Harris, Chuang, Scott, Stephen Pittard, et al., 2017; Rishishwar et al., 2017; Thung et al., 2014; Torene et al., 2020; Vendrell-Mir et al., 2019; Wijngaard et al., 2024). Recently it has been shown that their application on clinical cohorts provides an additional diagnostic yield (Gardner et al., 2021).

Other structural variants

Heterozygous deletions, insertions and duplications of intermediate size (between 50 and 1000 base pairs) are highly challenging to identify from sequencing data with read depth methods (Figure 2). Also copy-neutral events such as inversions and translocations do not result in an altered coverage profile. For the detection of these variants, methods were developed that exploit other properties of sequencing data. Some of these methods for example make use of the fact that an SV might result in an altered distance between the mates in a pair of reads (the insert size) or that the orientation of the mates in paired-end sequencing is altered. Split-read methods on the other hand aim to identify (breakpoints of) SVs by investigating incompletely aligned read pairs, and assembly-based methods try to *de novo* assemble reads resulting in longer contigs that encapsulate putative SVs (Kosugi et al., 2019; Tattini et al., 2015). All of these different methods and tools produce large lists of putative SVs. Although it is known that a large fraction of these variant calls are false positives, it has been demonstrated that the application of these tools on clinical ES and GS data combined with stringent filter protocols results in an additional diagnostic yield (Gardner et al., 2021; Palmer et al., 2021; Shashi et al., 2019).

Novel sequencing technologies

Despite all of the improvements in bioinformatics software, a substantial fraction of genetic variants cannot be identified from ES or GS because there is no trace present in the data. Although sequencing short DNA fragments can nowadays be done at relatively low cost, it is clear that the short length of these sequence fragments is a limiting factor for the identification of a large fraction of genetic variants such as SVs and genetic variants in duplicated genomic regions (Huddleston et al. 2017). In the past decade, a couple of companies developed and optimized technologies to sequence much larger segments of DNA. The two major players in the field of long-read sequencing (LRS) are Pacific Biosciences of California and Oxford Nanopore. With their technologies it is possible to generate accurate sequence reads of multiple tens of kilobases (Figure 2). In addition, there is no need for PCR amplification in the sequencing process. This step, which is necessary in short-read next-generation sequencing technologies, introduces biases as it cannot be done

perfectly uniform (Schadt et al., 2010). A major drawback of LRS technologies is the cost which still heavily exceeds that of a regular short-read sequencing experiment. For that reason, these novel sequencing experiments are yet not affordable for health-care systems (Branton et al., 2008; Deamer et al., 2016; Wenger et al., 2019). Nevertheless, more and more researchers report their success stories in pinpointing the disease-causative genetic variation by the use of LRS (Grosz et al., 2022; Ishiura et al., 2018; Merker et al., 2018; Sone et al., 2019).

The clinical interpretation of genetic variants

Every single individual has a genome that differs from the reference sequence in about 4 to 5 million sites (Auton et al., 2015). To be able to find the alteration that is causative for disease in a given patient it is essential to first associate genetic and biological information to each variant. This process is called variant annotation and allows for prioritizing genetic variants that are most likely involved in the patient's phenotype (Figure1).

Variant annotation

Among the most important annotations are known disease-gene annotations, known disease-variant annotations, population variant allele frequencies, the functional consequence of the variant on transcript level and computational variant effect predictions.

Known disease-causing variants and genes

In an automated variant interpretation workflow, one of the first steps is to check for variants with known pathogenicity. Two important databases that contain interpreted genetic variants are ClinVar and HGMD. ClinVar is a freely accessible and monthly updated inventory of clinically interpreted genetic variants (Landrum et al., 2016, 2020). When a specific laboratory has demonstrated the causative relationship between a genetic alteration and a phenotype it can be shared with the worldwide genetics community via ClinVar. As a result, other laboratories that identify the same genetic variant in a patient with the same phenotype can use this information to diagnose their patient. HGMD is a similar more extensive and better curated but commercial initiative (Stenson et al., 2020).

Patients are not only diagnosed based on previously known pathogenic variants. All potentially disease-causing variants in a complete set of relevant genes are considered. These sets of known disease genes are called gene panels and contain

genes known to cause one particular disease or phenotype (Bean et al., 2020). Some gene panels only contain a handful of genes while the largest panels contain multiple hundreds or even thousands of genes. Only when the clinical presentation is unspecific, the complete Mendeliome (the collection of all disease genes leading to a Mendelian or monogenic disease) needs to be considered. An example of a publicly available collection of curated disease gene panels is PanelApp from the Genomics England consortium (A. R. Martin et al., 2019).

To identify the disease-explanatory variant among all genetic variants in the applied gene panel, the variants are filtered on population allele frequency and their consequence on transcript level. To ensure that the subsequent variant interpretation analysis is as correct and uniform as possible, the American College of Medical Genetics and Genomics (ACMG) together with the Association for Molecular Pathology (AMP) and the College of American Pathologists developed standardized guidelines for the interpretation of sequence variants in known Mendelian disease genes (Richards et al., 2015). Their protocol uses population data, computational data, functional data and segregation data to classify variants into one of 5 categories: benign, likely benign, variant of unknow significance, likely pathogenic and pathogenic. The strongest evidence for variant pathogenicity are "null variants" in disease-relevant gene transcripts where loss-of-function is a known disease mechanism.

Population variant allele frequencies

The diseases for which genetic diagnostics is offered nowadays are typically monogenic. These diseases are rare and thus cannot be caused by common genetic variation. By using the frequencies of variant alleles in the general population, all common variants can be discarded in the search for the causal variant. The most important resource for allelic frequencies in the general population that is used by researchers and clinical investigators worldwide is the Genome Aggregation Database (gnomAD). Currently it includes genetic variants from more than 200,000 individuals from different ancestries (Karczewski, Francioli, Tiao, Cummings, Consortium, et al., 2020). In general, the vast majority of all variants in a human genome are common (circa 98% have allelic frequencies above 0.5%; (Auton et al. 2015)) and thus can be removed for further consideration in the search for the disease-causing variant.

Variant consequences on the transcript

Although rare genetic diseases can be caused by variants that do not alter the sequence of amino acids in a protein, both for ES and GS non-synonymous variants

(predicted loss-of-function and missense variants) will be investigated first since these variants have larger effects on the protein in general.

Predicted loss-of-function variants

Genetic variants that are predicted to disrupt gene function are called predicted loss-of-function (pLoF) variants. These variants can be single nucleotide substitutions that give rise to a premature stop codon (nonsense variants), small insertions or deletions disrupting the triplet reading frame of a transcript (frameshift variants), SNVs at positions ±1 or 2 at the splice sites (canonical splice site variants), substitutions leading to a loss of the initiation codon (start loss variants) or single or multiple exon deletions. As compared to genetic variants with other predicted consequences on the transcript and/or protein, the clinical interpretation of pLoF variants in well-established disease genes is relatively straightforward. The reason for this is that most of these variants have the same effect on the transcript: nonsense-mediated mRNA decay which ultimately leads to the absence of gene product. However, nonsense and frameshift variants in the last exon of a transcript do not lead to nonsense-mediated decay (Rivas et al., 2015) and may give rise to functionally different proteins, and thereby to different genetic disorders than for genuine pLoF variants. The clinical interpretation of canonical splice site variants also poses challenges. Although the ±1 or 2 splice donor and acceptor sites are ultra-conserved, canonical splice site variants do not automatically lead to a loss of the protein product. Such variants might for example cause an in-frame exon skipping without having an effect on protein function, or the splicing machinery might make use of cryptic splice sites to allow for splicing. Similarly, genetic alterations of the initiation codon do not always result in loss-of-function. Alternative ATG or non-ATG codons down- or upstream of the original start codon may initiate transcription (Bazykin & Kochetov, 2011). In general, a substantial portion of pLoF variants lead to functional protein products (Guo et al., 2013). For that reason, experimental assays that assess the impact of pLoF variants at the mRNA level can be useful in variant interpretation.

Missense variants

When an SNV or MNV results at the protein level in the substitution of an amino acid by another amino acid the variant is called a missense variant. In general, the clinical interpretation of these variants is highly challenging since their effect on transcript and protein level is difficult to predict. For that reason, the majority of rare missense variants are considered as variants of unknown significance (VUS). There are however some specific criteria for missense variants, according to the ACMG, which are considered as strong evidence for pathogenicity (Richards et

al., 2015). First, if the genetic variant leads to the same amino acid change as a known pathogenic genetic alteration. Second, if the missense variant has occurred *de novo*. Third, if the prevalence of the variant allele in a cohort of patients with the same phenotype is significantly increased as compared to a group of control individuals. Fourth, if functional studies demonstrate the damaging effect of the variant on transcript and/or protein level. Only if 2 of these criteria are true or if 1 of them is true combined with multiple moderate or supportive lines of evidence for pathogenicity the variant can be classified as pathogenic.

Computational variant effect predictions

The clinical interpretation of DNA variants heavily depends on knowledge about their effect on transcript and protein level. Because wet-lab assays are labour intensive, expensive and cumbersome, in the past decades, several computational tools were developed to predict the effect of genetic variants on molecular and human phenotypes. Most of these tools predict the deleteriousness of missense variants on protein level by feeding a machine-learning algorithm with information on evolutionary conservation, physico-chemical amino acid properties, population data, structural protein models for known disease causing and benign genetic variants. SIFT and PolyPhen-2 were among the first tools that were broadly used by the clinical genetics community (Ng & Henikoff, 2001; Ramensky et al., 2002). As more and more sequencing data and clinically interpreted variants became available, novel algorithms were developed, also making use of the latest advances in predictive modelling theory (Brandes et al., 2023; Frazer et al., 2021; Rentzsch et al., 2019). Although the accuracy of these tools substantially increased over the years, none have a 100% sensitivity or specificity. As a result, their outcomes are only used as a supportive argument in a diagnostic variant interpretation analysis (Richards et al., 2015). In research settings however, these tools can be highly valuable, as they allow for a meaningful variant prioritization before attempting laborious functional validation experiments (Figure 1b).

Variants of unknown significance

Due to our incomplete understanding of the molecular and phenotypic effect of genetic variation, the majority of variants are of unknown clinical significance, partly explaining the large number of genetically undiagnosed rare disease patients. The disease-explanatory genetic variant might be present in a known disease gene, but it might also reside in a gene or regulatory element which is not yet recognized as being involved in human disease. For either of these scenario's functional work is typically needed to prove variant pathogenicity. Although validated functional assays exist for many well-established disease genes, it is, especially in routine

diagnostics, infeasible to perform these assays for all rare variants in undiagnosed patients. In several laboratories however the genomes of genetically undiagnosed patients are further explored in a research setting. Attention and resources are focused on a handful of top disease-causing candidate variants by first prioritizing the genetic variants which are most likely involved in the patient's phenotype. These prioritization methods comprise rankings that result from bioinformatic tools (cf. Computational variant effect predictions), statistical analysis and family inheritance (Figure 1b; Eilbeck et al., 2017).

De novo variation and family inheritance

De novo mutations (DNMs) are, on average, more likely to be deleterious as compared to more common genetic variants (Veltman & Brunner, 2012). This is also apparent from the many genetic diseases that are (too a large degree) caused by *de novo* variation. For that reason, and given the fact that the number of new mutations in a genome is small, DNMs in yet undiagnosed rare disease patients (with healthy parents) are often of particular interest to rare disease researchers. When multiple independent patients with the same phenotype and with DNMs in the same gene are found this is considered strong evidence for disease causality. Further functional studies are then often required to provide additional support of pathogenicity or to decipher the biological pathways involved in disease. This strategy has been successful for a large number of diseases (Chong et al., 2016; Muona et al., 2015; Polla et al., 2021; Yuan et al., 2014).

For diseases that are inherited in large families, researchers used to do linkage studies to delineate one or a couple of regions in which genetic variants co-segregate with disease. The genes within these regions are then fully explored to find the cause of disease for the given family. Nowadays, sequencing the DNA of all family members in a large family is feasible and thus the need for dedicated linkage studies disappears. However, family segregation is still a very valuable and powerful prioritization strategy which each year aids in the identification of novel disease genes (Baetens et al., 2017; Karolak et al., 2019; Wu et al., 2015).

Statistical analysis

For some phenotypes, single genetic centers or research consortia are able to collect data for a large number of disease patients and control individuals. In that case, a statistical variant burden analysis can be conducted. The main purpose of these analysis is to identify genes in which rare variants are significantly more (potentially deleterious) or less (potentially protective) abundant in the case group as compared to the control group. Such analysis have been conducted by

several research groups and pinpoint genes which are highly likely involved in the respective phenotype (Lelieveld, Reijnders, Pfundt, Yntema, Kamsteeg, de Vries, et al., 2016). Another type of variant burden analysis is a *de novo* variant enrichment analysis. These studies have shown to be very successful in identifying genes implicated in intellectual disability and autism (Kaplanis, Samocha, Wiel, Zhang, Study, et al., 2020). The idea behind such analysis is that DNMs in a cohort of patients with the same disease might cluster in disease relevant genes or regions as compared to what is expected based on known *de novo* mutation rates.

Knowledge exchange in and between centers

Despite variant and gene prioritization strategies being relatively successful, a very large number of rare disease patients remain genetically undiagnosed. To further accelerate the identification of disease-explanatory variants and disease genes, in the last decade, several initiatives were founded to facilitate the transfer of data and knowledge between genetic laboratories within or across countries. Among the most important of such initiatives are Matchmaker Exchange (MME), Undiagnosed disease network (UDN) and Solve-RD (Philippakis et al., 2015; Ramoni et al., 2017; Zurek et al., 2021).

Aims and scope of this thesis

The aim of this thesis is to address some of the current shortcomings in variant identification and interpretation and to genetically diagnose previously undiagnosed rare disease patients.

In **chapter 2** we investigated the possibility of identifying genetic variation within the paralogous regions of the genome using exome sequencing data. We developed an algorithm called Chameleolyser in which sequencing reads from all paralogs are jointly analysed. We applied this novel method to a cohort of 41,755 exome datasets and obtained a genetic diagnosis in 25 previously undiagnosed patients.

Chapter 3 describes the approach and result of a large-scale reanalysis of existing exome and genome datasets of the Solve-RD consortium. We collected sequencing data from more than 6,000 previously undiagnosed rare disease families together with standardized phenotype and family information. In total, 37 institutions across 13 countries, affiliated to one of 4 European Reference Networks (ERN-ITHACA, ERN-RND, ERN-NMD and ERN-GENTURIS), contributed to the final dataset. By conducting

an in-depth variant analysis and interpretation we genetically diagnosed more than 500 previously undiagnosed rare disease patients.

In **chapter 4** we applied long-read sequencing to a unique cohort of 293 individuals from 114 genetically undiagnosed RD families selected by European Rare Disease Network (ERN) experts. We wanted to evaluate the extent to which rare diseases are caused by structural variants (SVs), small variants/single nucleotide variants (SNVs), or short tandem repeat (STR) expansions undetectable by standard technologies, including short-read exome and/or genome sequencing. In total, we identified and validated disease explanatory genetic variants in 13 families and in an additional four families we identified a candidate disease-causing variant.

In **chapter 5**, I discuss the relevance of these findings and their implications for future research into the causes of human genetic disease.

Chapter 2

Systematic analysis of paralogous regions in 41,755 exomes uncovers clinically relevant variation

Wouter Steyaert, Lonneke Haer-Wigman, Rolph Pfundt, Debby Hellebrekers, Marloes Steehouwer, Juliet Hampstead, Elke de Boer, Alexander Stegmann, Helger Yntema, Erik-Jan Kamsteeg, Han Brunner, Alexander Hoischen*, Christian Gilissen*

* These authors jointly supervised the work

Nature Communications, 2023

https://doi.org/10.1038/s41467-023-42531-9

Abstract

The short lengths of short-read sequencing reads challenge the analysis of paralogous genomic regions in exome and genome sequencing data. Most genetic variants within these homologous regions therefore remain unidentified in standard analyses. Here, we present a method (Chameleolyser) that accurately identifies single nucleotide variants and small indels, copy number variants and ectopic gene conversion events in duplicated genomic regions using whole-exome sequencing data. Application to a cohort of 41,755 exome samples yields 20,432 rare homozygous deletions and 2,529,791 rare single nucleotide variants and small indels, of which we show that 338,084 are due to gene conversion events. None of the single nucleotide variants and small indels are detectable using regular analysis techniques. Validation by high-fidelity long-read sequencing in 20 samples confirms >88% of called variants. Focusing on variation in known disease genes leads to a direct molecular diagnosis in 25 previously undiagnosed patients. Our method can readily be applied to existing exome data.

Introduction

Over 1,700 human protein coding genes partly or completely share a very high sequence identity with other genomic regions (Mandelker et al., 2016). These paralogous regions originate from small- or large-scale duplication events or retro-transpositions in the evolution of the human species. The sequence and function of these duplicated genomic regions typically diverge over evolutionary time by the accumulation of mutations at different rates. One of the copies might lose its function and evolve to a non-coding paralog (a pseudogene) or to a coding paralog with a different function (Michael & S., 2000; Walsh, 1995).

A well-known genetic mechanism that is relevant when studying paralogous regions is the ectopic gene conversion. A non-allelic or ectopic gene conversion is an event where a sequence is copied from a specific genomic region (the donor region) to a distant region (the acceptor region). When the donor and acceptor sequence differ, this introduces new genetic variation into the acceptor site (Dumont, 2015; Santoyo & Romero, 2005). Ectopic gene conversions occur in at least 1% of human genes associated with inherited disease (Casola et al., 2012). In several of these genes such as *STRC, OTOA* and *SMN1* gene conversions have previously been identified as a cause of genetic disease (Campbell et al., 1997; Laurent et al., 2021; Shearer et al., 2014).

Despite their clinical relevance, gene conversions remain unidentified in the analysis of short-read data such as whole-exome sequencing (WES) and whole-genome sequencing (WGS) data. Indeed, in case of an ectopic gene conversion, the sequencing reads that originate from the acceptor site will align onto the reference sequence corresponding to the donor site. As a result, no reads will be aligned to the acceptor site and single nucleotide variants and small insertions and deletions (SNVs/Indels) that are introduced by means of the gene conversion remain unidentified (Figure 1e). Copy number variant (CNV) callers however will typically identify such events as deletions despite the fact that no deletion is present in the patients DNA (from here the term 'deletion' refers to genetic events with the size of a single or multiple exons).

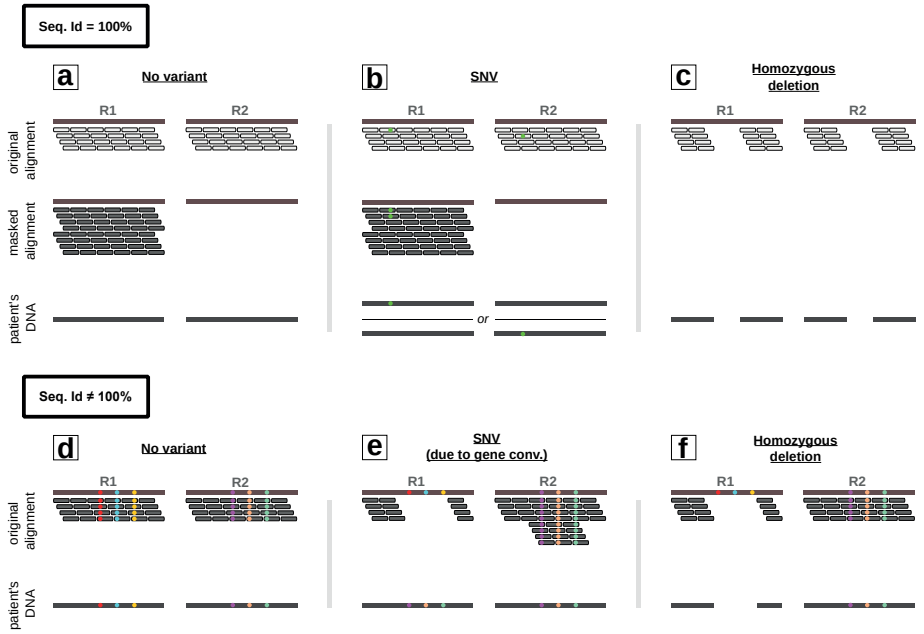


Figure 1: Schematic overview of genetic events that are identified by Chameleolyser. Regions R1 and R2 are 2 regions with a very high sequence identity. In panels a, b and c these 2 regions are completely identical (Seq. Id = 100%). As a consequence, reads that align onto these regions will have mapping qualities of 0 (when no masking is applied). To indicate this, reads are displayed white. Within Chameleolyser, reads are extracted and re-aligned onto a reference sequence in which R2 is masked. As a result, reads align uniquely onto R1 and will have mapping scores different from 0. This is indicated by representing them in grey. By applying a sensitive variant calling onto this masked alignment, Chameleolyser is able to identify single nucleotide variants and small indels (SNVs/Indels; green bullet in panel b). Nevertheless, the exact position of the variant remains ambiguous, hence we named them VAPs (variant with ambiguous position). In case R1 and R2 are identical in sequence, Chameleolyser limits the identification of homozygous deletions to events in which both R1 and R2 are deleted (panel c). Panels d, e and f illustrate the scenarios in which R1 and R2 are not completely identical (Seq. Id ≠ 100%). The 3 positions in which R1 differs from R2 are indicated with a coloured bullet. Since reads that align onto these regions will have sufficiently good mapping qualities, the identification of regular SNVs/Indels doesn't pose an issue for standard data analysis pipelines. Nevertheless, SNVs/Indels that result from a gene conversion typically remain unidentified. By only considering the coverage profile of R1, an ectopic gene conversion and a deletion look identical (panels e and f). Chameleolyser also considers the coverage at locus R2. As a result, gene conversions can be distinguished from deletions. Indeed, in case of an ectopic gene conversion, reads that originate from the acceptor site will align onto the reference sequence of the donor site resulting in a two-fold increase of the sequencing coverage as opposed to the scenario where no gene conversion is present.

The issue of variant discovery within paralogous regions is not limited to gene conversions. SNVs/Indels that are not introduced by means of a gene conversion also remain undetected, especially in genomic regions that have an identical paralog (100% sequence identity). In such cases, short sequencing reads align equally well to multiple locations in the genome and will typically be assigned a mapping quality of zero. These reads will be ignored by the variant calling algorithm as their alignment is deemed ambiguous. As a result, genetic variants that are supported by these reads are not detected (Figure 1b).

Among the methods that enable the identification of CNVs in WES and WGS data, a limited number is specifically designed to estimate copy numbers of paralogous genes (H. et al., 2010; Handsaker et al., 2015). By using the read depth at singly unique nucleotides (SUNs; the sequence differences between paralogs), it is possible to genotype the copy and content of paralogs within duplicated gene families. For regions that have an identical copy elsewhere in the genome (without SUNs), an estimate of the total copy number can be made. Despite these methods being accurate, they are designed to run on WGS data and they do not explicitly identify gene conversions. Furthermore, to our knowledge, there are currently no methods available to identify SNVs/Indels in identical paralogs. Ebbert et al., 2019 performed a thorough characterisation of paralogous regions in the human genome and suggested a strategy for rescuing variants in these regions based on re-alignment of reads to a masked reference genome, but their work did not provide a concrete solution (Ebbert et al., 2019). For these reasons, the accurate sequence analysis of paralogous regions still relies on experimental assays that only include one or a couple of genes. Typically, specific polymerase chain reaction (PCR) primers are designed to generate long-range PCR fragments which span the paralogous regions. Despite the fact that this approach has been successful for quite a number of genes these assays remain challenging to design and laborious to perform and are therefore not applied at scale (Borràs et al., 2017; Mandelker et al., 2016; Steyaert et al., 2021). Here, we present a method (Chameleolyser) that enables the identification of SNVs/Indels, CNVs and ectopic gene conversions in all paralogous regions in the coding portions of the human genome based on short-read sequencing data. By applying Chameleolyser to a cohort of 41,755 WES samples, we identify an average of 60 genetic variants per sample that could not be detected using standard WES analysis. Validation by high-fidelity long-read sequencing in 20 samples confirms >88% of called variants. Stringent filtering and clinical interpretation of these variants results in a genetic diagnosis for 25 previously undiagnosed rare disease patients. The wider application of our method might result in a new reservoir of genetic variation from which new

biological insights could be gained. Chameleolyser is implemented in Perl5 and requires a BAM or CRAM file (relative to GRCh37) as input. It runs about one hour on a single core for a single sample (depending on the enrichment kit and sequencing depth). Both raw and filtered variants are written to a tab separated file. The tool is freely available on GitHub (https://github.com/Genome-Bioinformatics-RadboudUMC/Chameleolyser) where also installation and usage instructions can be found (Steyaert, 2023).

Results

Chameleolyser works by extracting reads in the 3.5% of the exome that is affected by sequence homology (paralogous regions (Methods)) and re-aligning them to a reference genome in which all but one paralogs within each set of paralogs are masked (Ebbert et al., 2019). By masking all nucleotides in these regions in the reference genome, no sequencing reads will be aligned onto them. As a result, all reads that originate from a set of paralogous sequences are uniquely aligned onto a single region in the reference genome (the non-masked region; Figure 1b). Subsequently we perform sensitive variant calling to identify SNVs/Indels (Methods).

Homozygous deletions and ectopic gene conversion events are identified by analysing the coverage profile in the original alignment (without masking). In short-read sequencing data, a homozygous deletion and the acceptor site of a homozygous ectopic gene conversion appear identical: no reads are aligned onto that site of the reference genome. By also considering the number of reads that align onto the paralogous regions, it is possible to discriminate between deletions and gene conversions. In case of an ectopic gene conversion, the reads that originate from the acceptor site align onto the reference sequence of the donor site which results in a twofold increase in sequencing depth relative to what is expected (figure 1e-f). By applying this approach to a dataset of 41,755 exome samples we identified 2,191,707 SNVs/Indels which are not due to a gene conversion (cohort allele frequency (CAF) $\leq$ 10%; Supplementary Figure 1, Table 1, Supplementary data 1), 22,600 homozygous gene conversions that jointly introduce an additional 338,084 SNVs/Indels (CAF $\leq$ 10%; Supplementary Figure 2, Table 1, Supplementary data 2, Supplementary data 3) and 20,432 homozygous copy number losses (CAF $\leq$ 10%; Supplementary Figure 3, Table 1, Supplementary data 4). Importantly, none of the SNVs/Indels, either being the result of a gene conversion or not, were detected by a standard WES analysis (Methods).

Table 1: Observed number of variant calls, VAPs and deletions with 2 different cohort allele frequency thresholds (10% and 0.5%) in our cohort of 41,755 exome samples. All variants were annotated on Ensembl canonical transcripts. Loss-of-function (LoF) and missense (Miss) variants are relative to these transcripts.

Variant type	Due to ectopic gene conversion	Within or outside an OMIM disease gene	Transcript consequence	VAF threshold	Number of variants
deletion	NA	Within	NA	0.10	6,250
deletion	NA	Outside	NA	0.10	14,182
SNV/Indel	Yes	Within	LoF	0.10	1,043
SNV/Indel	Yes	Within	Miss	0.10	4,507
SNV/Indel	Yes	Within	Rest	0.10	56,279
SNV/Indel	Yes	Outside	LoF	0.10	341
SNV/Indel	Yes	Outside	Miss	0.10	10,970
SNV/Indel	Yes	Outside	Rest	0.10	264,944
SNV/Indel	No	Within	LoF	0.10	13,875
SNV/Indel	No	Within	Miss	0.10	142,324
SNV/Indel	No	Within	Rest	0.10	524,376
SNV/Indel	No	Outside	LoF	0.10	53,908
SNV/Indel	No	Outside	Miss	0.10	514,659
SNV/Indel	No	Outside	Rest	0.10	3,347,319
deletion	NA	Within	NA	0.005	1,182
deletion	NA	Outside	NA	0.005	3,885
SNV/Indel	Yes	Within	LoF	0.005	181
SNV/Indel	Yes	Within	Miss	0.005	1,279
SNV/Indel	Yes	Within	Rest	0.005	22,831
SNV/Indel	Yes	Outside	LoF	0.005	341
SNV/Indel	Yes	Outside	Miss	0.005	4,380
SNV/Indel	Yes	Outside	Rest	0.005	60,298
SNV/Indel	No	Within	LoF	0.005	2,000
SNV/Indel	No	Within	Miss	0.005	20,827
SNV/Indel	No	Within	Rest	0.005	79,037
SNV/Indel	No	Outside	LoF	0.005	8,760
SNV/Indel	No	Outside	Miss	0.005	79,962
SNV/Indel	No	Outside	Rest	0.005	482,720

Validation

To technically validate our variant call set, we performed whole genome high-coverage long-read sequencing (LRS) for 20 samples using PacBio high-fidelity technology (Wenger et al., 2019). Within this subset of samples, Chameleolyser identified 769 SNV/Indel calls that are not the result of a gene conversion. LRS data confirmed 678 of these calls (88.2%; Figure 2, Methods, Supplementary data 5). Of the 120/769 rare SNVs/Indels (CAF ≤ 0.5%), 111 (92.5%) are concordant with the LRS data (Supplementary data 5). Our analysis furthermore identified 8 homozygous gene conversions and 15 homozygous deletions within the subset of samples for which LRS data was generated. LRS data confirmed all ectopic gene conversions (100%) and 13/15 homozygous deletions (86.7%) (Figure 2, Supplementary data 6, Methods).

The quality of our variant call set was further evaluated by using the 6,980 parent-offspring trios that are present in our dataset. We observe that 99.0% of the SNVs/Indels that are present in the offspring is also called in one of the parents (Methods, Supplementary data 7). This suggests that only a small fraction of our variant calls are technical artifacts.

In addition to our in-house validation samples we also applied Chameleolyser to 5 genome-in-a-bottle samples (Methods). Since the identification of deletions and gene conversions requires a larger number of samples enriched with the same enrichment kit, the precision analysis was restricted to SNVs/Indels (not the result of a gene conversion). From the 118 SNV/Indel calls made by Chameleolyser, 98 are concordant with LRS (83.1%; Methods, Supplementary data 8). From the 39 calls corresponding to rare SNVs/Indels, 35 were concordant with LRS (89.7%; Supplementary data 8).

Comparison with other variant callers

Chameleolysers ability to identify SNVs/Indels (not the result of a gene conversion) was compared with GATK and DeepVariant (Poplin, Chang, et al., 2018). The sensitivity for both of these tools is exactly zero within genomic regions that are associated with zero mapping qualities in WES (Supplementary data 9, Figure 1b), With Chameleolyser a sensitivity of 43% is achieved (Methods). In regions onto which sequencing reads align uniquely, it has been shown that GATK and DeepVariant are excellent tools for the identification of SNVs/Indels (Lin et al., 2022). Within these regions, the added value of Chameleolyser is limited with a sensitivity of 88.0% compared to 86.3% for GATK (Methods).

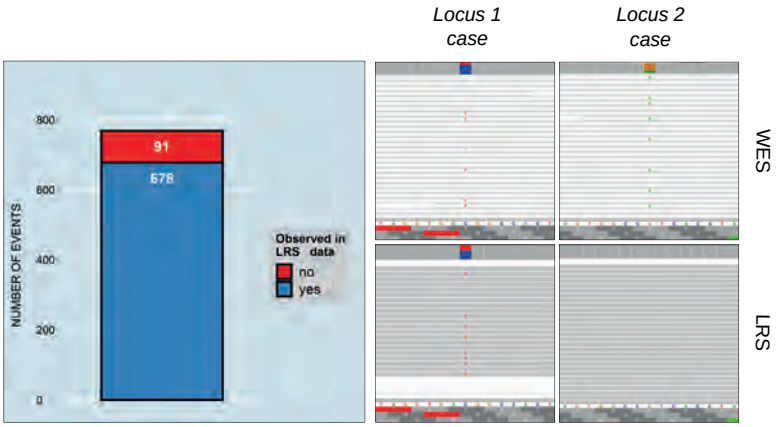
Sensitivity could not be assessed for homozygous deletions and ectopic gene conversions since we cannot, due to the availability of only a limited number of long-read sequencing samples, derive a call set of high-quality events with a population allele frequency ≤ 0.10 (Methods). The unique value of Chameleolyser can however be demonstrated by comparing its output with ExomeDepth (Plagnol et al., 2012) and Conifer (Krumm et al., 2012) (Methods, Supplementary Figure 4). Within the 20 in-house samples for which LRS alignments were generated, there are 4 events (3 deletions and 1 gene conversion) that are only called by Chameleolyser. Of these, 2 events (1 deletion and 1 conversion) were concordant with the LRS alignments. The other 12 deletions and 7 conversions that were identified by Chameleolyser are all called as deletions by ExomeDepth. As opposed to Conifer (for which there are no homozygous deletion calls within the validation samples), ExomeDepth made an additional 201 homozygous deletion calls which were not made by the other tools. Based on the LRS alignments we estimated the precision at 32.5% (Methods, Supplementary data 10).

Variants with ambiguous positions

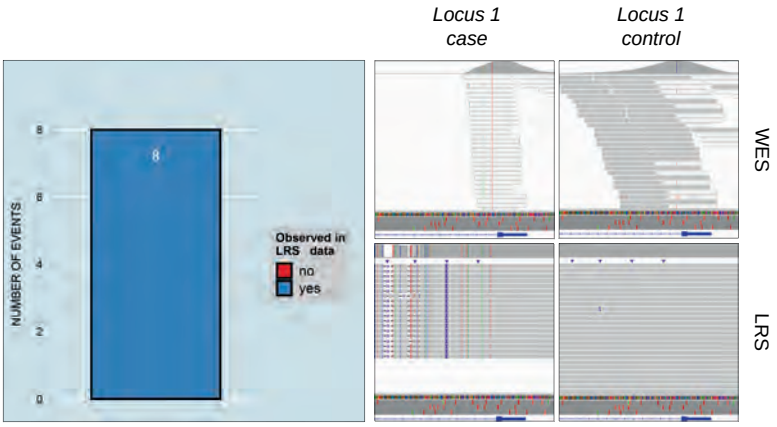
Heterozygous SNV/Indel calls (not due to a gene conversion and not corresponding to SUNs (methods)) result from a genomic alteration in one of the paralogs within the respective set of paralogs (Figure 1b). Since short-read data does not contain the information to discriminate between the different paralogs in an identical set of paralogs, all possible variants that could have caused the variant call are computed and annotated (Methods). In the remainder of the text we will call these "variants with ambiguous positions" (VAPs). This uncertainty is not applicable for variants which are homozygous in all paralogs nor is it relevant for gene conversions and deletions since these events are identified based on coverage data (Methods).

Approximately 10% of VAPs originate from protein-altering variants or from the corresponding alteration in only one possible non-coding paralog (the variant thus resides in a set of 2 paralogs of which one is coding and the other is not). In principle we would expect that half of these VAPs actually reside in the coding region. However, selection may act more on coding regions which could lead to an overrepresentation of VAPs that are actually present in non-coding regions. In order to derive the fraction of VAPs that originate from protein-altering variants we used two different approaches. Firstly, by using the LRS data that we used for validation purposes we can determine the actual location of these VAPs and thus determine the fraction. Within the 20 WES samples for which we generated LRS data, we identified 65 VAPs satisfying the aforementioned criteria. From these, 25 (38%) turned out to be present in the coding regions (p-valuebinom= 0.08; Supplementary data 11).

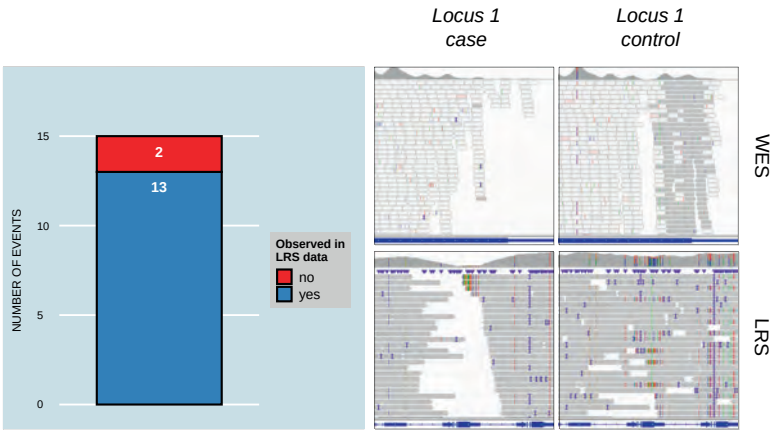
SNV/Indel not due to gene conversion



homozygous gene conversion



homozygous deletion



< Figure 2: Overview of validation successes with LRS. Each variant type (single nucleotide variant and small indel (SNV/Indel) not due to a gene conversion, homozygous gene conversion and homozygous deletion) is accompanied with a bar chart and 1 concrete example. Within each bar chart, the correspondence between Chameleolyser and long-read sequencing (LRS) is shown. The genomic coordinates of the SNV (not due to a gene conversion) in the IGV screenshots is chr2:96,692,489-C/T (locus 1) and chr2:96,463,586-G/A (locus 2). In whole-exome-sequencing (WES) we observe that roughly 25% of the reads at each locus support the variant allele. Based on LRS we clearly see that the genomic alteration is present as a heterozygous SNV in locus 1 (and not in locus 2). The genomic coordinate of the homozygous gene conversion that is shown in the IGV screenshot is chr1:22,338,347-22,339,613 (*CELA3A*; locus specific). WES data shows a clear difference between a sample with (a case) and a sample without the event (a control). In the case (as opposed to the control) there isn't any read that uniquely aligns onto the beginning of intron 1. Considering the LRS data, we see that this region is not deleted in the case sample. In contrast, we see several SNVs/Indels which are absent in the control sample. These alterations indeed correspond to sequence differences between *CELA3A* and *CELA3B* (Supplementary data 22). A conversion from *CELA3B* to *CELA3A* is responsible for these SNVs/Indels being present in *CELA3A* in the case sample. The genomic coordinate of the homozygous deletion that is shown in the IGV screenshot is chr9:84,545,162-84,547,705. The difference between a case and a control sample can be seen in WES. This corresponds to the LRS data.

A second approach to estimate the fraction of coding variants among VAPs uses the ratios of synonymous, missense and loss-of-function (LoF) variants in the paralogous and non-duplicated (unique) regions of the exome. Using our standard WES analysis pipeline (Lelieveld, Reijnders, Pfundt, Yntema, Kamsteeg, de Vries, et al., 2016) we find exome-wide ratios of 1.19 and 0.043 for missense to synonymous and LoF to synonymous variants respectively (Supplementary data 12). Within the homologous regions of the exome Chameleolyser identified on average 6.3 synonymous, 11.1 missense and 1.40 LoF VAPs per sample (Supplementary data 13). We assume that synonymous variants are not under strong selection and thus that half of synonymous VAPs actually originate from variants residing in protein coding regions. If we further assume that the ratios of missense to synonymous and LoF to synonymous variants are comparable between paralogous and non-paralogous regions, we can calculate the proportion of missense and LoF variants among VAPs as 33.8% and 10.0%, or 3.75 and 0.14 variants per sample respectively. As such, we provide two lines of evidence that roughly 30-40%, of protein-altering VAPs resides in protein coding regions.

Systematic analysis of SNVs/Indels results in 14 diagnoses

In order to investigate SNVs/Indels that could be of clinical interest we only considered variants in exomes of patients that were molecularly undiagnosed (n=17,650; Supplementary data 14). We selected missense and LoF VAPs with a CAF ≤ 0.5%, and occurring in clinically relevant genes according to predefined gene panels for which an investigation was requested for the particular patient. In addition, we included a single synonymous variant in *SMN1* (chr5:g.70,247,773C>T(GRCh37)) that is known to lead to a truncated protein product (Lorson et al., 1999).

The application of the aforementioned filter criteria to our variant call set resulted in 1,071 heterozygous VAPs (131 LoF and 940 missense; Supplementary data 15) as well as 57 homozygous variants (5 LoF, 46 missense and 6 synonymous; Supplementary data 16). All of the homozygous variants are introduced in the gene of interest by means of gene conversions that most likely occurred in a proximal or distant ancestor (a total of 21). Importantly, the genomic positions of these homozygous variants are not ambiguous (hence these are not VAPs), but clear site-specific calls (Figure 1d-e).

Among the 1,071 rare VAPs that we identified in our cohort there were 7 alterations in the *STRC* gene that occur in patients in which we also identified a heterozygous multi-exonic deletion (Supplementary data 17, Supplementary Figure 5). Validation

experiments consisting of multiplex ligation-dependent probe amplification (MLPA) and long-range polymerase chain reaction (PCR) followed by sequencing were conducted for all of the 7 individuals. This confirmed that all of the 7 deletions and 4 out of the 7 SNVs/Indels (1 LoF, 3 missense – all in trans with the deletion) were present in the *STRC* gene (and thus not in its pseudogene; Table 2; Figure 3; Supplementary Figure 6), resulting in 4 genetic diagnoses. The other 1,064 VAPs did not reveal any additional diagnosis. Either the phenotype that is associated with the gene of interest did not sufficiently match the clinical presentation of the patient or the disease gene is recessive where only a heterozygous variant is identified.

Out of the 21 ectopic gene conversions that were interpreted, 11 were considered as not causal for disease due to their frequency among the patients for which the specific gene was not a gene of interest. The other 10 conversions provided a direct diagnosis (Table 2; Figure 3; Supplementary Figure 7; Supplementary Figure 8). All of these events were found in one of only three genes: *STRC* (n=1), *OTOA* (n=3) and *SMN1* (n=6). The conversion from *STRCP1* to *STRC* causes a LoF variant to be introduced in *STRC* and thus leads to a null allele (Shearer et al., 2014). The 3 gene conversions that affect the *OTOA* gene also lead to null alleles as a result of a LoF variant being introduced. This conversion that affects exon 22 of the *OTOA* gene (ENST00000646100) has previously been discussed by Laurent et al., 2014 (Laurent et al., 2021). The conversion from *SMN2* to *SMN1* which was found in 6 patients with spinal muscular atrophy (SMA) is causative for disease as a result of a synonymous variant that is introduced in the *SMN1* gene. This variant leads to altered splicing and, as a consequence, results in a non-functional protein product (Lorson et al., 1999). By using MLPA we confirmed the bi-allelic losses of the *STRC* and *SMN1* alleles. Using long-range PCR and long-read PacBio sequencing we confirmed the bi-allelic losses of the *OTOA* alleles.

Importantly, among the individuals for which the deafness disease gene panel was not requested we did not identify any homozygous LoF-introducing gene conversion in *STRC* or *OTOA*. The same holds true for *SMN1*: all of the identified pathogenic gene conversions were exclusively found amongst SMA patients. This illustrates the very high precision of our calls (100%; p-value$_{\chi^2}$,*STRC* = 6.26e-2; p-value$_{\chi^2}$,*OTOA* = 1.51e-11; p-value$_{\chi^2}$,*SMN1* = 1.13e-11).

Table 2: Overview of new genetic diagnosis in our study cohort as a consequence of disease-causing variations identified with Chameleolyser. The first column indicates in which sample the variant was identified. The second, third and fourth column respectively represent the chromosome, genomic start and end of the event (hg19). In the next column, the type of genetic event can be found. The sixth column indicates the respective gene symbol and the last column the associated disease is displayed.

Sample	Chrom	Start	End
SAMPLE_24323	chr16	21747381	21747911
SAMPLE_29813	chr16	21747381	21747911
SAMPLE_30025	chr16	21747381	21747911
SAMPLE_26907	chr5	70247601	70248925
SAMPLE_28821	chr5	70247601	70248925
SAMPLE_36286	chr5	70247601	70248925
SAMPLE_37053	chr5	70247601	70248925
SAMPLE_39455	chr5	70247601	70248925
SAMPLE_20848	chr5	70247601	70248925
SAMPLE_23606	chr15	43890861	43897797
SAMPLE_37062	chr16	21747381	21747911
SAMPLE_37080	chr16	21747381	21747911
SAMPLE_23649	chr5	70247601	70248925
SAMPLE_6943	chr5	70247601	70248925
SAMPLE_27880	chr5	70247601	70248925
SAMPLE_9901	chr5	70247601	70248925
SAMPLE_29108	chr5	70247601	70248925
SAMPLE_30394	chr5	70247601	70248925
SAMPLE_31929	chr5	70247601	70248925
SAMPLE_31987	chr5	70247601	70248925
SAMPLE_40265	chr5	70247601	70248925
SAMPLE_21563	chr15	43908399	43908399
		43890861	43894856
SAMPLE_32502	chr15	43906154	43906154
		43890861	43894856
SAMPLE_38648	chr15	43908409	43908409
		43890861	43894856
SAMPLE_36262	chr15	43908184	43908184
		43890861	43894856

TypeOfEvent	GeneSymbol	OMIM
Conversion	*OTOA*	Deafness, autosomal recessive 22
Conversion	*OTOA*	Deafness, autosomal recessive 22
Conversion	*OTOA*	Deafness, autosomal recessive 22
Conversion	*SMN1*	Spinal muscular atrophy-1-4
Conversion	*SMN1*	Spinal muscular atrophy-1-4
Conversion	*SMN1*	Spinal muscular atrophy-1-4
Conversion	*SMN1*	Spinal muscular atrophy-1-4
Conversion	*SMN1*	Spinal muscular atrophy-1-4
Conversion	*SMN1*	Spinal muscular atrophy-1-4
Conversion	*STRC*	Deafness, autosomal recessive 16
Deletion	*OTOA*	Deafness, autosomal recessive 22
Deletion	*OTOA*	Deafness, autosomal recessive 22
Deletion	*SMN1*	Spinal muscular atrophy-1-4
Deletion	*SMN1*	Spinal muscular atrophy-1-4
Deletion	*SMN1*	Spinal muscular atrophy-1-4
Deletion	*SMN1*	Spinal muscular atrophy-1-4
Deletion	*SMN1*	Spinal muscular atrophy-1-4
Deletion	*SMN1*	Spinal muscular atrophy-1-4
Deletion	*SMN1*	Spinal muscular atrophy-1-4
Deletion	*SMN1*	Spinal muscular atrophy-1-4
Deletion	*SMN1*	Spinal muscular atrophy-1-4
hemizygous SNV/Indel: G>C heterozygous deletion	*STRC*	Deafness, autosomal recessive 16
hemizygous SNV/Indel: G>C heterozygous deletion	*STRC*	Deafness, autosomal recessive 16
hemizygous SNV/Indel: G>- heterozygous deletion	*STRC*	Deafness, autosomal recessive 16
hemizygous SNV/Indel: C>G heterozygous deletion	*STRC*	Deafness, autosomal recessive 16

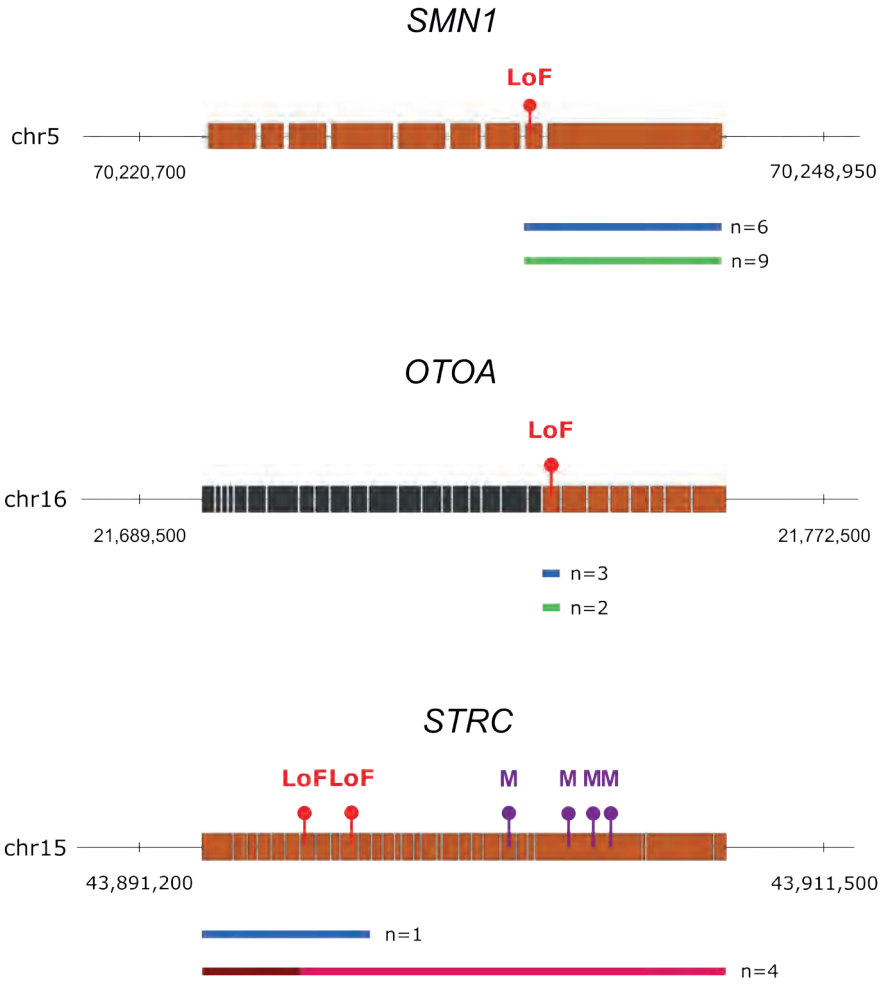


Figure 3: Overview of previously unidentified pathogenic variants. The 3 genes in which we identified pathogenic variants are represented by a model of their Ensembl canonical transcript. The orange parts of the gene are affected by sequence homology (thus incorporated in our analysis). The black parts are not. Homozygous gene conversions are illustrated with blue rectangles (n=10). Loss-of-function variants (LoFs) which are introduced by means of these gene conversions are indicated with a red bullet. Homozygous deletions are indicated with green rectangles (n=11). The bordeaux rectangles underneath *STRC* represent a heterozygous deletion (n=4). The darkest part indicates the genomic region that was inspected for heterozygous deletions. This rectangle is extended with a lighter coloured rectangle to indicate the actual span of the deletion (based on MLPA). Each of these 4 respective deafness patients have an ultra-rare hemizygous missense variant (M), indicated with a purple bullet.

In total, the analysis of SNVs/Indels within paralogous coding regions of known disease genes in previously undiagnosed patients resulted in 14 new diagnoses.

Systematic analysis of homozygous deletions results in 11 diagnoses

Analogous to the SNVs/Indels, homozygous deletions were filtered prior to clinical interpretation. Only events with a CAF ≤ 0.5% that affect a gene that is present in the disease gene panel of interest for an undiagnosed patient were considered. Application of this filter resulted in 147 homozygous deletions (Supplementary data 18). Among these were several known genetic causes for disease, such as a bi-allelic loss of *OTOA* exon 22 identified in two patients with deafness (Laurent et al., 2021), and 9 homozygous deletions of *SMN1* exon 7 in SMA patients (Lefebvre et al., 1995) all of which were confirmed with MLPA.

In the group of individuals for which the deafness disease gene panel was not requested, no homozygous *OTOA* deletions were identified. Among the individuals for which *SMN1* was not present in the disease gene panel of interest, only one homozygous *SMN1* deletion was found. This may represent a case with a very mild phenotype as has been reported in literature (Brahe et al., 1995). When we conservatively assume that this call is false positive, the precision of our *OTOA* and *SMN1* deletions remains high (91.7%; p-value$_{\chi 2}$,*OTOA* = 9.02e-7;p-value$_{\chi 2}$, *SMN1* = 3.51e-16).

Overall, the analysis of homozygous copy number variants in known disease genes revealed 11 pathogenic deletions leading to a diagnosis in previously undiagnosed patients (Table 2; Supplementary Figure 7, Supplementary Figure 8).

Distinguishing ectopic gene conversions from deletions

By using *STRC* as an example, we wanted to investigate whether any patient was diagnosed with a pathogenic deletion but in which the real underlying genetic event is most likely an ectopic gene conversion (Figure 1e-f). Our in-house diagnostic pipeline identified 58 homozygous deletions in the subcohort of patients with hearing impairment. All of these events were confirmed by using MLPA. Using Chameleolyser we also found homozygous losses of *STRC* alleles for these 58 patients. However, in only 37 of these we actually detected a homozygous deletion. In the remaining 22 (37%) we identified, based on coverage profiles, a homozygous gene conversion from *STRCP1* to *STRC* (Supplementary data 19). All of these gene conversions are predicted to affect at least exons 19-23 (ENST00000450892) and therefore introduce LoF variation into *STRC*. As a consequence, the pathogenicity of the identified deletions and gene conversions is the same and thus, the

genetic diagnosis of a homozygous *STRC* deletion in the 22 patients in which we identified a gene conversion does not pose an issue. Nevertheless, the ability of Chameleolyser to distinguish homozygous deletion events from gene conversions events is clinically very relevant since the vast majority of gene conversions is benign. For example, in our cohort of 41,755 samples we identified 47 homozygous gene conversions from *STRCP1* to *STRC* that do not introduce LoF variation (Supplementary data 20). Since these events are present in patients with all kinds of different phenotypes as well as in healthy parents of patients we can reasonably assume that these events are benign. This can only be true in case the alleles are indeed converted and not deleted. However, we note that using ExomeDepth all of these events are called as homozygous deletions of *STRC* exons. This potentially poses a risk for making an erroneous molecular diagnosis.

Discussion

We developed a bioinformatics method to systematically analyse all coding paralogous regions in 41,755 individuals using existing WES data. We identified an average of 60 variants per sample that could not be detected using standard WES analysis. Of these, about 1% is a missense or LoF variant with an allele frequency ≤ 0.5% in one of the 332 OMIM disease genes that are affected by sequence homology (Supplementary data 21). We carefully interpreted a subset of these variants, namely the variants within the genes in the requested disease gene panels. Doing so, we could establish a genetic diagnosis for 25 previously undiagnosed patients by either SNVs/Indels, gene conversions or CNVs, or the combination thereof. All of these pathogenic variants were identified in 1 of 3 genes: *STRC*, *OTOA* and *SMN1*. For the respective patient groups (patients with hearing impairment and patients with spinal muscular atrophy) our method solved > 1% of previously undiagnosed patients. As our approach identifies causal variants in known disease genes, we believe that it may also be used to find novel disease genes.

We noted that using standard data analysis approaches, CNV callers that are applied on WES or WGS data are unable to discriminate between gene conversions and deletions. Indeed, gene conversions are falsely called as deletions as a consequence of a reduced number of reads at the acceptor site of the conversion (Figure1). Sometimes, the pathogenicity of a gene conversion and the corresponding deletion is the same, e.g. a LoF-introducing gene conversion in *OTOA* or *STRC*. In such a scenario there is no risk in making a wrong molecular diagnosis, and only the exact genomic alteration that is responsible for the patient's phenotype will be

wrong. However, most gene conversions are benign, and a genomic deletion may be inferred by standard WES tools, where Chameleolyser could provide an accurate diagnosis. Our technical validation efforts demonstrate that large part of the issues related to the analysis of duplicated genomic sequences are resolvable with novel sequencing technologies (roughly 90% of called variants is concordant with HiFi PacBio data). Undoubtedly, the generalized usage of these novel technologies will further help the field to characterize these difficult genomic regions – much beyond to what Chameleolyser can offer based on short-read data. It also provides input to generate a more complete and higher quality human reference genome (T2T (Sergey et al., 2022)) which in turn improves variant discovery for both short and long-read data (Noyes et al., 2022; Nurk et al., 2022). We foresee that these feedback loops will continue to accelerate the quality of sequence analysis in the next decades. Currently however these novel sequencing technologies are highly expensive and therefore not affordable for a health-care system. As a result, the rate at which short-read data is produced is still much higher as compared to LRS. For all of these short-read data our method offers an effective way to query the difficult parts of the exome and genome. In this study we applied Chameleolyser to the large number of WES datasets that are currently available in our medical genetics center (van der Sanden et al., 2021; Yauy et al., 2020). Chameleolyser could equally well be applied to short-read WGS data. Direct application would however only consider homologous coding regions. A future update of Chameleolyser for WGS could also incorporate homologous regions that only affect non-coding regions, although the interpretation of identified variants in such regions would be very challenging.

In conclusion, we present a bioinformatics method to identify genetic variation in paralogous genomic regions. By analysing 41,755 WES samples we identified a genetic diagnosis in 25 previously undiagnosed patients. We expect that Chameleolyser can substantially contribute to future discoveries based on genome variation that has so far remained hidden.

Methods

Samples

The analysis was applied on 41,755 WES samples including 6,980 patient-parent trios (20,940 samples (50%)). All samples were sequenced either using Illumina HiSeq2000, Illumina HiSeq4000 or BGI DNBSEQ short reads sequencing platforms. 6,894 exomes were enriched by using the Agilent SureSelect Human All Exon V4 kit

while for the other 34,861 Agilent SureSelect Human All Exon V5 was used. All of these data were generated between 2002 and 2020 as part of the routine genetic investigation from Genome Diagnostics Nijmegen. As such, data processing was conducted with our standard diagnostics WES pipeline (Lelieveld, Reijnders, Pfundt, Yntema, Kamsteeg, de Vries, et al., 2016). Depending on the patient's phenotype a specific gene panel is requested in which genetic variants are inspected and interpreted. After this diagnostic screening, 17,650 patients remained molecularly undiagnosed (Supplementary data 14).

Identification of paralogous regions

Two different sets of paralogous regions were derived after which they were merged. For the first set, all protein coding genes with one or multiple pseudogenes were used as a starting point. For the second set, the genomic coordinates of reads with low mapping quality were used.

Set I: Regions in protein coding genes with known pseudogenes

Starting from all pseudogenes in the comprehensive gene annotation file from Gencode 31 (lift37; https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_31/GRCh37_mapping/gencode.v31lift37.annotation.gff3.gz), those with a corresponding protein coding gene were selected (n=1,680). This correspondence was based on the HGNC gene name in the file. Next, by using MAFFT v7.407 (Katoh & Standley, 2013), a multiple sequence alignment was generated between each protein coding gene and its pseudogenes. To only keep the regions for which the protein coding gene exactly has 1 paralogous pseudogene and for which the sequence identity between the protein coding gene and the pseudogene is 90% or more, a sliding window approach was used (window length = 100 bp). This resulted in a set of regions in 989 protein coding genes with their respective pseudogenes (Supplementary Figure 9).

Set II: Regions corresponding to low mapping qualities

For 250 randomly chosen WES samples (SureSelect Human All Exon V5), low quality reads (i.e. mapping quality (MQ) < 10) were extracted using samtools 1.9 (Danecek et al., 2021). To avoid that poorly covered regions (and thus uninformative in terms of variant identification) are included in the region set, bedtools v2.28.0 (Quinlan & Hall, 2010) was used to only keep the genomic positions with a sequence depth ≥ 10. Because many regions were fragmented, i.e. separated by only a few bases, the resulting regions were merged using three different slopping distances: 250 bp, 500 bp and 5% of the region length (bedtools). In the last list we removed regions < 50 bp (Supplementary Figure 10a). To find the homologous relations between the

regions in the lists of regions we performed pairwise sequence alignments (PWA) with EMBOSS Needle v 6.6.0.0 (Needleman & Wunsch, 1970) as follows. Per list (250bp, 500bp and 5%), PWAs were made between each region and each other region in the file, as well as with its reverse complement (Supplementary Figure 10b). We then defined sets of paralogs in the following way: all regions within a set should mutually have an alignment score ≥ 0.9 (it has been shown that the sequence identity between a linked donor and acceptor site is ≥ 90% with very few exceptions (J.-M. Chen et al., 2007)). Furthermore, paralogs are only tolerated in a set when they do not have an alignment score ≥ 0.9 with a region in another set of paralogs (Supplementary Figure 10c). Before merging the 3 lists (each corresponding to a different slopping distance), sets without any exonic overlap and with > 5 members were removed. The actual merging process starts with the region list corresponding to a slop distance of 500. To that list, regions from the 2 other lists were added (first 250bp, next 5%) with the following rule: only if a region doesn't overlap with regions which are already in the merged region list, the region is added. This resulted in 1,334 regions with their paralogs.

Merging region set I and II

To combine region set I and II into 1 final set of regions to operate the paralogy analysis on, the same iterative procedure as above (i.e. merging set II regions with different slopping distances) was used. Here, we start from the full list of set I regions. To that list, set II regions were added only if there is no overlap. Doing so, 177 sets of regions were removed. The output of this step corresponds to the variant calling regions in the paralogy analysis (Supplementary data 23). In order to limit the number of broken read pairs in the extraction procedure (cf. mapping and short variant calling), the region list for read extraction consists of the same regions extended with 500 bp up and downstream (Supplementary data 24).

Generating the masked reference genome

When re-aligning the sequencing reads, all regions except 1 in a set of paralogs should be masked in the reference genome. For this, we choose to mask the regions that had the least overlap with a protein coding sequence (Supplementary data 25). Masking was conducted with bedtools v2.28.0.

Mapping and short variant calling

All reads overlapping the read extraction regions were extracted using samtools 1.9. On the resulting alignments, variant calling using GATK 4.1.2 (Van der Auwera et al., 2013) was performed. Next, the aligned reads were converted to FASTQ using samtools 1.9. BBmap v 38.56 was used to remove broken read pairs. Alignment of the reads to the

masked reference sequence was done with bwa 0.7.17 (Li & Durbin, 2009). To remove duplicate reads, Picard v. 2.20.8 was used. Because reads from paralogous regions are aligned to a single region, we do not expect 50% allele ratios for variants. Therefore we used LoFreq 2.1.3.1 (Wilm et al., 2012) for sensitive variant calling on this newly generated alignment (parameters: no-default-filter, use-orphan, no-baq, no-mq, sig=1).

Identification of deletions and conversions

Determination of subregions

Region set I contains regions that consist of multiple exons and introns. Since we want to call deletions and conversions at the resolution of a single exon, we split these regions into subregions. This was achieved by intersecting the list of regions with the list of all protein coding exons from Gencode 31 (including 200 bp intronic flank; lift37). For region set II this operation is not needed because these regions are small by design (*i.e.* they do not contain large introns). Nevertheless, in order to ensure high quality deletion and conversion calls, region set II was filtered to only keep sets consisting of 2 paralogs. In total we end up with 4,921 subset of regions (Supplementary data 26).

Readcounts per subregion

Bedtools v2.28.0 was used to derive, per sample, the number of reads that align to the different subregions. This coverage calculation was applied on the original alignment (no masking applied).

Kernel density estimation

In order to accurately identify deletions and ectopic gene conversions it is important to consider a large set of samples at once. Doing so, it can be seen from the coverage distribution if a poorly covered sample is part of a wide distribution originating from samples without deletion or it is not part of such a distribution and thus it is likely a sample with aberrant coverage due to a genetic event. The identification of these coverage peaks is done in 2 steps. First we estimated the density of the data with the technique of kernel density estimation (KDE). This technique smoothens the discrete data, *i.e.* it results in a continues curve with aligns with the density of the data. For this scikit-learn was used (exponential kernel; default parameters; Pedregosa et al., 2011). After having estimated the density we applied the argrelextrema function from the scikit-learn software package to determine the local minima and local maxima of the curve. This results in peaks or KDE clusters. This operation was separately done for deletions were both paralogs in a set were deleted and deletions plus gene conversions affecting only 1 region in a set.

Deletions of both paralogs within a set of paralogs

To detect the events where all regions in a set of paralogs are deleted, a vector with the per-sample number of reads that align onto one of the regions in the set is used as the input for the KDE. This is only done for sets for which the median (in the cohort; per enrichment kit) of the total number of reads is 120 or more. To find rare deletions, all KDE clusters (peaks) having > 10 reads or corresponding to > 10% of samples in the cohort were excluded (taking into account that some read ends might have an alignment in the deleted region, we tolerate some reads to be aligned). The analysis was done separately for Agilent V4 and Agilent V5 samples and, importantly, for this analysis both uniquely and non-uniquely aligned reads were used. This resulted in 1,962 calls in our full cohort.

Gene conversions or deletions of 1 region in a set

To identify gene conversions or deletions of 1 region in a set of regions, a vector of per-sample read count ratios is used. We call this ratio *R*: for a set of paralogs consisting of region X and region Y it is the number of reads that uniquely align to X divided by the number of reads that uniquely align to X or Y. To be able to technically discriminate between a low (possible acceptor) and a high (possible donor) covered region, within this vector, samples with < 30 uniquely aligned reads (sum of both paralogs) were excluded. Furthermore, regions for which the median number of reads (in the cohort; per enrichment kit) in one of the two paralogs is below 60 were excluded in the analysis (To be able to discriminate samples with low from samples with high coverage). To identify rare genetic events, all KDE clusters (peaks) corresponding to > 10% of the samples in the cohort were excluded for further analysis. Furthermore, all alterations not overlapping a protein coding gene were discarded. This resulted in 49,151 calls. Within this call set we defined an event to be either the homozygous deletion of X or the conversion from Y to X if 1) maximum 10 uniquely aligned reads onto X 2) $R \leq 0.05$ (*e.g.* we tolerate (acceptor) regions with 10 reads aligned to them only if the number of reads on the possible donor site is at least 20 times higher).

The distinction between a deletion and conversion is based on the number of reads that uniquely align onto Y (the donor site in case of a conversion). First, the number of uniquely aligned reads onto X and Y are normalized per sample (based on the total number of reads over all paralogs). Next, for each sample, a one-sided percentile was calculated. For sample S and region X we call this metric $\text{PercDPN}_{X,S}$, for region Y it is $\text{PercDPN}_{Y,S}$. We now calculate a threshold

$$T_{X,S} = \text{PercDPN}_{X,S}^{\frac{1}{3(1-3\text{PercDPN}_{X,S})}} \quad (1)$$

An event is predicted to be a gene conversion if $PercDPN_{Y,S} < T_{X,S}$ (2) and if the normalized number of reads that align onto $Y > \frac{4}{3} Median\ Y_{norm}$ (3). If these conditions are not met, we predict region X to be deleted. Formula (2) is implemented to accommodate for the fact that not all ectopic gene conversions are very rare. The basic idea is to require a more extreme coverage (i.e. an extreme observation in the subcohort of all samples enriched with the same enrichment kit) on the possible donor site when the coverage on the acceptor site is very extreme. When for example 1,000 samples only have few reads aligned onto the possible acceptor site then it is possible and likely that there is at least in part of the samples an ectopic gene conversion. And as a result quite some samples will have an elevated coverage on the donor site. If we would then require a very extreme coverage observation in order to predict the event as an ectopic gene conversion, we would be wrong for most of the samples (overly conservative).When only few samples have a very low coverage on the possible acceptor site, it is impossible that there are many samples with an ectopic gene conversion. So in that case we can be more strict. By replacing $PercDPN_{X,S}$ by a very small value (e.g. 10^{-5}) in the formula, we can see that it the formula can be approximated by the cubic square of $PercDPN_{X,S}$. When we on other hand substitute the parameter by 10^{-1} it can be approximated by the square root and thus a less stringent read depth requirement for the possible donor site.

Combining single exon CNVs

After having derived the deletions and conversions per subregion we merged the ones that are in direct proximity. For this, all filtered calls were annotated with their overlapping gene name using Gencode 31 (lift37). First, the calls were combined per gene (*e.g.* if for a certain patient 2 different exons are deleted in the same gene, these calls are merged into 1 deletion). Next, deletions and conversions in neighbouring genes were merged. Also when 1, 2 or 3 coding genes (but not necessarily part of a paralogous set of sequences) exist between the 2 different CNV calls, it was assumed that these originate from the same genetic event and therefore these calls were merged into a single call.

Short variant processing

After having called variants in the original and newly generated alignments (*i.e.* after masking) we made a raw variant call set with variants of interest.

Raw call set

The raw variant call set consist of variants satisfying the following criteria:

- The variant is not an alteration of a singly unique nucleotide (SUN; *i.e.* sequence difference between homologs)
- Depth of overage at the position of interest in the masked alignment ≥ 60. With 30x read depth an almost optimal sensitivity is achieved in WES for SNV/Indel identification (Meynert et al., 2013). A threshold of 60x is chosen since the vast majority of paralogous sets consist of 2 paralogs.
- Variant allele fraction (VAF) is ≥ 0.15. In a pool of reads originating from 2 alleles, a heterozygous variant has an expected read ratio of 0.5. This ratio becomes 0.25 for a heterozygous variant in 1 of 2 paralogs, each having 2 alleles. If we consider all variants with a read ratio of 0.15 and higher, we obtain 97.9% sensitivity under a binomial model assuming 60x read depth (and probability 0.25). As shown before, the distribution of the ratio of reads that support the variant allele approximately follows a binomial distribution (Heinrich et al., 2012).
- The variant is not present in the original VCF (GATK variant calling; no masking)

This resulted in 56,156,453 calls.

Expansion: from variant calls to variants

Except for calls corresponding to homozygous variants in all paralogs, it is unknown in which region (within the group of paralogous regions) the variant is present. For that reason we need to compute all possible variants (*i.e.* VAPs) corresponding to a variant call. The number of possible variants equals the number of paralogs in the set. The actual computation is done with an in-house Perl script and is based on the MSA between the different paralogs. The 56,156,453 calls correspond to 119,551,166 VAPs.

Annotation

All variants were annotated on canonical transcripts using Ensembl VEP 97.

Filtering

Several filters were applied in order to transform the raw variant call set to a high accuracy call set with variants of interest:

- Sine the major focus is on (relatively) rare variants, we excluded variants which were observed in > 10% of the analysed samples for further analysis. This resulted in 4,064,684 calls corresponding to 8,753,075 VAPs.

- Variants with an apparent good quality in a particular sample but which are of low quality in most other samples were filtered out. This is implemented as follows: the variant is removed from the call set if the number of samples having the variant with a VAF > 0.05 is more than twice the number of samples having the variant with a VAF > 0.15. With that we filter out variants for which most samples have a VAF between 0.05 and 0.15. This resulted in 2,474,765 calls corresponding to 5,152,554 VAPs.
- It has been shown that the Illumina sequencing technology is prone to small indel errors within homopolymer tracts (Ross et al., 2013) and post-homopolymer substitutions (Stoler & Nekrutenko, 2021). In general, the longer the mononucleotide run, the more sequencing bias is introduced, but from a tract length of 6 and more, the errors become most apparent. For that reason, we took genomic coordinates for all homopolymers in the genome > 5 mononucleotides from https://github.com/ga4gh/benchmarking-tools. These intervals were extended up and downstream with 2 base pairs. All variants in these regions were excluded for further analysis. This resulted in 2,337,271 calls corresponding to 4,859,954 VAPs.
- Because the pairwise alignments of paralogous regions (which was used to derive the singly unique nucleotides) and the alignment of the short next-generation sequencing reads (based on which variants were identified) can be slightly different in regions with several sequence differences between paralogs, we ignored these subregions because we otherwise would have an inflation of false positive variants due to unrecognised singly unique nucleotide. For this we excluded variants in subregions containing 5 singly unique nucleotides in a stretch of 10 bp or less This resulted in 2,191,707 calls corresponding to 4,596,461 VAPs.

Validation

Technical validation

HiFi PacBio sequencing reads were generated for 20 samples using the Pacific Biosciences Sequel II instrument with Chemistry 2.0. DNA for all samples was sheared using a Megaruptor 3 instrument aiming fragments of 18kb. SMRTbell Express 2.0 was used to prepare the library, the PippinHT instrument for fragment size-selection >10 kb. Finally, sequencing was conducted with 3 SMRT Cells per sample targeting 30x coverage. Sequencing reads where aligned to the GRCh38/Hg38 genome with minimap2 (Li, 2018). Variant calling was conducted with DeepVariant (Poplin, Chang, et al., 2018).

Chameleolyser initially identified 15 homozygous deletions, 8 homozygous gene conversions and 847 SNV/Indel calls not due to a gene conversion within these 20 samples based on the short-read data. The coordinates of these variants were converted to hg19 with CrossMap (Zhao et al., 2014). For SNVs/Indels not due to a gene conversion, if only 1 possible variant could be converted (and thus the other(s) failed to be converted) we discarded the variant for validation purposes. Doing so, we ended up with 769 variant calls to be validated. Deletions and gene conversions were manually checked using the Integrative Genomics Viewer (IGV). For SNVs/Indels (not due to a gene conversion), a 2-steps approach was followed. Firstly, the variant was checked in the VCF. This resulted in 557 calls that were concordant with the long-read data. Next, we manually checked (using IGV) the variants that could not be validated using the VCFs. This resulted in an extra 121 validated variant calls (Supplementary data 5).

All 15 deletion calls were visually inspected in the LRS alignments. For 8/15 deletion calls there is a maximum of 1 LRS read aligning onto the region that Chameleolyser claims to be deleted (as opposed to samples without the deletion). For 5 of the remaining 7 deletions the reads that align onto the region corresponding to the deletion call are all read tails with a large number of non-matching bases. Read tails were blasted to the reference genome, showing that these read portions actually correspond to the region up or downstream of the actual deletion (Supplementary Figure 11). The 2 remaining deletion calls are in complex regions (the KIR gene cluster) and we could not unequivocally come to the same conclusion. By comparing the alignments with samples without the deletion call we found that there is a genetic event, but not necessarily a homozygous deletion. For that reason we conclude to have 13/15 deletion calls which correspond to LRS alignments.

All conversion calls were visually inspected in the LRS alignments for absence of a deletion or coverage drop. In addition, we checked for homozygous variants that correspond to the sequence differences between the linked donor and acceptor site. We confirmed this for all 8 ectopic gene conversion calls.

We downloaded exome and PacBio LRS data from https://github.com/genome-in-a-bottle/giab_data_indexes for 5 genome-in-a-bottle (GIAB) samples (HG002, HG003, HG004, HG005 and NA12878). The SNVs/Indels (not the result of gene conversions) were validated with the same procedure as described above for in-house generate validation samples.

Trio-validation

We considered all genomic sites with a minimal read depth of 60 in the masked alignment. For these sites, we counted the number of variants in the child for which the variant allele fraction in the mother and in the father is below 1%. These variants were considered *de novo*. If, on the other hand, the variant was identified in the father or mother (thus having a variant allele fraction above 15%), we considered this variant as inherited.

Comparison with other variant callers

SNVs/Indels (not the result of gene conversions)

GATK is run within the Chameleolyser method (cf. Mapping and short variant calling). DeepVariant 1.5.0 was run in a Docker container as described in the readme (https://github.com/google/deepvariant). To calculate the sensitivity of Chameleolyser, GATK and DeepVariant for SNVs/Indels (not the result of gene conversions) we first derived a set of high quality SNVs/Indels. To do so, we applied DeepVariant on 25 samples for which 30x High-Fidelity LRS alignments were available (20 in-house + 5 GIAB). All variant calls with a quality score > 30 which are present in the homologous regions that are used in this study (cf. Identification of paralogous regions) were considered true positive. Since all LRS samples were aligned to hg38 and exome data were mapped against hg19, a lift over of the coordinates was needed. This was done with CrossMap (Zhao et al., 2014). Because some of these true positive genetic variants might be present on genetic sites that are not or insufficiently covered in the exome experiment, we only considered variants for sensitivity analysis if the read depth for the corresponding base in exome data is ≥ 20. A variant is considered to be present in a zero mapping quality region if all reads that cover the respective position have mapping quality = 0.

Homozygous deletions and gene conversions

ExomeDepth and Conifer were applied on all exome samples in the study cohort. For ExomeDepth, capture target files were subdivided according to Parrish et al. 2017 (Parrish et al., 2017). Reference pools were created each consisting of 500 samples from healthy sex-matched individuals which were sequenced on the same sequencing machine and for which exome capture was done using the same enrichment kit. For Conifer, the same initial approach for reference pool selection was used. Here, bad quality reads (average quality score < 20) were removed from the samples. Next, the standard analysis steps were undertaken for both ExomeDepth and Conifer. All deletions in the output of ExomeDepth that overlap with a paralogous region and that have an observed/expected read ratio < 0.1 were

considered for further analysis. For Conifer, deletions with a SVD-ZRPKM ≤ -3 were selected. Next, in analogy with Chameleolyser (and thus for comparability), we removed all deletion calls with a cohort frequency > 10%. The remaining deletions that were present in the 20 samples for which we generated LRS data were used for the comparison between the CNV callers (Supplementary Figure 4). Sensitivity could not be assessed for deletions and gene conversions because of two reasons. 1) we cannot derive a set of true positive (or approximated by high-quality) events with a cohort frequency ≤ 10% because we only have a limited number of LRS samples available. This would be needed because the CNV calling within Chameleolyser is restricted to events ≤ 10% which is part of the method itself. 2) There are no tools to identify ectopic gene conversion events from LRS data. To estimate the number of true positive events among the deletion calls from ExomeDepth we first converted the regions from hg19 to hg38 by CrossMap. Only 151 regions could unambiguously be converted. If the number of LRS reads in the complete region corresponding to the deletion call is 20 or less, we presumed the region to be deleted. The reason for allowing 20 reads is that for the deletions which were manually validated by visual inspection up to 20 reads can be present in the deleted region as a result of suboptimal alignment of read endings (Technical validation, Supplementary Figure 11). Nevertheless by choosing a different threshold, the difference between Chameleolyser and ExomeDepth remains the same.

Variants of clinical interest

OMIM disease gene annotations were fetched from Ensembl Version 97 (MIM morbid 12/04/2019).

Variants with ambiguous positions

Exome-wide VCF files for all 41,755 WES samples were available through our in-house diagnostic pipeline (Radboud University Medical Center). This includes read alignment with bwa-mem 0.5.9-r16 and variant calling with GATK 3.2-2. In order to only retain high-quality variants, we filtered variants based on GATK's quality score: only substitutions with a GATK quality score ≥ 300 and indels with a quality score ≥ 1,000 were taken into consideration (Khazeeva et al., 2022; Xicola et al., 2019).

P value calculations

The p-values in the paragraphs 'Validation' and 'Variants with ambiguous positions' are derived from a binomial test in R (two sided). All other statistical tests in this manuscript are chi-squared tests, corrected for multiple hypothesis testing per series (Bonferroni's method; R 3.5.1).

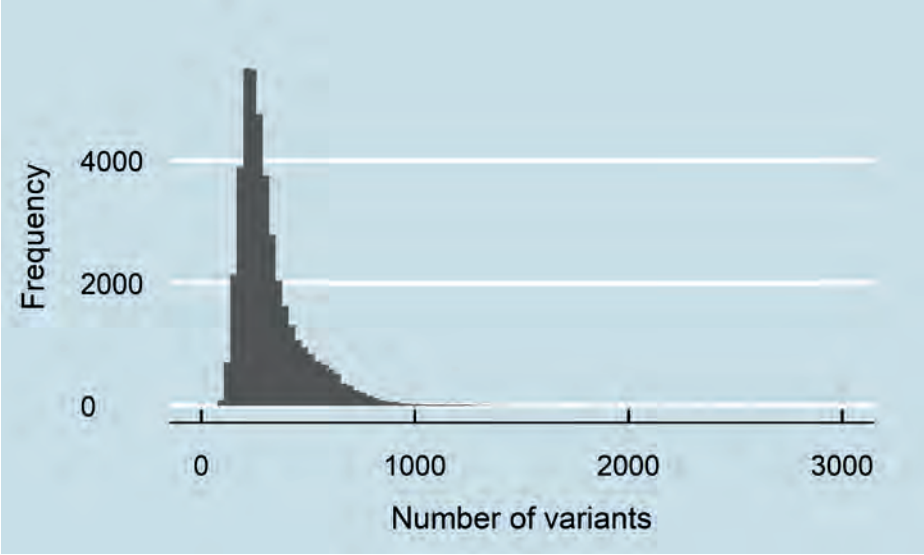
Data availability

The validation data generated in this study have been deposited in EGA under accession codes EGAS00001006479 (long-read genome sequencing for individuals with biobank consent (https://ega-archive.org/studies/EGAS00001006479)) and EGAS00001007513 (*STRC* amplicon sequencing (https://ega-archive.org/studies/EGAS00001007513)). These datasets are available under restricted access. Re-use of the data will be evaluated by a data access committee whether the proposed re-use is in line with the consent. Supplementary table 1 describes the mapping between the EGA sample identifiers and the identifiers that were used in this manuscript. The data onto which Chameleolyser is applied in this study is collected through routine genetic investigation. A diagnostic laboratory can use (de-identified) samples from archived clinical samples to validate and implement novel diagnostic assays. The derived clinically relevant variants can be shared, but in absence of explicit data sharing consent at individual patient level, complete FASTQ, BAM and VCFs cannot be disclosed unless specifically consented to by individual patients. These methods are also in accordance with relevant guidelines and regulations and approved by the institutional review board of the Radboud University Medical Center (2020-7142) and the Declaration of Helsinki. Source data are provided with this paper. The processed data that support the findings of this study are available as Supplementary data 1–26. The genome-in-a-bottle data used in this study are publicly on NCBI (URLs available on GitHub (https://github.com/genome-in-a-bottle/giab_data_indexes)) and/or the PacBio cloud (https://downloads.pacbcloud.com/public/). A list of download URLs per sample is also available as Supplementary Note 1. Source data are provided with this paper.

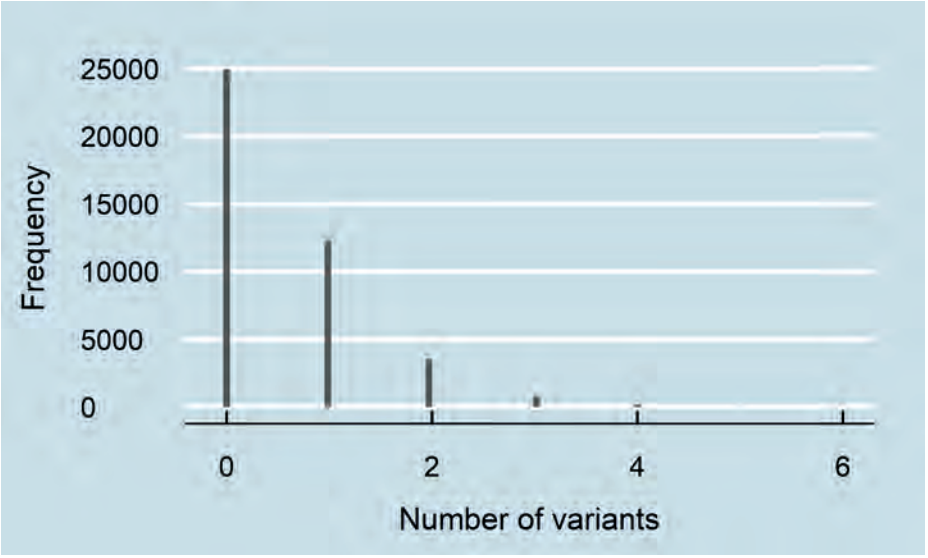
Code availability

The tool (Chameleolyser) as well as all other code that was used to produce tables and figures is available on GitHub (https://github.com/Genome-Bioinformatics-RadboudUMC/Chameleolyser; Steyaert, 2023).

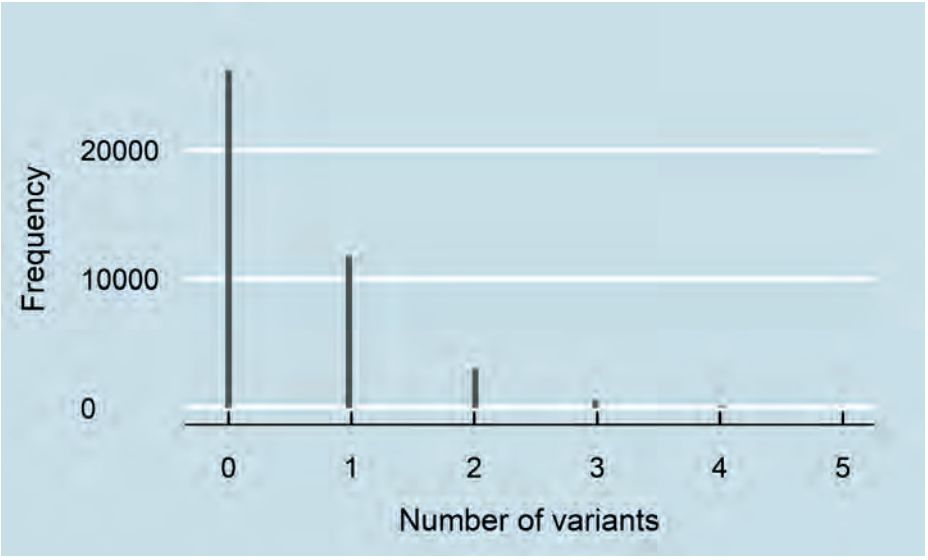
Supplementary Information



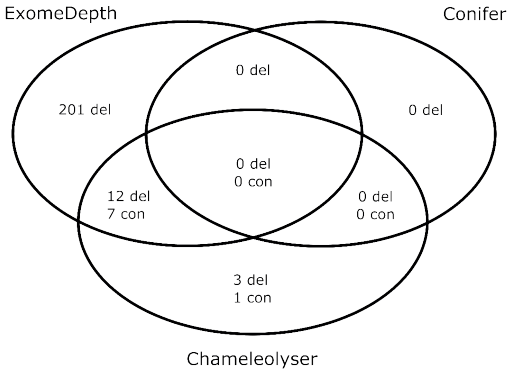
Supplementary figure 1: Distribution of the number of SNVs/Indels (not the result of an ectopic gene conversion) per individual in the study cohort (n=41,755). The horizontal axis indicates the number of SNVs/Indels (not the result of an ectopic gene conversion) that were identified per individual. The vertical axis represents the number of individuals with that number of variants identified. Clearly, the number of SNVs/Indels that were identified is between 0 and 1,000 for almost all individuals. Source data are provided as a Source Data file.



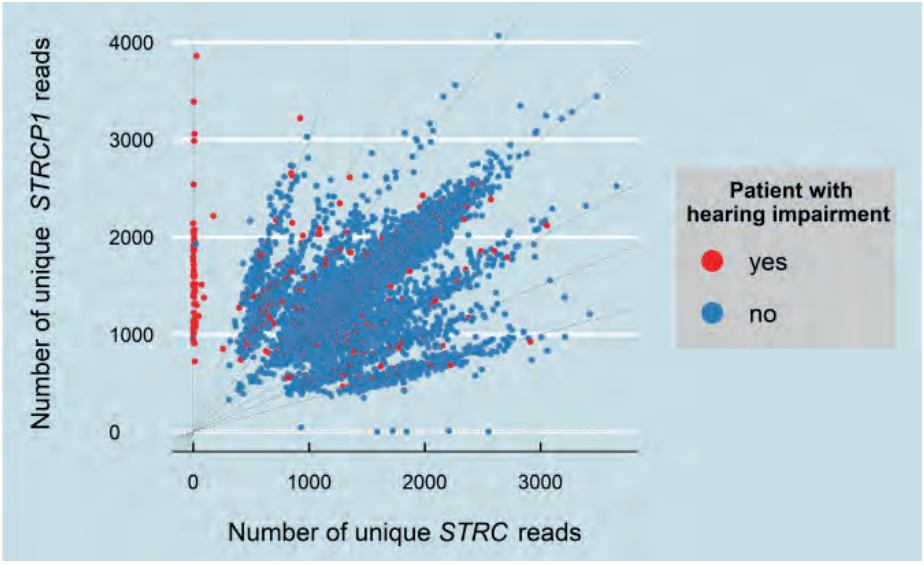
Supplementary figure 2: Distribution of the number of homozygous gene conversions per individual in the study cohort (n=41,755). The horizontal axis indicates the number of homozygous gene conversions that were identified per individual. The vertical axis represents the number of individuals with that number of gene conversions identified. In more than half of the studied individuals we did not identify a homozygous gene conversion. Source data are provided as a Source Data file.



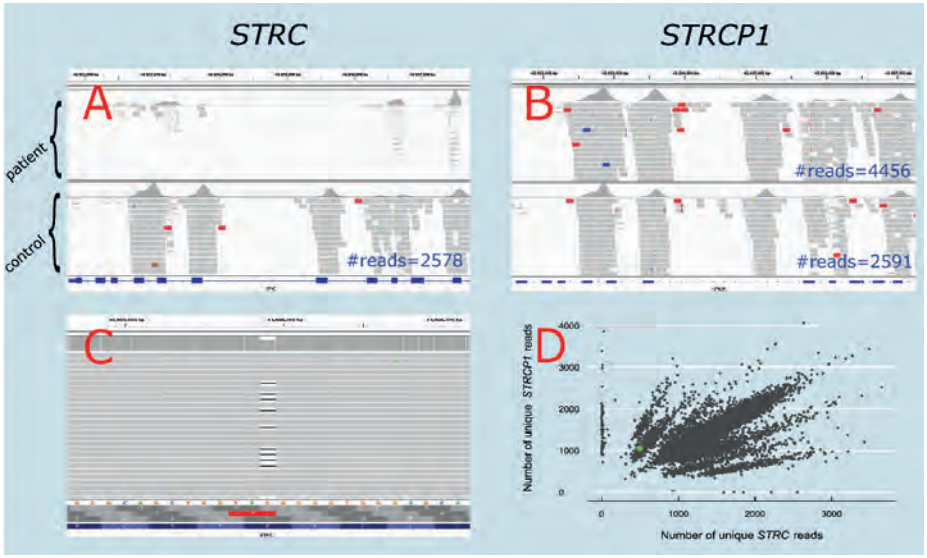
Supplementary figure 3: Distribution of the number of homozygous deletions per individual in the study cohort (n=41,755). The horizontal axis indicates the number of homozygous deletions that were identified per individual. The vertical axis represents the number of individuals with that number of gene conversions identified. In more than half of the studied individuals we did not identify a homozygous deletion. Source data are provided as a Source Data file.



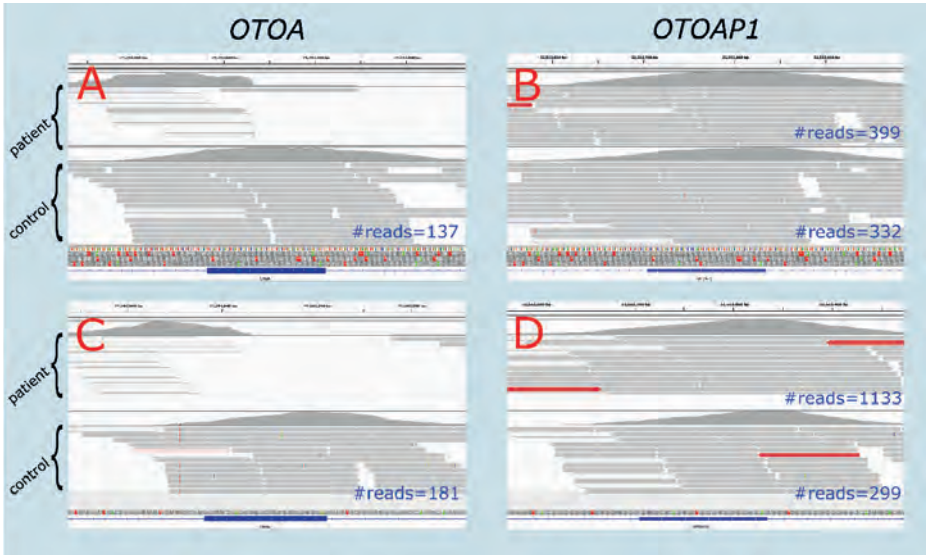
Supplementary figure 4: Comparison between Chameleolyser, ExomeDepth and Conifer for the identification of homozygous deletions and gene conversions within the paralogous regions of 20 validation samples. The Venn diagram shows the number of homozygous deletions (del) and the number of homozygous ectopic gene conversions (con) that are identified in the paralogous regions of the 20 exome samples for which LRS data was generated. The 7 conversions in the intersection between ExomeDepth and Chameleolyser are called as homozygous deletions by ExomeDepth.



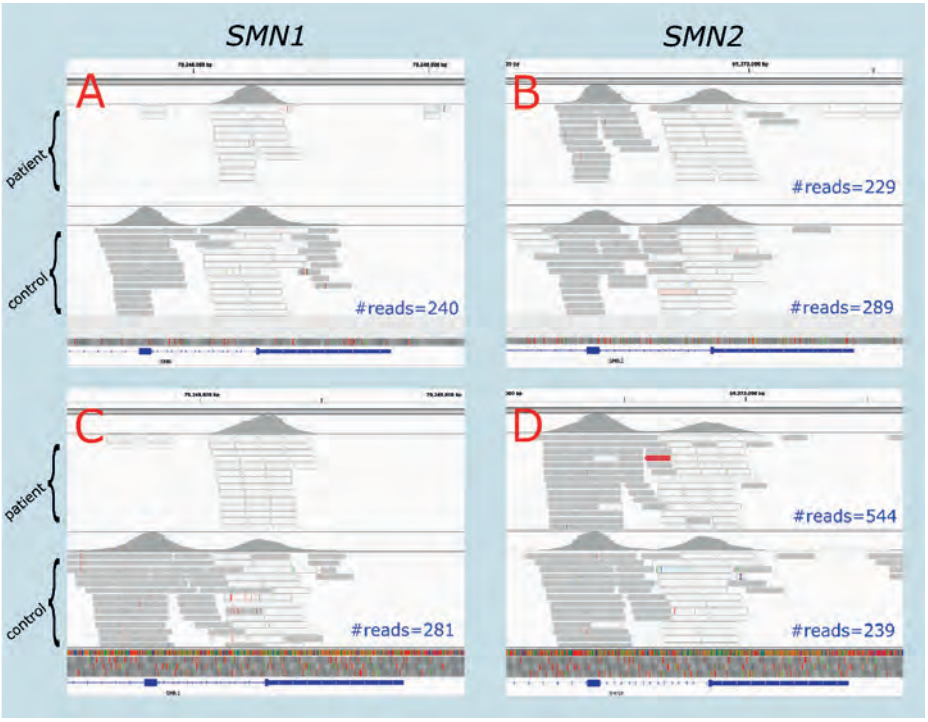
Supplementary Figure 5: The number of unique *STRC* versus *STRCP1* reads. The horizontal axis represents the absolute number of uniquely aligned reads onto the last 6 exons of *STRC* (ENST00000450892). The vertical axis corresponds to the number of reads that uniquely align onto the homologous region of *STRCP1*. Each dot is an individual in the study cohort. The visualization is illustrative for the number of *STRC* and *STRCP1* copies each individual has. Patients with hearing impairment are coloured red. Point clouds are formed due to the (discrete) genetic nature of the events. Source data are provided as a Source Data file.



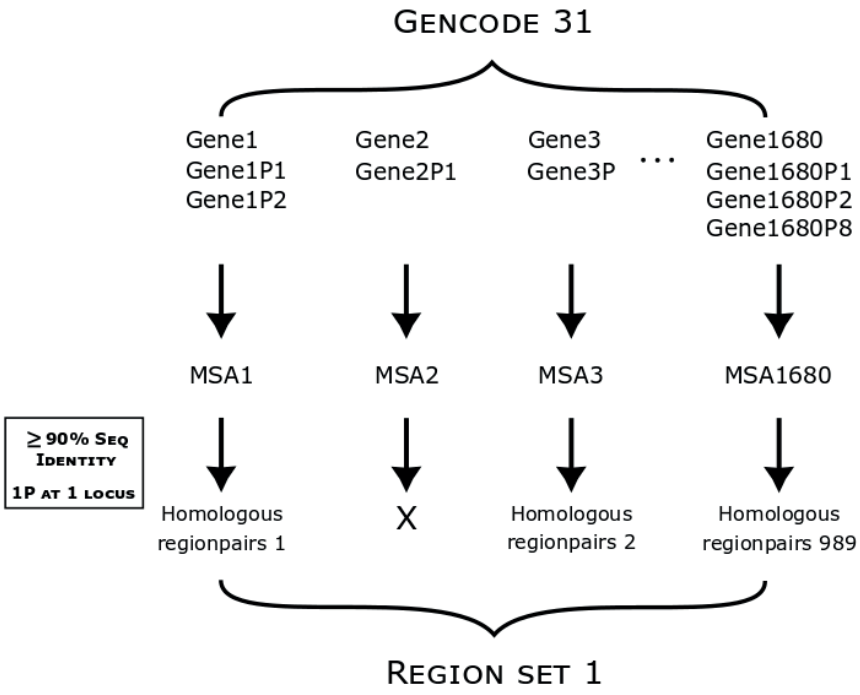
Supplementary Figure 6: Examples of pathogenic *STRC* variations identified by Chameleolyser. Panels A and B illustrate a pathogenic homozygous ectopic gene conversion that was identified in *STRC*. Reads and a coverage track for both a patient with hearing impairment (SAMPLE_23606) and a control individual are displayed for the affected *STRC* exons (panel A) and for the homologous exons in *STRCP1* (panel B). The absolute number of reads that align to the reference sequence in the displayed window (shown in purple) is highly indicative for the presence of an ectopic conversion from *STRCP1* to *STRC* in the patient. Panels C and D illustrate the pathogenic variants in SAMPLE_38648. In panel C a single nucleotide deletion is shown in the masked alignment. Since reads from both *STRC* and *STRCP1* are aligned to the same locus, a heterozygous *STRC* variant has a variant allele fraction of roughly 0.25. Panel D is a replicate of supplementary figure 5, but here we coloured SAMPLE_38648 in green. Clearly, the patient is present in the point cloud corresponding to 1 *STRC* allele and 2 *STRCP1* alleles (a heterozygous *STRC* deletion).



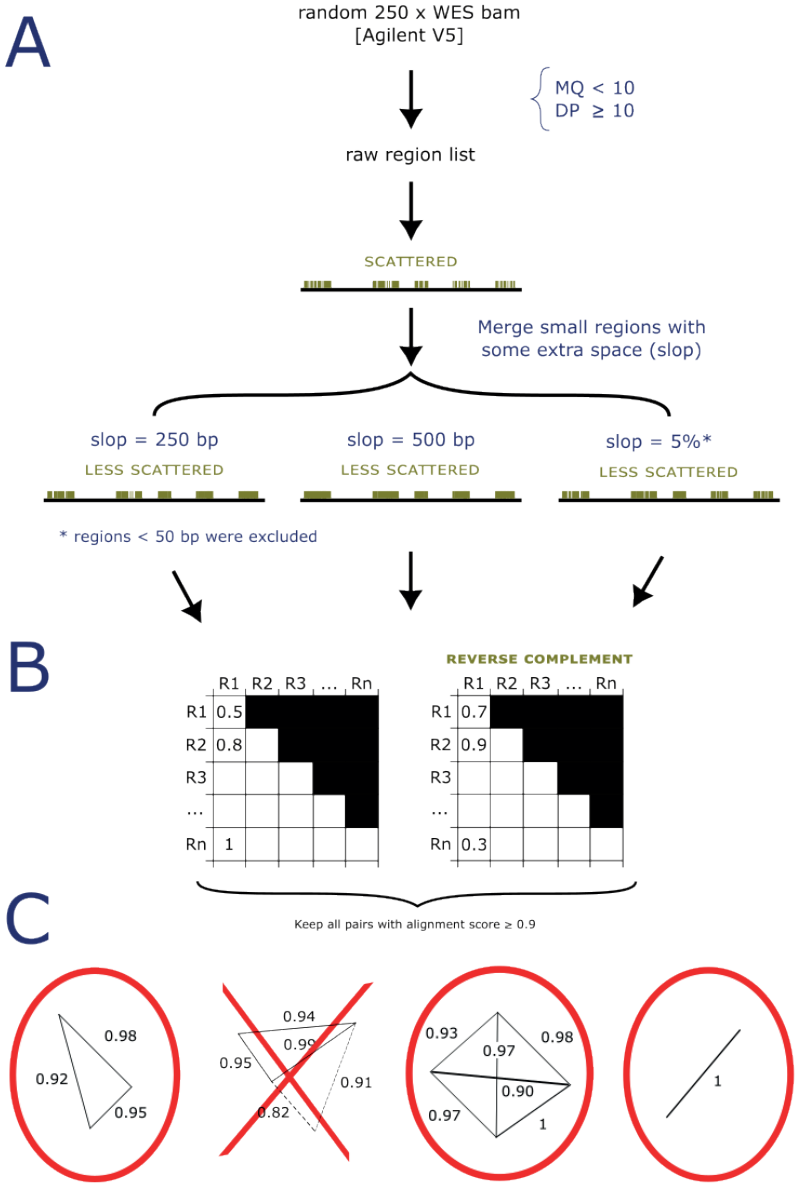
Supplementary Figure 7: Examples of pathogenic *OTOA* variations identified by Chameleolyser. Panels A and B illustrate a pathogenic homozygous deletion that was identified in *OTOA*. Reads and a coverage track for both a patient with hearing impairment (SAMPLE_37062) and a control individual are displayed for the affected *OTOA* exon (panel A) and for the homologous exon in *OTOAP1* (panel B). The absolute number of reads that align to the reference sequence in the displayed window (shown in purple) is highly indicative for the presence of a homozygous deletion in the patient (no reads aligning onto the *OTOA* exon whereas the number of reads aligning onto the *OTOAP1* exon is comparable between the patient and the control). In panels C and D the same type of IGV screenshots are shown, but here the pathogenic event is a homozygous ectopic gene conversion which is strongly suggested by the absolute number of reads that are aligned onto *OTOAP1* in the patient (SAMPLE_24323).



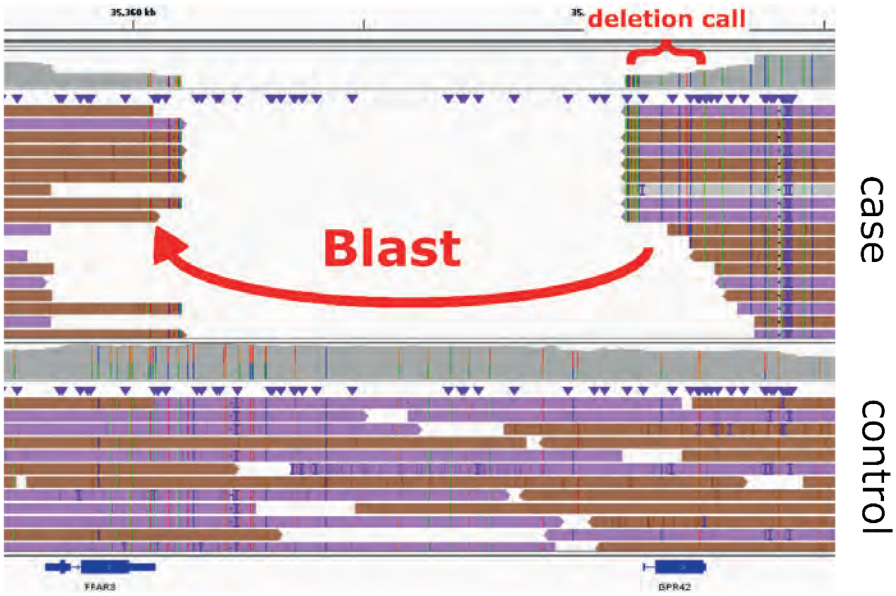
Supplementary figure 8: Examples of pathogenic *SMN1* variations identified by Chameleolyser. Panels A and B illustrate a pathogenic homozygous deletion that was identified in *SMN1*. Reads and a coverage track for both a patient with spinal muscular atrophy (SAMPLE_31987) and a control individual are displayed for the affected *SMN1* exons (panel A) and for the homologous exons in *SMN2* (panel B). The absolute number of reads that align to the reference sequence in the displayed window (shown in purple) is highly indicative for the presence of a homozygous deletion in the patient (no reads aligning onto the *SMN1* exon whereas the number of reads aligning onto the *SMN2* exon is comparable between the patient and the control). In panels C and D the same type of IGV screenshots are shown, but here the pathogenic event is a homozygous ectopic gene conversion which is strongly suggested by the absolute number of reads that are aligned onto *SMN2* in the patient (SAMPLE_20848).



Supplementary figure 9: Graphical summary of the derivation of region set I. This figure is a graphical representation of the derivation of region set I: regions in protein coding genes with known pseudogenes (Methods).



Supplementary figure 10: Graphical summary of the derivation of region set II. This figure is a graphical representation of the derivation of region set II: regions that are associated with low mapping qualities (Methods). Panel A illustrates how low mapping quality regions were found in exome datasets. Panel B displays how these regions were pairwise aligned in order to identify sets of paralogous regions. Panel C shows how these initial sets of paralogous sequences were filtered to obtain reliable groups of paralogs.



Supplementary figure 11: Read alignment within deleted region. LRS alignments (and coverage track) are displayed for 2 individuals, a case and a control. In the case, as opposed to the control, there is a homozygous deletion. Chameleolyser called this homozygous deletion with the genomic region corresponding to the brace. The intergenic region between the 2 protein coding genes (*FFAR3* and *GPR42*) could not be included in the deletion call since Chameleolyser starts from WES data. Based on the LRS alignments, at first sight, the *GPR42* gene is not homozygously deleted since > 10 reads align onto it. However, the reads that align onto this gene have a large number of non-matching bases and the sequence perfectly corresponds to the *FFAR3* gene (*i.e.* without non-matching bases). Based on this we could conclude that the read alignment is suboptimal and that the region that Chameleolyser claims to be deleted is indeed deleted.

Supplementary Note 1. URLs for the genome-in-a-bottle data used in this study

HG002

- LRS: https://downloads.pacbcloud.com/public/revio/2022Q4/HG002-rep1/analysis/
- WES: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG002-EEogPU_v02-KIT-Av5_AGATGTAC_L008.posiSrt.markDup.bam

HG003

- LRS: https://downloads.pacbcloud.com/public/revio/2022Q4/HG003-rep1/analysis/
- WES: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG003-EEogPU_v02-KIT-Av5_TCTTCACA_L008.posiSrt.markDup.bam

HG004

- PacBio: https://downloads.pacbcloud.com/public/revio/2022Q4/HG004-rep1/analysis/
- WES: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG004-EEogPU_v02-KIT-Av5_CCGAAGTA_L008.posiSrt.markDup.bam

HG005

- LRS: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/analysis/PacBio_CCS_15kb_20kb_chemistry2_12072020/HG005/
- WES: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/ChineseTrio/HG005_NA24631_son/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG005-EEogPU_v02-KIT-Av5_CGCATACA_L008.posiSrt.markDup.bam

NA12878/HG001

- LRS: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/analysis/PacBio_CCS_15kb_20kb_chemistry2_042021/
- WES: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/Nebraska_NA12878_HG001_TruSeq_Exome/NIST-hg001-7001-b-ready.bam

Supplementary Table 1: Sample identifier mapping between EGA and this manuscript for HiFi genome sequencing validation samples

Manuscript sample identifier	EGA sample identifier
SAMPLE_17545	EGAN00003613106
SAMPLE_17971	EGAN00003613113
SAMPLE_17973	EGAN00003613116
SAMPLE_17972	EGAN00003613117
SAMPLE_18244	EGAN00003613114
SAMPLE_18246	EGAN00003613115
SAMPLE_18245	EGAN00003613110
SAMPLE_18601	EGAN00003613111
SAMPLE_18603	EGAN00003613108
SAMPLE_18602	EGAN00003613109
SAMPLE_14580	EGAN00003613119
SAMPLE_14579	EGAN00003613121
SAMPLE_15442	EGAN00003613123
SAMPLE_15444	EGAN00003613105
SAMPLE_15443	EGAN00003613107

Chapter 3

Genomic Reanalysis of a Pan-European Rare Disease Resource Yields >500 New Diagnoses

Steven Laurie[a], Wouter Steyaert[a], Elke de Boer[a], Kiran Polavarapu[a], Nika Schuermans[a], Anna K. Sommer[a], German Demidov[b], Kornelia Ellwanger[b], Ida Paramonov[b], Coline Thomas[b], Stefan Aretz, Jonathan Baets, Elisa Benetti, Gemma Bullich, Patrick F. Chinnery, Jill Clayton-Smith, Giulia Coarelli, Enzo Cohen, Daniel Danis, Jean-Madeleine de Sainte Agathe, Anne-Sophie Denommé-Pichon, Jordi Diaz-Manera, Stephanie Efthymiou, Laurence Faivre, Marcos Fernandez-Callejo, Mallory Freeberg, José Garcia-Pelaez, Lena Guillot-Noel, Tobias B. Haack, Mike Hanna, Holger Hengel, Rita Horvath, Henry Houlden, Adam Jackson, Lennart Johansson, Mridul Johari, Erik-Jan Kamsteeg, Melanie Kellner, Tjitske Kleefstra, Didier Lacombe, Hanns Lochmüller, Estrella López-Martín, Alfons Macaya, Anna Marce Grau, Aleš Maver, Heba Morsy, Francesco Muntoni, Francesco Musacchia, Isabelle Nelson, Vincenzo Nigro, Catarina Olimpio, Carla Oliveira, Jaroslava Paulasová Schwabová, Martje G. Pauly, Borut Peterlin, Sophia Peters, Rolph Pfundt, Giulio Piluso, Davide Piscia, Manuel Posada, Selina Reich, Alessandra Renieri, Robin Wijngaard, Lukas Ryba, Karolis Šablauskas, Marco Savarese, Ludger Schöls, Leon Schütz, Verena Steinke-Lange, Volker Straub, Marc Sturm, Morris A. Swertz, Marco Tartaglia, Iris te Paske, Rachel Thompson, Annalaura Torella, Christina Trainor, Bjarne Udd, Liedewei Van de Vondel, Bart van de Warrenburg, Jeroen van Reeuwijk, Jana Vandrovcova, Antonio Vitobello, Janet Vos, Emílie Vyhnálková, Carlo Wilke, Doreen William, Jishu Xu, Burcu Yaldiz, Luca Zalatnai, Birte Zurek, Solve-RD DITF-GENTURIS, Solve-RD DITF-EURO-NMD, Solve-RD DITF-ITHACA, Solve-RD DITF-RND, Solve-RD consortium, Anthony J. Brookes[c], Teresinha Evangelista[c], Christian Gilissen[c], Holm Graessner[c], Nicoline Hoogerbrugge[c], Stephan Ossowski[c], Olaf Riess[c], Rebecca Schüle[c], Matthis Synofzik[c], Alain Verloes[c], Leslie Matalonga[c], Han G. Brunner[c], Katja Lohmann[d], Richarda M. de Voer[d], Ana Töpf[d], Lisenka E.L.M.Vissers[d], Sergi Beltran[d] , Alexander Hoischen[d].
These authors contributed equally: [a]Shared first; [b]shared second, [c]shared penultimate, [d]shared last

Nature Medicine. In press.

Abstract

Genetic diagnosis of rare diseases (RD) requires accurate identification and interpretation of genomic variants.

Clinical and molecular scientists from 37 expert centres across Europe created the Solve-RD resource encompassing clinical, pedigree, and genomic RD data (94.5% exomes, 5.5% genomes), and performed a systematic reanalysis for 6,447 previously undiagnosed individuals (3,592 male, 2,855 female), affected by RD from 6,004 families.

We established a genetic diagnosis in 506 (8.4%) families. Of 552 disease-causing variants, 464 (84.1%) were single nucleotide variants or short insertions/deletions. These variants were either located in recently published novel disease genes (n=67), recently reclassified in ClinVar (n=187), or reclassified by consensus expert decision within Solve-RD (n=210). Bespoke bioinformatics analyses identified the remaining 15.9% of causative variants (n=88). Ad-hoc expert review parallel to the systematic reanalysis diagnosed 249 (4.1%) additional families for an overall diagnostic yield of 12.6%.

The Solve-RD resource is open for the global RD community, allowing phenotype, variant and gene queries, as well as genome-wide discoveries.

Introduction

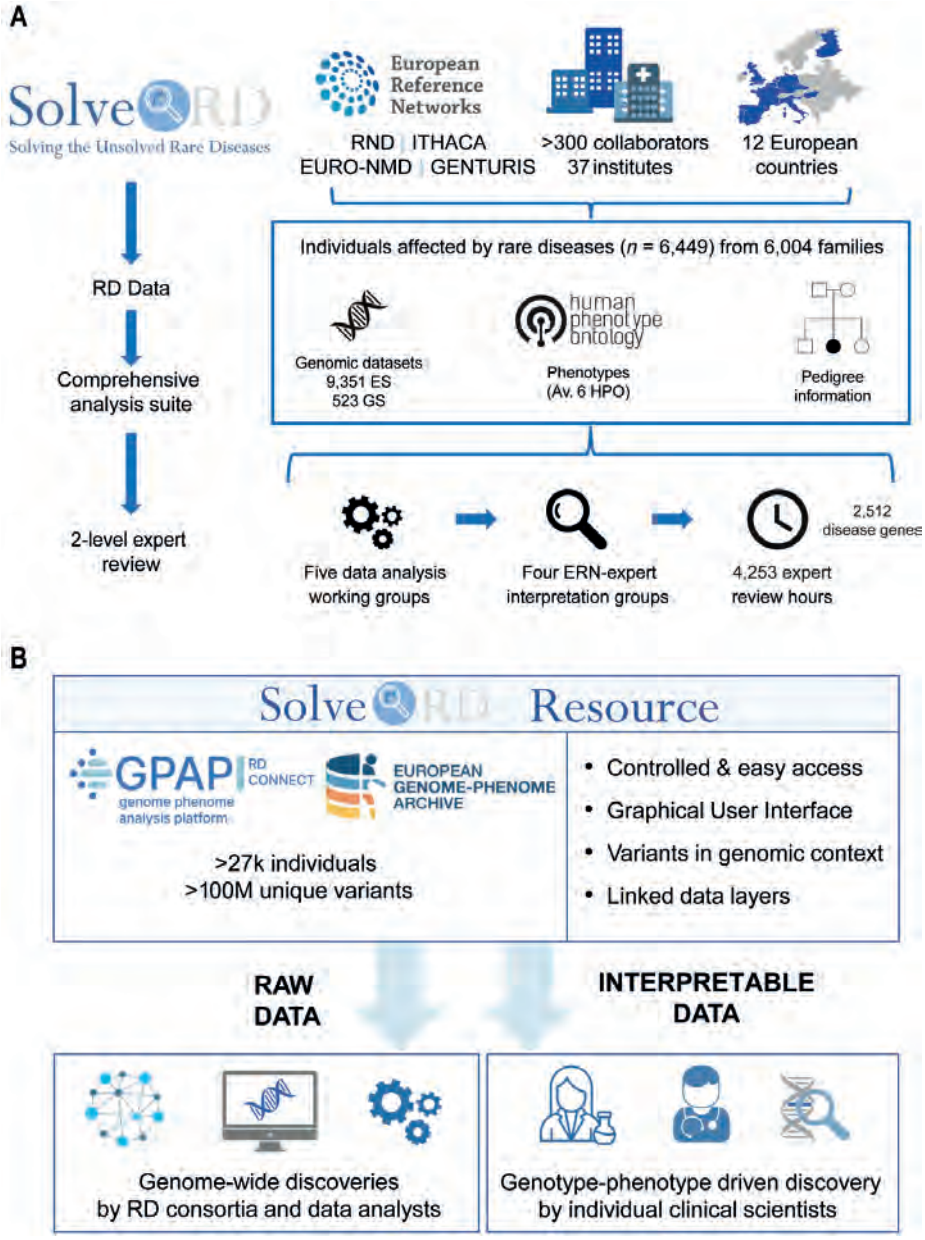
While the definition of what constitutes a rare disease (RD) is arbitrary, and thus varies by jurisdiction, the European Union has adopted a definition of an RD being an ailment that affects less than 50 individuals per 100,000. More than 70% of the >6,000 unique RDs are genetic, and collectively they constitute a major health issue, with 3.5-6% of individuals affected by a RD over their lifetime (Nguengang Wakap et al., 2020).

Despite improvements in diagnostics and research options for RDs, many individuals remain without a molecularly proven genetic diagnosis. In healthcare systems, where exome sequencing or genome sequencing are becoming standard of care, the diagnostic yield varies between 20-70% depending on the type of RD, inclusion criteria, sequencing strategy, and analysis standards (Smedley et al., 2021; Turro et al., 2020; Wright et al., 2023).

As reviewed in Dai et al (Dai et al., 2022), it has been shown that reanalysis of existing genomic data can lead to novel diagnoses, both as a result of newly described disease genes and due to improvements in the identification, annotation, and interpretation of genomic variants. However, reanalysis of such data is not routinely undertaken due to the time and multidisciplinary expertise required, and associated costs.

In 2017, the European Union brought together expertise on RDs into 24 thematic European Reference Networks (ERNs). Each ERN has multiple national centres across the 27 member states, which have all been vetted for their clinical, diagnostic and research expertise. These collaborations provide a pan-European framework to improve care for individuals with RDs.

Solve-RD is a pan-European -omics project that brings together (i) clinicians, geneticists and translational researchers from four ERNs, including Rare Neurological Diseases (RND, https://www.ern-rnd.eu/), Intellectual disability, TeleHealth, And Congenital Anomalies (ITHACA, https://ern-ithaca.eu/), Neuromuscular Diseases (EURO-NMD, https://ern-euro-nmd.eu/), and Genetic Tumour Risk Syndromes (GENTURIS, https://www.genturis.eu), as well as the Spanish Undiagnosed Disease Program (SpainUDP, https://spainudp.isciii.es/home); (ii) patient organisations represented by EURORDIS (Zurek et al., 2021) (https://www.eurordis.org/); (iii) genomic data sharing and analysis resources, such as the RD-Connect Genome-Phenome Analysis Platform (Laurie et al., 2022) (RD-Connect GPAP; https://platform.rd-connect.eu/) and the European Genome-Phenome Archive


A
SolveRD
Solving the Unsolved Rare Diseases
European Reference Networks
RND | ITHACA
EURO-NMD | GENTURIS
>300 collaborators
37 institutes
12 European countries
RD Data
Comprehensive analysis suite
2-level expert review
Individuals affected by rare diseases (n = 6,449) from 6,004 families
Genomic datasets
9,351 ES
523 GS
human phenotype ontology
Phenotypes
(Av. 6 HPO)
Pedigree information
Five data analysis working groups
Four ERN-expert interpretation groups
4,253 expert review hours
2,512 disease genes
B
Solve RD Resource
GPAP genome phenome analysis platform
RD CONNECT
EUROPEAN GENOME-PHENOME ARCHIVE
>27k individuals
>100M unique variants
Controlled & easy access
Graphical User Interface
Variants in genomic context
Linked data layers
RAW DATA
INTERPRETABLE DATA
Genome-wide discoveries by RD consortia and data analysts
Genotype-phenotype driven discovery by individual clinical scientists

< Figure 1: Overview of the Solve-RD analysis and interpretation framework and community resource established. A) Solve-RD brought together rare disease (RD) data and expertise. Central to Solve-RD are four core European Reference Networks (ERNs) on RD; via these expert disease networks, RD patients were recruited from 43 research groups from 37 institutes from 12 European countries (Belgium, the Czech Republic, Finland, France, Germany, Hungary, Italy, the Netherlands, Portugal, Slovenia, Spain, the United Kingdom) and Canada. The work involved >300 collaborators in the submission, analysis and interpretation of RD data. The RD-REAL framework allows sharing of data and expertise on a continental scale, consisting of a) expert curated data, b) a comprehensive analysis suite, c) two-level (i.e. molecular and clinical) expert review. The complete dataset comprises 9,645 individuals, from 6,004 families and includes phenotypes in Phenopacket format (average of six HPO terms per affected individual), pedigrees, and genomic data (genomes and exomes). Av., average. *For 419 families, disease causing SNVs or short insertions/deletions were identified; for 87 families disease-causing non-SNV variants were identified, including 3 cases of compound-heterozygosity involving an SNV and a CNV/SV. † In 114 of 147 cases where we could confirm the variant identified ad hoc, it would also have been found by the standard analysis B) Illustration of the utility of this resource to the global RD community. In total RD-REAL data of >23k individuals with >100M unique genomic variants are available via the RD-Connect GPAP and the EGA. This represents a growing resource, which contains data that has been submitted since the start of Solve-RD. The interpretable data (genetic variants, phenotypes and pedigrees) is standardised and annotated, and made available for querying, analysis, and interpretation in the RD-Connect GPAP for authorised users. In addition, all raw and processed data are available for download at the EGA under a controlled access model.

(Freeberg et al., 2022) (EGA, https://ega-archive.org/); (iv) European networks aiming to improve and harmonise the quality of genetic testing services, such as EuroGentest (http://www.eurogentest.org/); and (v) experts in the field of -omics technologies, bioinformatics, knowledge management and rare disease ontology such as the Orphanet Rare Disease Ontology (ORDO, https://www.orphadata.com/ontologies/) and the Human Phenotype Ontology (Köhler et al., 2021) (HPO).

One of Solve-RD's core aims is to improve the rate of genetic diagnosis for individuals affected by a RD (Graessner et al., 2021). A specific objective of Solve-RD is to systematically collate and reanalyse existing exome/genome datasets and corresponding structured ontology-based phenotype and pedigree information across the disease areas of its ERN partners (Figure 1). Here we report the results from the systematic reanalysis of data from 6,004 undiagnosed RD families recruited from across Europe by Solve-RD. The entire dataset is available as a resource for the global RD research community.

Results

A pan-European genomic data collection of individuals affected by rare diseases

Solve-RD involves over 300 clinicians, laboratory geneticists, and translational researchers from 43 research groups associated with 37 institutes located in 12 European countries and Canada. In total, we collected 10,276 genomic datasets, as well as phenotypic descriptions and pedigrees from 10,039 individuals, all previously analysed through local diagnostic or research efforts. The collection includes 554 genomes and 9,722 exomes enriched using 28 different exome-enrichment kits and generated on several short-read sequencing platforms. Following quality control (see Online Methods) 9,874 datasets (523 genomes, 9,351 exomes) from 9,645 individuals remained. These represent 6,449 individuals affected by RDs, and 3,196 unaffected relatives, from 6,004 families (Figure 1; Table 1; Supplementary Table S1). Disease categories comprise rare neurological diseases (RND, n=2,271 families), (multiple) malformation syndromes, intellectual disability, and other neurodevelopmental disorders (ITHACA & SpainUDP, n=1,857), rare neuromuscular diseases (EURO-NMD, n=1,517), and suspected hereditary gastric and bowel cancer (GENTURIS, n=359).

Phenotypic information was collected using standardised Human Phenotype Ontology (HPO) terms, with a median of six terms (range 0-74 terms) assigned per affected individual (Supplementary Figure S1), varying from a median of four terms for GENTURIS to ten for ITHACA, reflecting the phenotypic complexity of probands from the respective RD. In addition, for 2,126 (35.4%) probands, a clinical diagnosis was encoded using Orphanet ORPHAcode(Lagorce et al., 2024), of which 338 were unique.

Table 1: Solve-RD reanalysis data: Number of datasets after QC filtering (see Online Methods), representing the number of previously undiagnosed families/probands. Numbers are given for the entire project and for each European Reference Network (ERN) separately. We provide the overall yield of newly diagnosed RD cases for both the multi-centre systematic reanalysis, and the parallel ad hoc expert review. The table also indicates the number of (likely) pathogenic variants that led to 'candidate diagnoses'.

Solve-RDRD-REAL data	ERN RND	ERN ITHACA	ERNEURO-NMD	ERN GENTURIS	Sum across ERNs
Experiments (exomes/ genomes)	2852 (2,692/160)	4470 (4,231/239)	2162 (2,059/103)	390 (369/21)	**9,874**
Participants (affected individuals)	2,799 (2,453)	4,331 (1,933)	2,125 (1,685)	390 (378)	**9,645 (6,449)**
Families	2,271	1,857	1,517	359	**6,004**
Diagnosed probands (*systematic reanalysis*) [%]	242 [10.7%]	158 [8.5%]	96 [6.3%]	10 [2.8%]	**506 [8.4%]**
Diagnosed probands (*ad hoc* expert review) [%]	61 [2.7%]	145 [7.8%]	42 [2.8%]	1 [0.3%]	**249 [4.1%]**
Probands with 'candidate diagnoses' [%]	119 [5.2%]	139 [7.5%]	41 [2.7%]	45 [12.5%]	**344 [5.7%]**

>500 new genetic diagnoses upon systematic reanalysis

A two-level expert analysis strategy (data-expert and clinical-expert levels) was applied as detailed in the Online Methods. All datasets were reanalysed for a broad range of genomic variants, including Single Nucleotide Variants and short Insertions-Deletions (SNVs/InDels), non-canonical splice variants predicted in-silico, homoplasmic and heteroplasmic mtDNA variants, Copy Number Variants (CNVs), Structural Variants (SVs), Mobile Element Insertions (MEIs) and Short Tandem Repeat expansions (STRs) (Supplementary Figure S2). Each ERN generated a list of established disease genes for their respective conditions, resulting in gene lists ranging from 230 genes for GENTURIS to 1,820 genes for RND (Online Methods; Supplementary Table S2). Systematic reanalyses resulted in 506 new genetic diagnoses, by (likely) pathogenic variants that explained the phenotype, representing 8.4% of probands.

New molecular diagnoses

SNV/InDel reanalysis revealed 461 (likely) pathogenic variants, enabling a diagnosis in 419 families. In order to retrieve the 461 (likely) pathogenic SNV/InDel variants from the >50k prioritised variants, a total of 4.8 minutes was spent per variant (with an average of 9 variants per sample) on molecular and clinical expert review (Supplementary Table S3).

The 461 SNV/InDel variants identified, in 419 probands, consisted of 282 heterozygous variants with dominant effect, 85 homozygous and 76 compound heterozygous variants with recessive effect, and 18 hemizygous variants. Functionally, these represented 187 nonsense/frameshift variants, 249 missense variants, 11 in-frame deletions, ten splicing variants (eight intronic, and two synonymous variants), two 5′ UTR variants, one promoter region variant and one complex InDel variant (Figure 2; Supplementary Table S4). 41 of the 461 (9.1%) variants could be confirmed as *de novo* mutations, due to the availability of proband-parent trios for 1,320 (22%) families, primarily from ERN ITHACA (1,081).

We evaluated why the 461 SNV/InDel variants were not classified as disease-causing in prior analyses. We found that 67 of these variants affect genes which were established as a novel disease gene after data submission to Solve-RD (i.e. appeared in OMIM after 01/01/2018; Supplementary Figure S3, Supplementary Table S4) while the remaining 394 were in established disease genes at the time of data submission. Of these, 117 variants have been reclassified in the interim (i.e. novel or modified ClinVar entry since 2018), and 70 had initially been deemed not to fully explain disease, despite the variant being classified as pathogenic in ClinVar, as a result of perceived insufficient clinical concordance at the time. The remaining 207 variants were not in ClinVar and were only classified as (likely) pathogenic by the experts involved in this project.

We applied a suite of analysis tools for calling and annotating variants. These included queries for non-canonical splice variants, mtDNA variants, CNVs, SVs, MEIs, and STRs. These additional analyses yielded a diagnosis in 87 RD families through a total of 88 variants, with CNVs in 44 probands (45 variants) as the most prevalent variant type. This included three cases where biallelic pairings of an SNV with a CNV/SV formed a compound heterozygous variant, and one case where two CNVs affecting different genes led to a digenic diagnosis (Supplementary Figure S2; Supplementary Table S4).

The diagnostic yield across disease groups (i.e. ERNs) ranged from 2.8% (genetic tumour risk syndromes, GENTURIS) to 10.6% (rare neurological disorders, RND), with yields correlating with the number of established disease genes provided by the ERNs (Figure 2; Supplementary Table S2). Overall, for the 506 newly diagnosed probands, the inheritance pattern was autosomal dominant for 306 probands, autosomal recessive for 137, X-linked for 42, mitochondrial for 16, dual-diagnoses in four individuals, and digenic inheritance in one individual (Supplementary Table S4).

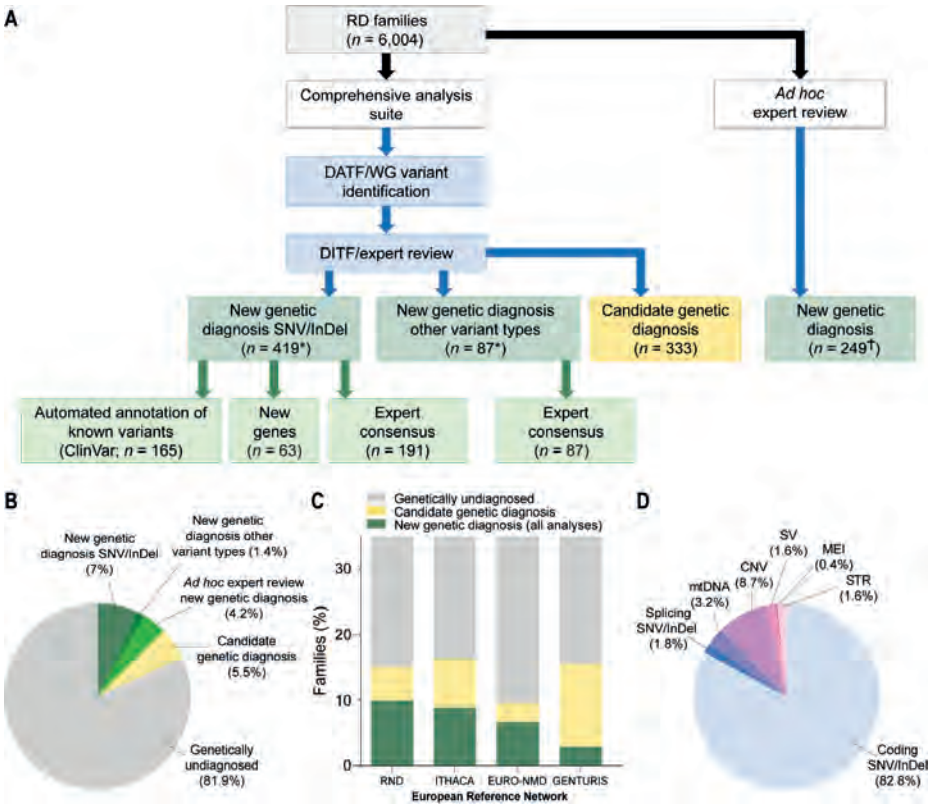


Figure 2: Systematic reanalysis of genomic datasets for the genetic diagnosis of RDs. A) Flowgram of systematic analysis of 6,004 families. Yield per analysis type (new genetic diagnoses by SNV/InDel and other variant types; candidate genetic diagnoses and new genetic diagnoses by ad hoc expert review) is shown. For SNV/InDels we evaluated why the 464 variants identified in 419 families had not been classified as disease-causing previously. B) Chart summarising the diagnostic yield across 6,004 families in Solve-RD. C) Chart summarising the yield per disease category (ERN), denominator is 6,004 families. D) Chart summarising the different variant types that led to a molecular diagnosis in 506 of 6,004 families as part of the systematic reanalysis effort of Solve-RD. Abbreviations: RD: rare disease; DATF: data analysis task force; WG: work group; DITF: data interpretation task force; SNV: single nucleotide variant; InDel: short insertions and deletions; mtDNA: mitochondrial DNA; CNV: copy number variant; SV: structural variant; MEI: mobile element insertion; STR: short tandem repeat; Splicing SNV/InDel: non-canonical splicing sites; RND: rare neurological diseases; ITHACA: rare malformation syndromes, intellectual and other neurodevelopmental disorders; EURO-NMD: rare neuromuscular diseases; GENTURIS: genetic tumour risk syndromes. *Three families were diagnosed due to a combination of an SNV/InDel and a variant identified through the "other variant type" analyses and are counted only under 'New genetic diagnosis other variant types'.

Next to the overall yield across the cohort, the importance of new diagnoses can be illustrated by individual RD cases and families, each benefitting from technical and interpretational improvements, leading to the closure of diagnostic odysseys. We present examples of previously missed SVs/CNVs in RND (*B4GALTNT1*) and GENTURIS (*APC*), missed mosaic *de novo* mutations in ITHACA (*PIK3CA*), as well as novel disease-gene associations (ITHACA, *MN1*) (for details see Case Reports in Supplementary Information).

One other example on how variant annotation pipelines can aid in variant interpretation is exemplified by the diagnostic path of a girl (P0012491) who was clinically suspected to have Rett syndrome (MIM#312750). WES performed in 2014 did not yield a diagnosis, despite specific attention for variants affecting *MECP2*, the gene associated with Rett syndrome. Almost eight years later, the re-analysis presented here uncovered a pathogenic *de novo MECP2* variant from the same data. Retrospective analysis of previous interpretation steps revealed that the variant was initially annotated to a less relevant isoform of *MECP2* (MECP2-e2; ENST00000303391.11), in which the variant located to an intron. However, reannotation here revealed that the variant truncates the brain-specific isoform of *MECP2* (MECP2-e1; ENST00000453960.7) and hence is indeed explanatory for the Rett syndrome in this girl.

Cases diagnosed by ad hoc expert review

During the course of Solve-RD, many contributing partners continued to performed analysis on specific families of interest, both locally and using the RD-Connect GPAP. This ad hoc expert review provided 249 additional diagnoses (4.1%), some of which have been included in individual reports (de Boer et al., 2021; Matalonga et al., 2021; Pauly, Brüggemann, et al., 2023; Pauly, Korenke, et al., 2023; Schüle et al., 2021; te Paske et al., 2021; Töpf et al., 2021) and novel disease gene discovery efforts (Kaplanis, Samocha, Wiel, Zhang, Retterer, et al., 2020; Weihl et al., 2023) published previously. Cases solved through ad hoc expert review were reported to Solve-RD and not interpreted further as part of the systematic reanalysis. For 197 (79%) of these ad hoc diagnoses the causative variants were SNVs. For 147 (75%) of these SNVs we could assess post hoc whether the variants would also have been identified by the systematic reanalyses performed. We found that in 114 of 147 (78%) cases the SNVs would have been identified, while the remaining cases were diagnosed due to the discovery of variants located in novel disease genes not included in the ERN gene lists, or initially discounted for technical reasons e.g. having insufficient coverage (<10 reads) or being deep intronic variants.

Candidate disease-causing variants

In addition to variants that were deemed causative for disease, we identified 378 further variants (in 333 affected individuals) in established disease genes that have not yet been confirmed as causative, either because the variant does not fully explain the individual's phenotype, or because the variant's pathogenicity cannot yet be conclusively determined (Figure 2; Supplementary Table S4).

Cross-ERN analysis, genetic recurrences and clinical actionability

Cross ERN de novo mutation analysis

Systematic reanalyses were performed by each of the four ERNs, thus maximising disease-specific expertise. Because the clinical spectrum may occasionally cross ERN boundaries, we assessed all *de novo* mutations across all genes included in any of the ERN gene lists (2,512 unique genes), irrespective of which ERN originally submitted the case. This led to a molecular diagnosis in an additional three probands through the identification of (likely) pathogenic *de novo* variants in *CSDE1* (Gangfuß et al., 2022), *EP300*, and *SYT1* (Supplementary Table S5; Supplementary Information), which would have been missed without this cross-ERN analysis.

Recurrent variants

We observed recurrence for 21 (likely) pathogenic variants, together accounting for 41 diagnoses (Supplementary Table S6). These 21 variants occurred in 18 genes, with three genes harbouring two different recurring variants (*SPG7*, *KCNA2*, and *SPAST*).

One of the recurring variants was identified across three ERNs: an identical *MT-ATP6* missense variant (chrM:9185T>C (ENST00000361899:c.659T>C (p.Leu220Pro))) was observed in five affected individuals from three unrelated families submitted by ERNs EURO-NMD, RND and ITHACA. The variant was observed with a heteroplasmy of 77% and 90% in the EURO-NMD and RND probands, respectively, while it was homoplasmic in the ITHACA proband, in line with the variable phenotypic presentation (Supplementary Table S7; Supplementary Information).

Beyond diagnosis: clinical actionability

We investigated the number of diagnosed individuals that would potentially benefit from therapy or other actionability by considering medications or interventions included in three databases: IEMbase (Ferreira et al., 2019), Treatabolome (Atalaia et al., 2020), ClinGen (Rehm et al., 2015), and international cancer guidelines.

We identified 73 affected individuals (14.4% of diagnosed individuals) that harboured variants in a potentially actionable gene (Supplementary Figure S4).

Implementation, and feedback to referring clinicians and eventually families and patients is following local guidelines which differ between centres. Actual actionability has already happened and is continuously ongoing. To date we have received feedback for a subset of the aforementioned cases, with details of 16 examples summarised in Supplementary Table S8.

An example from ERN EURO-NMD, is provided by the case of two young-adult patients from different families who had presented with limb-girdle muscle weakness and fatigability from 2 years of age, and subsequently developed ptosis and difficulty swallowing, leading to a suspected diagnosis of limb-girdle myasthenic syndrome. While prior ES analyses were negative, reanalysis within Solve-RD using SpliceAI (Jaganathan et al., 2019) led to the identification of a homozygous intronic variant with a potential splice donor effect, c.1023+5G>A proximal to the exon 5 / intron 5 junction of *DES* in both patients. In parallel, but outwith Solve-RD, a female with a similar phenotype, among a cohort of patients suspected of having congenital myasthenic syndrome (CMS) being treated in the same hospital, was also found to be homozygous for this mutation. Subsequent laboratory analyses indicated reduced production of normal desmin transcript and protein. Administration of the standard CMS treatment of Pyridostigmine and Salbutamol was initiated and while one of the two patients showed no improvement after 3 months, the other exhibited a 50% improvement in measures of fatigable weakness (for further details see Example Case Reports in Supplementary Information).

Discussion

Genomic data from RD cases which have been extensively analysed by experts in the past can still yield a large number of new diagnoses, with prior studies reporting success rates commonly in the range of 6-13% (Dai et al., 2022). We previously reported on preliminary ClinVar focussed reanalyses undertaken within Solve-RD which resulted in molecular diagnoses being provided for 111 families (Denommé-Pichon et al., 2023; Matalonga et al., 2021). The value of an in-depth systematic reanalysis is supported by our success in newly diagnosing 8.4% of affected individuals through our systematic reanalysis, and the further 4.1% diagnosed in parallel by local reanalysis in individual centres through ad hoc expert review. In total, we have achieved 12.6% new diagnoses to date. While a few recent studies

have reported higher diagnostic rates following reanalysis, ranging from 15-21% (Baker et al., 2019; Bullich et al., 2022; Liu et al., 2019; Wright, McRae, et al., 2018), it should be noted that those data sets were more homogeneous in nature, usually originated from one country, and were of substantially smaller scale and breadth, ranging from 240-1,684 probands, facilitating interpretation and diagnosis. Nevertheless our diagnostic yield is in the top end of the typical range of that found across the 29 reanalysis studies described in Dai et al (Dai et al., 2022). All expert-curated (likely) pathogenic variants are being uploaded to ClinVar (Landrum et al., 2020).

The example of Solve-RD is instructive because it provides a real-world investigation of the current diagnostic potential of systematic reanalysis of exome and genome sequencing data for rare diseases at scale (Baker et al., 2019; Bullich et al., 2022; Liu et al., 2019; Wright et al., 2018). As such, comprehensive reanalyses are typically infeasible for individual centres, but will lead to many new diagnoses. Implementation of initiatives similar to Solve-RD at a national or transnational level may prove beneficial and cost-effective.

Limitations

The data presented here, pertain to only four RD domains, i.e. rare neurological disease, intellectual disability and malformation syndromes, rare neuromuscular disease, and hereditary tumour risk syndromes. It remains to be seen if other types of rare disease will have the same return of new diagnoses by systematic reanalysis. Nonetheless, the approach that we applied in Solve-RD is generic and can easily be implemented across all 24 ERNs representing the full gamut of rare diseases.

Further, the previously generated exome and genome sequencing data was highly heterogeneous since this is a pan-European project aiming to provide diagnoses for individuals across Europe. This heterogeneity, both in terms of the quality of the historic ES data, and the breadth of phenotypic descriptions, impacted upon our ability to confidently identify potentially pathogenic variants. Another limitation was that for two-thirds of the families analysed (4103/6004), we only had sequencing data from the affected proband, thus limiting supporting segregation information during downstream variant interpretation, especially with respect to the identification of pathogenic *de novo* variants.

Key findings

After more than a decade of diagnostic exome sequencing (de Ligt et al., 2012; Rauch et al., 2012), our knowledge of the spectrum of genes and variants causing

monogenic rare disease is clearly still steadily increasing. This is exemplified by the large number (n=394) of SNV/InDel variants that could now be correctly interpreted, based on the availability of new gene-level or variant-level information.

Our finding that systematic reanalysis yielded many new diagnoses also reflects the fact that analysis pipelines and methods are not harmonised across countries and centres. Consequently, a substantial proportion (15.9%; n=87) of novel diagnoses were a result of individually rare variant types that are not SNV/InDels, many of which were not detectable by standard diagnostic bioinformatics pipelines.

With increasing size of RD datasets, we shall identify an increasing number of identical variants in multiple individuals. This is evident here, as we identified 21 (likely) pathogenic variants that occurred two or three times across a total of 41 unrelated probands from the 6,004 families analysed, sometimes straddling different clinical disease categories.

The proposed framework, RD-REAL (see Online Methods for more information) with its two-level expert review represents a practical blueprint for reanalysis efforts on a global scale and in part overcomes heterogeneity in data type and origin. The amount of time that was invested in expert reanalysis was manageable at 4.8 minutes per variant, or 42.8 min on average per proband. In Europe, the existing organisation of rare disease expertise in 24 ERNs should enable the implementation of the Solve-RD approach to the full gamut of Rare Diseases.

Value of a genetic diagnosis

Providing a diagnosis to individuals affected by a rare condition and their families brings an end to their diagnostic odyssey, often after many years. It has been shown that parents place intrinsic value on information and knowledge regarding their child's condition, suggesting that changes in clinical outcomes alone may not be a perfect measure of the full benefit of diagnostic tests (Marshall et al., 2019).

We examined clinical actionability for the new diagnoses in the series, using a definition that only considered approved medication or (preventive) interventions. This is a more restrictive definition than that applied in a previous study (Smedley et al., 2021), which also considered informed future reproductive choice and additional surveillance of other family members. Even with this limited definition, there was potential for medical actionability in 14.4% of those receiving a diagnosis in our series, with ongoing implementation and the first concrete examples shown in Supplementary Table S8.

Data and tools for the entire RD and genomics community

Global data sharing is essential for the discovery of rare disease genes (Rehm, 2022). Increasingly, novel gene discovery is driven by very large datasets (Fu et al., 2022; Zhou et al., 2022) that cover specific disease-domains, such as neurodevelopmental disorders or autism spectrum disorder (ASD). In some cases, the analysis may be restricted to a specific type of variant such as *de novo* mutations (Kaplanis, Samocha, Wiel, Zhang, Retterer, et al., 2020). Ideally, rare disease resources should contain complete phenotypic and genomic datasets in an accessible format. One such successful resource is the Simon-Simplex Collection (Fischbach & Lord, 2010) (SSC), which has detailed phenotype and genotype data (*e.g.* SFARI | Simons Simplex Collection), for 2,600 individuals affected by ASD and their unaffected parents.

Solve-RD offers a resource that comprises a wider range of phenotypes and diseases across four disease-domains for 6,004 families affected by a RD. We anticipate this resource will continue to grow as part of Solve-RD and beyond (Figure 1). Authorised users can use either the RD-Connect GPAP to search and analyse phenotype (HPO, ORPHAcodes), gene and variant level data, or the EGA to access all data simultaneously. The detection of gene-level recurrence in other individuals globally affected by a rare condition is further facilitated through connection to the MatchMaker Exchange network (Boycott et al., 2022).

In line with the successful demonstration of the reanalysis undertaken here and throughout the Solve-RD project, several countries, such as Germany, have recently built infrastructure and implemented formal processes to facilitate periodic reanalyses of data from undiagnosed RD patients (https://translate-namse.de/). Furthermore, the tools and infrastructure developed within Solve-RD have been adopted as the core framework for undiagnosed RD case reanalysis within the ERDERA project, which aims to extend out to all 24 ERNs, and reanalyse over 100,000 datasets from rare disease families across all disease types (https://www.ejprarediseases.org/erdera/).

Outlook

Our current effort focussed on diagnoses in established RD genes. However, this resource and the datasets in Solve-RD should be well suited for the generation of continued insights. This effort will allow the community to continue to make new diagnoses. As an example, since the systematic analysis presented her was completed, we have already promoted 2 SVs and 7 CNV from candidate to disease causing (Demidov et al. (Demidov et al., 2023, 2024)) and an additional 10 SNV/InDel variants (Supplementary Table S9). This resource shall also allow the discovery

of novel disease genes or loci, and discovery of new disease mechanisms and causes is an ongoing part of Solve-RD (Graessner et al., 2021; Zurek et al., 2021). As a first example, the most recent discovery by Genomics England of variants in the non-coding RNA gene *RNU4-2* that cause a complex NDD phenotype (Y. Chen et al., 2024; Greene et al., 2024) led to one further solved case in Solve-RD (P001996), in addition to the Solve-RD case (P0007197) which contributed to the original discovery, Supplementary Table S9; Figure 3, Panel I). As a further example, we would like to highlight the gene *RAB14*, member RAS oncogene family (*RAB14*), which encodes a protein involved in intracellular membrane trafficking during early embryonic development. Although the importance of *RAB14* for neurodevelopment had already been suggested by a statistically significant enrichment of *de novo* variants in a developmental disorder cohort in 2020 (Kaplanis, Samocha, Wiel, Zhang, Retterer, et al., 2020), the associated neurodevelopmental phenotype remains to be fully characterised. The Solve-RD dataset includes data from two male individuals with neurodevelopmental phenotypes harbouring *de novo* variants in *RAB14*, thus enabling clinical characterisation as a result of the comprehensive HPO description collected as part of this effort (Figure 3, Panels G & H). Thus, for *RAB14*, the Solve-RD resource serves as a starting point for establishing a new genotype-phenotype association, and accordingly, many additional genotype-phenotype and/or mechanistic studies have been initiated from the Solve-RD datasets and are currently followed up within the Solve-RD RDMM-Europe initiative (see Model matchmaking via the Solve-RD Rare Disease Models & Mechanisms Network (RDMM-Europe); Ellwanger et al., Nature Lab Animal, in press).

New genomic technologies such as optical mapping and long-read genome sequencing will undoubtedly add novel molecular mechanisms of rare disease in the future (Beyter et al., 2021; Cohen et al., 2022; Mantere et al., 2019; Merker et al., 2018; Sabatella et al., 2021; Te Paske et al., 2022). Similarly, combinations of genomic data with other -omics technologies may lead to further improvements in the RD diagnostic yield (Cummings et al., 2017; Wortmann et al., 2022; Yépez et al., 2022). It is also likely that developments in AI will also have a significant impact (Cheng et al., 2023).

Expansion to other types of RD through their respective ERNs or future international data sharing initiatives will further enhance the Solve-RD resource for the global RD research community.

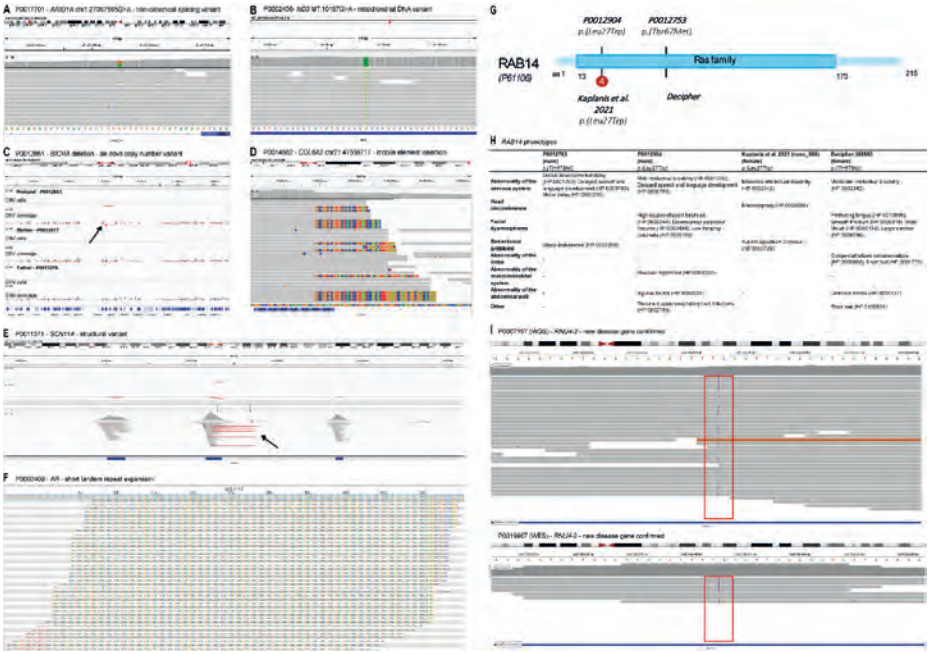


Figure 3: Examples of 'beyond standard' variant types, and new discoveries by Solve-RD. Panels A-F provide illustrative examples of previously unsolved RD probands for which a new variant other than a coding SNV/InDel resulted in a new diagnosis, while panels G-I provide examples of new discoveries enabled by the Solve-RD resource. A) Non-canonical splicing variant (individual P0017701); B) mtDNA variant (P0002456); C) *de novo* copy number variant (P0012861); D) Mobile element insertion variant (P0014682); E) Structural variant (P0011371); F) Short tandem repeat expansion (P0002409). G-H) RAB14 *de novo* variants in two cases from this project contribute to the establishment of a new genotype-phenotype relationship. The first individual (P0012753) presents with mild global developmental delay in the absence of any facial dysmorphism or congenital anomalies and carries a *de novo* variant in *RAB14* (chr9:123952916G>A; NM_016322.3:c.200C>T; p.(Thr67Met)) which is rare (not observed in gnomAD v2.1.1), likely to be deleterious (CADD score of 29), and observed *de novo* in at least 4 additional individuals with developmental disorders in the literature (Kaplanis, Samocha, Wiel, Zhang, Retterer, et al., 2020). The second individual (P0012904) presents with mild ID, subtle facial dysmorphisms comprising a high square-shaped forehead, downslant of palpebral fissures and a low hanging columella, in the absence of congenital anomalies. The *de novo* variant found in this individual (chr9:123954475A>C; NM_016322.3:c.80T>G; (p.(Leu27Trp)) is also absent from gnomAD, predicted to be deleterious (CADD score of 28), and observed *de novo* in at least one additional individual with a neurodevelopmental disorder in DECIPHER (Firth et al., 2009) (https://www.deciphergenomics.org/patient/305550/phenotypes/person/62257). The female individual reported in Decipher presents with moderate ID, facial dysmorphism consisting of large earlobes, smooth philtrum, a wide mouth and protruding tongue, short feet with congenital talipes calcaneovalgus, thick hair and an umbilical hernia. I) The new discovery of recurrent *de novo* variants in *RNU4-2* led to likely new diagnoses in two Solve-RD cases. Both variants have been validated, and the phenotypes match the recently published phenotypic descriptions (Y. Chen et al., 2024; Greene et al., 2024).

Box 1: Practical recommendations for large scale distributed genomic re-analysis initiatives.

1) Harmonise pheno-clinical data and metadata, and make sure it is accessible together with the corresponding genomic data.
2) For heterogeneous collections use raw sequencing data as input.
3) Perform quality control of all data as early as possible and define strict inclusion criteria. *e.g.* make sure samples are biologically related in the manner described in the phenotypic submission. Require a minimum on-target coverage - we recommend 80-fold for ES and 30-fold for GS.
4) Apply genome-wide variant calling, irrespective of enrichment kit used for exome sequencing.
5) Use multiple variant calling pipelines for each variant type, with the possible exception of SNV/short InDels, for which variant calling is relatively robust and reproducible. Of all other variant types, CNVs promise the highest yield from exome data, as found here and by Lemire et al (Lemire et al., 2023)
6) Consider reducing stringency with respect to observed alternative allele frequency for heterozygous calls (*i.e.* allow values below 20%), or apply bespoke somatic mutation calling algorithms, if variants are observed in genes commonly associated with the observed phenotype, in order to allow detection of mosaicism or true heterozygotes with poor allele balance.
7) Prioritise variants according to their occurrence in clinical interpretation databases such as ClinVar, HGMD, and similar local/national resources.
8) Reverse phenotyping can be key to re-evaluate the clinical diagnosis in some cases, especially for syndromic disease.
9) Update bioinformatic workflows regularly to incorporate new tools and the latest versions of key databases such as ClinVar.
10) If it is necessary to prioritise among cases for re-analysis, focus first on cases which were (re-analysed) further in the past, since diagnostic yield is likely to be higher.
11) Collect feedback on disease-causing and prioritised candidate variants and solved cases in an accessible database.
12) To facilitate feedback on variant interpretation, favour specificity over sensitivity, and share short-lists of variants for each individual once, and only once.

Online methods

Family Recruitment

Any undiagnosed individual with an apparent genetic RD that falls under the umbrella of conditions in which one of the four partner ERNs specialize, and for whom a prior ES analysis had been undertaken and proven inconclusive, was a candidate to be included in this study. The pan-European recruitment effort involved over 300 clinicians with expertise in RD working in 43 research groups across 37 institutions located in 13 countries. To facilitate data submission and sharing we implemented a pragmatic approach to collect datasets to allow efficient reanalysis across centers. We refer to these datasets as RD-REAL: Rare Disease – RE-Analysis Logistics datasets. An RD-REAL dataset must include genomic data, family information, and phenotypic descriptions. The RD-REAL framework facilitates sharing of data and expertise at a continental scale, consisting of a) expert curated data, b) a comprehensive analysis suite, c) two-level (*i.e.* molecular and clinical) expert review (Figure 1). Informed consent for data sharing within Europe for the purpose of research was obtained from all recruited individuals.

Data pertaining to 10,039 individuals from 6,246 undiagnosed families was initially assembled, which was reduced to 9,645 individuals (6,447 affected) in 6,004 families following application of quality control measures, as described below. Of the 6,447 affected individuals, 3,592 (56%) were male and 2,855 (46%) female. 6,215 (96.4%) were alive at the start of the study, 84 (1.3%) were deceased, and for 148 (2.3%) the vital status was unknown.

Pseudonymized phenotypic data collation for all individuals was facilitated using the PhenoStore module of the RD-Connect GPAP. PhenoStore promotes deep phenotyping of affected individuals using HPO terms and disease classification using Orphanet Rare Disease Ontology (ORDO) ORPHA codes (http://www.orphadata.org/cgi-bin/index.php) and/or OMIM identifiers (https://www.omim.org/) as appropriate, and can import/export this information using the GA4GH Phenopackets format (Jacobsen et al., 2022).

ERN Cohort descriptions

For all families recruited to Solve-RD, local standard of care genetic diagnostic work-up and/or research-based analyses had failed to identify any molecular genetic cause underlying the proband's rare condition.

ERN RND

The ERN RND cohort consists of 2,799 individuals from 2,271 families with previously unsolved rare neurological diseases. Genomic and phenotypic data for all affected individuals, and family members where available (~20% of families), were submitted for reanalysis by nine ERN RND partner institutions located in eight European countries: Belgium, France, Germany, Hungary, the Netherlands, Slovenia, Spain, and the United Kingdom. Individuals had been recruited and sequenced either as part of standard diagnostic care, or through participation in large European rare neurological disease research projects such as NeurOmics (https://rd-neuromics.eu/) and Treat-HSP (https://www.treathsp.net/). The 2,271 families comprised 1,924 singletons, 168 duos, 141 triples (103 of which were parent-child trios), and 38 families with four or more members, including a total of 2,453 affected individuals. The most frequently used HPO terms to describe the phenotypes were ataxia, gait disturbance, dysarthria, and spastic paraplegia (Supplementary Table S10). ORPHA codes were assigned to 1,294 (57%) probands, the most common of which were Hereditary Spastic Paraplegia (n=397, 31%), Rare Hereditary Ataxia (n=353, 27%), and Paroxysmal Disorders (n=189, 15%).

ERN ITHACA

The ERN ITHACA cohort consists of 4,405 individuals from 1,836 families, submitted for reanalysis by twelve partner institutions located in six countries: the Czech Republic, France, Germany, Italy, the Netherlands, and the United Kingdom. A further 65 individuals from 21 families from the Spanish Undiagnosed Disease Program (SpainUDP) (López-Martín et al., 2018) were included in this cohort for analysis, due to the similarity of the underlying phenotypes. The clinical spectrum of the ERN ITHACA cohort consisted of individuals with intellectual disability (ID) with or without additional phenotypic features, and individuals with (multiple) congenital anomalies without ID. Given the importance of *de novo* mutations underlying the rare conditions within ERN ITHACA (de Ligt et al., 2012; Gilissen et al., 2014), unaffected parents and/or unaffected siblings were also included, wherever possible, to allow for direct segregation of variants. The 1,857 families comprised 632 singletons, 38 duos, 1,138 triples (1,081 parent-child trios), and 49 families with four or more members, including a total of 1,933 affected individuals. The most frequently used HPO terms to describe affected individuals related to global developmental decay, intellectual disability, and autism (Supplementary Table S10). ORPHA codes were assigned to 242 (13%) probands, the most common of which were Rare Intellectual Disability (n=41, 17%), and Pseudohypoparathyroidism with Albright Hereditary Osteodystrophy (n=22, 9%).

EURO-NMD

The ERN EURO-NMD cohort consists of 2,125 individuals from 1,517 families, submitted for reanalysis by sixteen partner institutions located in eight countries: Belgium, Canada, Finland, France, Germany, Italy, Spain and the United Kingdom. Previously unsolved RD-REAL datasets submitted to Solve-RD had either been recruited and sequenced as part of large international neuromuscular research projects such as NeurOmics (https://rd-neuromics.eu/), SeqNMD, Myocapture (Bauché et al., 2016), MYO-SEQ (Töpf et al., 2020), UK10K (https://www.uk10k.org/), Unravel-CMS, BBMRI-LPC (https://cordis.europa.eu/project/id/313010), CMS CMG (https://cmg.broadinstitute.org/), Consequitur (Hiz Kurul et al., 2022), or through participating centres' own diagnostic or research pipelines. Samples incorporated from the MYO-SEQ project were recruited from 50 specialised neuromuscular disease centres across Europe and the Middle East, and some datasets incorporated from the Unravel-CMS, BBMRI-LPC, and CMS CMG projects were from privately sequenced undiagnosed individuals followed at Nimhans, India (https://nimhans.ac.in/). The 1,517 families comprised 1,202 singletons, 90 duos, 156 triples (135 parent-child trios), and 69 families with four or more members, including a total of 1,685 affected individuals. The most frequently used HPO terms to describe affected individuals related to muscle weakness, myopathy, and abnormal muscle morphology (Supplementary Table S10). ORPHA codes were assigned to 338 (22%) probands, the most common of which were Limb-Girdle Muscular Dystrophy (n=56, 17%), Congenital Myasthenic Syndrome (n=49, 14%), Distal Hereditary Motor Neuropathy (n=44, 13%),and non-Limb-Girdle Muscular Dystrophy (n=44, 13%).

ERN GENTURIS

The ERN GENTURIS cohort consists of 390 individuals, from 359 families, with a suspected genetic tumor risk syndrome, submitted for reanalysis by seven partner institutions located in four countries: Germany, the Netherlands, Portugal, and Spain. All individuals were either recruited and sequenced as part of daily diagnostic care, or as part of research projects. The 359 families comprised 345 singletons, six duos, four triples (one parent-child trio), and four families with four or more members, including a total of 378 affected individuals. The most frequently used terms to describe affected individuals related to colorectal cancer, followed by gastric cancer and pheochromocytoma (Supplementary Table S10). ORPHA codes were assigned to 252 (70%) of the probands, the most common of which were Familial Adenomatous Polyposis (n=122, 48%), Hereditary Gastric Cancer (n=65, 26%), Hyperplastic Polyposis Syndrome (n=56, 22%), and Intestinal Polyposis Syndrome (n=33, 13%).

Depth of standardised phenotypic and clinical diagnosis annotations

A median of six HPO terms (range 0 - 74) were used to describe each affected individual across this Solve-RD cohort. This drops to five HPO terms (range 0 - 45) when removing HPO redundancies. To remove annotation redundancy, only the most specific HPO terms were considered by counting terms from leaf nodes, or nodes without selected parent or child entities. Overall quality of phenotypic descriptions was assessed using the Monarch Initiative annotation sufficiency score (Shefchek et al., 2020) (maximum possible value of 5.0). The median annotation sufficiency value across the Solve-RD cohort was 3.61 (Supplementary Figure S1). Clinical diagnosis was reported using ORDOcodes for 2,126 affected individuals.

Generation of ERN-specific Candidate Gene Lists

To facilitate the potential for clinicians to confirm a diagnosis based upon identified variants, findings returned to the ERN DITFs for interpretation were restricted to those in disease genes of interest to the specific ERN, apart from any potentially pathogenic variants encountered in the mitochondrial genome, all of which were returned. Each of the four ERNs generated a curated list of genes implicated in diseases studied by their ERN, exploiting their pan-European disease expertise. The RND list was primarily based upon genes associated with neurological disease which have a green review status in Genomics England PanelApp (A. R. Martin et al., 2019), with the addition of 25 further genes based upon recommendations by clinical experts (n=1,821 genes). For ITHACA a consolidation of gene lists pertaining to intellectual disability (ID) from a variety of resources was undertaken, followed by evaluation based upon occurrence in multiple resources and the quality of curation of said resources, resulting in a list of diagnostically relevant genes (n=1,645). In the case of GENTURIS the list included all genes routinely screened in the partners' diagnostic laboratories (n=230). For EURO-NMD, the manually curated, and annually updated, Gene Table of Muscular Disorders (Benarroch et al., 2023) was used (n=615 in 2021). These ERN gene lists were used as a primary filter in the identification of potentially pathogenic variants of any type in affected individuals submitted to Solve-RD by collaborators from the corresponding ERN, irrespective of the individual's phenotype. This resulted in a list of 2,512 distinct genes implicated in rare diseases of interest to the four ERNs, many of which were identified by more than one ERN (Supplementary Table S2). Further details on the criteria used to define these lists are provided in the Supplementary Information.

Identification of Clinically Actionable Genes

Potentially clinically actionable genes in affected individuals were identified from three independent initiatives: ClinGen (Rehm et al., 2015) (n=77), IEMbase (J. J. Y. Lee et al., 2018) (n=214), and Treatabolome (Bonne, 2021) (n=154; https://treatabolome.cnag.crg.eu). This provided a total of 392 unique genes, of which 311 (79%) were included in at least one of the curated gene lists from the ERNs. For the assessment of clinically actionable genes in individuals affected by a hereditary cancer disposition we searched GeneReviews® and the National Comprehensive Cancer Network Clinical Practice Guidelines in Oncology (Ajani et al., 2022) for actionability based on surveillance for cancer advice.

Data Submission and Analysis Workflow

Raw sequencing data, associated metadata, phenotypic and pedigree descriptions, were collated from 43 research groups across Europe using the RD-Connect GPAP (Laurie et al., 2022). In order to ensure secure, rapid, and robust transfer of the large quantity of raw genomic data (FASTQ, BAM or CRAM) to be reanalysed (approximately 100 terabytes in total), each research group was provided with access to a dedicated private space, in which to upload their sequencing data, on an Aspera server hosted by RedIRIS, the Spanish national research and education network (https://www.rediris.es/). From here the sequencing data were downloaded to the Centro Nacional de Análisis Genómico in Barcelona, which develops and hosts the RD-Connect GPAP.

All genomic data submitted to Solve-RD was analysed in identical fashion to minimise any batch effects, using the RD-Connect GPAP standard analysis pipeline (Laurie et al., 2016). Briefly, reads were aligned to the decoy version of GRCh37 (hs37d5) using BWA-MEM v0.7.8 (Li, 2013). Short variants, *i.e.* SNVs, and insertions and deletions less than 50nt in length (referred to here as InDels), were identified across the genome, independent of the target capture region of interest, using the GATK HaplotypeCaller (v3.6) in accord with the GATK Best Practices workflow (DePristo et al., 2011). The output of the pipeline for each experiment is an aligned, base quality score recalibrated BAM, and a gVCF per chromosome and for the mitochondrion. All variant positions covered by at least eight reads and a GATK assigned Genotype Quality of at least 20, are uploaded to the RD-Connect GPAP, as are any non-variant positions for which at least one other experiment in the uploaded batch has a variant position at the same genomic location. SNVs, InDels, and mitochondrial variants received detailed annotations provided by Ensembl Variant Effect Predictor (McLaren et al., 2016), gnomAD (Karczewski, Francioli, Tiao,

Cummings, MacArthur, et al., 2020), and ClinVar (Landrum et al., 2020) (May 2021 release), among other resources.

In addition to the above described annotations available through the RD-Connect GPAP, all gVCFs derived from affected individuals were converted to VCFs and annotated by a custom annotation pipeline at RadboudUMC, as described previously. (Lelieveld, Reijnders, Pfundt, Yntema, Kamsteeg, De Vries, et al., 2016) This comprises variant-based annotations including nucleotide conservation scores (phyloP, CADD), RadboudUMC in-house database allele frequencies, and gene-based annotations including e.g. mouse knockout model phenotypes, and pLI/LOEUF scores, among others. These annotated VCF files were made available to the Solve-RD consortium through the Solve-RD Sandbox, hosted by UMC Groningen, Netherlands (see below).

Raw sequencing data (FASTQ), and newly generated alignment (BAM or CRAM) and variant call (gVCF) files for each experiment, accompanied by the corresponding phenotypic description in Phenopackets and pedigree descriptions in PLINK PED format, were submitted to the European Genome-Phenome Archive (EGA) (Freeberg et al., 2022) in Hinxton, UK for long-term archival and controlled access to the wider human genomics community.

To facilitate further analyses within the Solve-RD consortium, files submitted to the EGA were shared with authorised Solve-RD partners from the University of Nijmegen, Netherlands and the University of Tübingen, Germany. In addition, the files could be downloaded by authorised users to the Solve-RD sandbox. The Solve-RD sandbox is a cloud environment used by project partners to conduct bespoke analyses and to securely share analysis and interpretation results. For more detailed information on the Solve-RD information technology infrastructure, please see Johansson et al. 2024 (manuscript accepted for publication, GigaScience).

Quality Control

A total of 10,276 ES and GS RD-REAL datasets from 10,039 individuals were initially submitted to Solve-RD for reanalysis. Preliminary quality control of sequencing data required a median coverage of at least ten reads over at least 70% of the defined target region of interest for the corresponding enrichment kit, or across the entire genome in the case of GS data. Furthermore, with respect to phenotypic data, each submitted family was required to have an affected proband with associated HPO terms. Misassigned relationships were identified, and subsequently corrected where possible, using KING (Manichaikul et al., 2010). Following application of

these quality control measures the final number of RD-REAL datasets taken forward for reanalysis comprised data from 9,645 individuals from 6,004 families, of which 6,447 (66.9%) are affected by a rare disease. Of these, ES data was available for 9,124 (94.6%) individuals, GS data for 333 (3.5%), and both ES and GS data for the remaining 190 (2.0%).

Variant Identification and Prioritisation

RD-REAL Data Analysis and Interpretation

We applied two-level expert analysis and interpretation to the RD-REAL datasets comprising firstly of bioinformatic and molecular genetics experts working together in dedicated working groups within a Data Analysis Task Force (DATF), and secondly, clinical and RD experts from each ERN who jointly prioritised and interpreted all variants returned by the DATF, working in four separate Data Interpretation Task Forces (DITF). To maximise generalisability of this effort, the entire dataset of 6,004 families, was included in a comprehensive analysis suite comprised of firstly, a centralised analysis of each different variant type: SNVs and InDels; *de novo* mutations; mitochondrial variants, non-canonical splice variants, copy number variants (CNVs), structural variants (SVs), short tandem repeat expansions (STRs) and mobile element insertions (MEIs). Secondly, general filters were applied with respect to variant quality, control population allele frequencies, and predicted consequence, followed by further ERN and disease specific filters (see below). Details of all tools applied in these analyses are provided in Supplementary Table S11.

As Solve-RD processed data in multiple data freezes over time, subsets of experiments continued to undergo analyses in parallel, some of which resulted in new diagnoses prior to the results of the centralised systematic analyses being returned to submitters. This includes the preliminary analysis of a smaller dataset (Denommé-Pichon et al., 2023; Matalonga et al., 2021). Furthermore, many datasets underwent parallel or additional analyses in the laboratories of the respective submitters, resulting in the identification of (likely) pathogenic, or candidate disease-causing variants in established or novel genes. These results are labelled as ad hoc expert review (Figure 2, Supplementary Table S12), though the majority of these variants were also prioritised in the systematic analyses.

Taken together this resulted in either diagnosed individuals, i.e. those harbouring (likely) pathogenic variants which fully explain the proband's phenotype, unequivocally allowing a molecular diagnosis of a rare condition, or affected individuals with candidate variants, worthy of further follow-up and/or functional

studies, which may prove to be diagnostic in the future, as adjudged by the referring clinicians and/or expert ERN partners.

SNVs/InDels

Programmatic reanalysis was undertaken on annotated variants from the RD-Connect GPAP using Application Programming Interface (API) endpoints as previously described (Matalonga et al., 2021). Two different sets of parameters were used: firstly, a low-hanging fruit analysis to identify (likely) pathogenic variants already listed in ClinVar; secondly, identification of rare variants of high or moderate impact in ERN genes of interest, matching the expected mode(s) of inheritance.

1. **Low-hanging fruit analysis**: depth of coverage (DP) >7; GATK genotype quality (GQ) >19; Minor Allele Frequency (MAF) <0.01 in gnomAD; observed allele frequency <0.02 in the internal RD-Connect GPAP database; affecting a gene in the corresponding ERN gene list, and annotated as pathogenic (class 5) or likely pathogenic (class 4) for any disorder in ClinVar as of May 2021.
2. **High-Moderate impact variant analysis**: DP >7; GQ >19; MAF <0.01 in gnomAD; observed allele frequency <0.02 in the internal RD-Connect GPAP database; affecting a gene in the corresponding ERN gene list, and predicted to have a high or moderate consequence at the protein level according to Ensembl VEP, and matching the expected inheritance pattern i.e. autosomal dominant, autosomal recessive or X-linked.

Variants passing the above filtering criteria were returned in a single table to the respective DITF for each ERN to facilitate evaluation and provision of feedback. Across the Solve-RD cohort we identified a mean of nine short variants per affected individual for interpretation, ranging from one to thirteen across ERNs, the difference largely reflecting differences in the number of genes included in the corresponding ERN gene lists (Supplementary Tables S2 and S13).

De novo ***Mutations***

For all families for which parent-child trios were available (n=1,320; 22% overall), *de novo* mutation calling was undertaken using both HaplotypeCaller (DePristo et al., 2011) and DeNovoCNN (Khazeeva et al., 2022). DNM calls with a probability >0.85 of being a bona fide event and any apparent *de novo* mutations identified by HaplotypeCaller which were located in a gene on the respective ERN gene list were returned to DITFs for variant interpretation.

Mitochondrial Genome Variants

Mitochondrial DNA variants were identified using MToolBox (Calabrese, Simone, Diroma, Santorsola, Gutta, et al., 2014) (version 1.2). The workflow includes mapping reads to the revised Cambridge Reference Sequence mitochondrial genome and annotation using the MITOMAP (Lott et al., 2013) database (accessed 28th June 2021). Both homoplasmic and heteroplasmic variants were identified.

Identification of Non-canonical Splicing Variants

To identify variants potentially affecting splicing at sites other than canonical splice sites, two novel tools were applied, SpliceAI (Jaganathan et al., 2019) and SQUIRLS (Danis et al., 2021). Rare variants receiving a strong splice-altering prediction from both tools, *i.e.* a delta-score >0.8 in SpliceAI and a pathogenic classification by SQUIRLS, which would potentially alter splicing of any gene in the corresponding ERN gene list were returned to DITFs for interpretation.

Large Copy Number and Structural Variants

Three different tools were used to maximise the likelihood of identifying pathogenic CNVs as described in Demidov et al. (Demidov et al., 2023): ClinCNV (Demidov et al., 2022), Conifer (Krumm et al., 2012), and ExomeDepth (Plagnol et al., 2012). Variants observed to have a frequency >0.01 across the cohort were discarded and the remaining rare CNVs were intersected with the corresponding ERN gene list, and annotated using AnnotSV (Geoffroy et al., 2018), before being returned to DITFs for interpretation. In parallel Manta (X. Chen et al., 2016; Demidov et al., 2024) was run in exome mode to look for signatures of split reads, which may indicate the presence of balanced structural variants such as inversions. To facilitate interpretation, Integrative Genomics Viewer (Robinson et al., 2011) tracks were generated for all large variants, indicating the exons, the position and type of call produced by the tools, and beta-allele frequency.

Short Tandem Repeat Expansions

The identification of potentially pathogenic STR expansions was largely based on the work of van der Sanden et al. (van der Sanden et al., 2021). ExpansionHunter (Dolzhenko et al., 2017) (version 3.1.2) was used to screen 21 genomic loci previously described as harbouring pathogenic repeat expansions in both ES and GS data, from a total of 5,983 families. Following retrieval of predicted pathogenic genotypes across all samples, any frequently observed events were discarded, and the remaining variants affecting genes on the corresponding ERN gene list were manually curated by visual inspection, before being returned to DITFs for interpretation.

Mobile Element Insertions

MEI identification was undertaken using both MELT (Gardner, Lam, Harris, Chuang, Scott, Stephen Pittard, et al., 2017) and SCRAMble (Torene et al., 2020), to identify any MEIs potentially affecting ERN genes of interest (see Wijngaard et al (Wijngaard, Demidov, O'Gorman, Corominas-Galbany, Yaldiz, Steyaert, de Boer, Vissers, Kamsteeg, Pfundt, et al., 2024), for further details).

Inclusion and ethics statement

All individuals were recruited via four ERNs. Inclusion criteria were a clinical rare disease diagnosis in at least one family member by one of the associated expert centres and a not conclusive exome or genome analysis at time of submission. We did not exclude anyone based on sex, gender, ethnicity, race, age or any other socially relevant groupings.

Each patient entry was associated with its submitting investigator or clinician and linked to its corresponding ERN or UDP. The responsibility of checking the data was suitable for submission to the RD-Connect GPAP and Solve-RD lay with the data submitter as required by their Code of Conduct (institution: Fundació Centre de Regulació Genòmica) and Data Sharing Policy (institution: Solve-RD general assembly), respectively. In some cases, individuals had to be re-consented prior to data submission. The individuals described in Supplementary Figure 6 gave permission for their photos to be used in this publication, for which we thank them and their families. This study adheres to the principles set out in the Declaration of Helsinki.

Data Availability

Pseudonymised phenotypic information for all individuals and their genetic variants are accessible through the RD-Connect GPAP (https://platform.rd-connect.eu/) upon validated registration. All raw and processed data files are available at the EGA (Datasets EGAD00001009767, EGAD00001009768, EGAD00001009769, and EGAD00001009770, under Solve-RD study EGAS00001003851). All novel and expert curated variants have been submitted to ClinVar (n=207).

Acknowledgements

The Solve-RD consortium is grateful to all involved RD patients and their families as well as other contributors to Solve-RD.

The Solve-RD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779257. This

research is supported (not financially) by four ERNs: (1) The ERN for Intellectual Disability, Telehealth and Congenital Anomalies (ERN ITHACA)—Project ID No 869189; (2) The ERN on Rare Neurological Diseases (ERN RND)—Project ID No 739510; (3) The ERN for Neuromuscular Diseases (ERN Euro-NMD)—Project ID No 870177; (4) The ERN on Genetic Tumour Risk Syndromes (ERN GENTURIS)—Project ID No 739547. The ERNs are co-funded by the European Union within the framework of the Third Health Programme.

Additional funding and support is acknowledged in the Online Material.

Disclosures

Matthis Synofzik has received consultancy honoraria from Janssen, Ionis, Orphazyme, Servier, Reata, GenOrph, and AviadoBio, all unrelated to the present manuscript.

Supplementary materials

Abbreviations Used

Array CGH	Microarray-based Comparative Genomic Hybridisation
CNV	Copy Number Variant
DATF	Solve-RD Data Analysis Task Force
DITF	Solve-RD Data Interpretation Task Force
DNM	*De novo* Mutations
DP	Depth of Coverage, assigned by GATK
EGA	The European Genome-Phenome Archive
ERN	European Reference Network
ERN GENTURIS	The ERN for Genetic Tumour Risk Syndromes
ERN ITHACA	The ERN for Intellectual disability, TeleHealth, And Congenital Anomalies
EURO NMD	The ERN for Rare Neuromuscular Diseases
ERN RND	The ERN for Rare Neurological Diseases
ES	Exome Sequencing
GATK	The Genome Analysis Tool Kit
gVCF	Genomic Variant Call Format file
GQ	Genotype Quality, assigned by GATK
GS	Genome Sequencing
HPO	The Human Phenotype Ontology
ID	Intellectual Disability
InDels	Short Insertions or Deletions (<50 nucleotides in length)
IGV	The Integrative Genomics Viewer
MAF	Minor Allele Frequency
MEI	Mobile Element Insertion
MLPA	Multiplex Ligation-dependent Probe Amplification
NDD	Neurodevelopmental Disorders
ORDO	The Orphanet Rare Disease Ontology
RD	Rare Disease
RD-Connect GPAP	The RD-Connect Genome-Phenome Analysis Platform
RD-REAL	Rare Disease REAnalysis Logistics
SNV	Single Nucleotide Variant
SpainUDP	The Spanish Undiagnosed Rare Disease Program
STR	Short Tandem Repeat
SV	Structural Variant

Definitions

Pathogenic variants, as described in this manuscript, are defined as genomic variants with a direct, proven link between the exact change, the gene it occurs in, and a particular condition or syndrome.

Likely pathogenic variants, as described in this manuscript are defined as genomic variants for which there is a high likelihood for a causal relation between the change, the gene it occurs in, and a particular condition or syndrome.

Variants of unknown significance (VUS), as described in this manuscript, are defined as either i) variants affecting a disease gene on the corresponding ERN disease gene list, but for which establishing variant pathogenicity requires further analysis (such as by experimental function follow-up, segregation analysis etc.), or alternatively, ii) variants meeting criteria for pathogenicity based on molecular deleteriousness to protein function, but for which the gene-disease association has not been established (e.g. a *de novo* Loss of Function variant in candidate disease gene).

Disease genes, as described in this manuscript, are defined as genes for which, based on expert opinion by the ERNs, (well)-established genotype-phenotypes exist. Consequently, the gene was included for interpretation by the ERNs in this study.

Candidate disease genes, as described in this manuscript, are defined as genes for which a genotype-phenotype association has not been formally established. Based on gene function and ERN expert opinion, it may, however, be expected that (likely) pathogenic variants disrupt this function, resulting in phenotypic consequences in line with the ERNs disease domain.

Disease-causing variants, as described in this manuscript, are defined as (likely) pathogenic variants in disease genes that fully explain the RD phenotype observed in the individual. For dominant diseases, this involves a single (likely) pathogenic variant affecting one allele. For recessive diseases, (likely) pathogenic variants are observed for both alleles.

Candidate disease-causing variants, as described in this manuscript, are defined as either i) variants of unknown significance in established disease genes, ii) (likely) pathogenic variants in disease genes which do not fully explain the RD phenotype observed, or iii) (likely) pathogenic variants in a candidate disease gene.

Supplementary Results

Example Case Reports

The following case reports illustrate a spectrum of individuals who have been diagnosed as a result of their inclusion in Solve-RD. They describe the clinical courses of disease, previous test results, indicate the newly identified (likely) pathogenic variants identified by Solve-RD, and indicate why the original analyses failed to provide a diagnosis.

ERN RND

P0015028

This 58-year-old male attended special education in childhood but was reported not to have physical problems at that time. However, at the age of 42, he developed progressive gait disturbances, eventually requiring the use of a walking aid, approximately eight years after the initial onset of symptoms, and also complained of urinary urge incontinence. He underwent nerve conduction studies including electromyography at 44 years of age, which showed signs of sensory neuronopathy or sensory polyneuropathy. Neurological examination at 48 years of age revealed a spastic ataxic gait and confirmed the presence of signs of peripheral neuropathy. His lower limbs were hypertonic with peroneal muscle weakness. Vibration sense in his ankles was impaired, his Achilles tendon reflexes were absent, and there was a bilateral extensor plantar response. Neurological examination of the upper limbs was normal. Magnetic Resonance Imaging (MRI) of his brain and spine at 48 years of age showed no abnormalities, but his eye movements showed saccadic hypermetria with increased latency of pursuit.

Systematic reanalysis of ES data in Solve-RD identified two variants in the *B4GALNT1* gene, one being an intragenic heterozygous deletion of a large part of the gene (commencing in exon 5, removing exons 6-11 (NM_001478.5), and ending in the 3'UTR (Chr12(GRch37): g.58014705-58024263del; Supplementary Figure S5) and the other a missense SNV leading to an amino acid change of unknown pathogenicity (c.451G>A (p.(Gly151Ser)). Presence of the deletion was confirmed with multiplex amplicon quantification (MAQ) analysis, and segregation analysis showed that the proband's mother was a heterozygous carrier of the deletion (*B4GALNT1* Chr4(GRch37): g.(58016984-?)_(58019959+?)del), but not of the missense variant. Unfortunately, the proband's father was not available for genetic testing. Functional confirmation was obtained via glycomics analysis of plasma glycolipids, indicating reduced levels of *B4GALNT1* glycolipid products. Pathogenic variants

in the *B4GALNT1* gene are known to cause autosomal recessive spastic paraplegia type 26 (OMIM #609195; https://www.omim.org/entry/609195), which was the diagnosis reached for this individual, given the combination of phenotypic match, genetic data, and biochemical profile. Thus, the final diagnosis was achieved via a combination of the variant identification undertaken in Solve-RD and subsequent follow-up experiments performed by the local research team.

ERN ITHACA

P0012716

This 24 year old male was referred for genetic diagnostics at 15 years and 10 months of age with recurrent luxation of the left patella and asymmetry of legs and face, described as underdevelopment of the left side. During pregnancy, his mother had a large splenic cyst, but pregnancy and delivery were otherwise uncomplicated. At birth, asymmetry of the legs and face was evident and there was a postaxial rudimentary digit on the right hand that regressed to a small nodule over time. The asymmetry of the face and legs was reported to be stable over time and his cognitive development was within the normal range (IQ of 89). He was affected by complex partial seizures with continuous spike-and-wave during sleep from childhood, however the seizures had a good clinical progression and medication could be discontinued at eleven years of age. Other medical problems included scoliosis, autism spectrum disorder, clumsy motor skills, and sleeping problems. Physical examination at 15 years and 10 months of age revealed normal overall growth (height 172.6 cm (SD -1), weight 58.3 kg (SD=0), and head circumference 55 cm (SD -1)), with asymmetry of the legs (thigh circumference right 46 cm, left 41.5 cm; calf circumference right 25 cm, left 26 cm; ankle circumference right 38 cm, left 36 cm; smaller shoe size of left foot). Additionally, there was gynaecomastia on the right side, a small postaxial nodule on the right hand, scoliosis, and a pelvic tilt as a result of the difference in leg length. Facial characteristics included full hair with a low posterior hairline, synophrys, long eyelashes, and facial asymmetry, with a smaller left than right side and the left eye positioned lower than the right eye. Re-examination at 23 years and six months showed leg length difference of 2 cm (right longer than left), and a thigh circumference difference (right 54 cm, left 49.5 cm), similar facial asymmetry to that seen at age 15 years and 10 months, scoliosis with winging of the right scapula, hypermobility of the hand joints and one café au lait spot on the right leg.

Previous genetic investigations included Affymetrix CytoScan HD array, which did not identify any relevant variants, and trio-based ES (Agilent SureSelectXT

Human All Exon 50Mb Kit, Illumina HiSeq2000, BGI-Europe, with a mean target coverage of 115-fold). ES data were subjected to several local reanalyses, but only a heterozygous *de novo* variant of uncertain significance in DLG1 was identified (Chr3(GRCh37):g.196807944T>C; NM_004087.2:c.1982A>G; p.(Asn661Ser)) as being of potential interest.

Systematic reanalysis of ES data in Solve-RD using the *de novo* variant calling pipeline led to the identification of a mosaic *de novo* variant in *PIK3CA*: Chr3(GRCh37):g.178916876G>A; NM_006218.4:c.263G>A; p.(Arg88Gln), present in 13% of the reads (Supplementary Figure S6 Panel A). The variant had been previously reported in PIK3CA-related overgrowth syndrome (McDermott et al., 2018; Rivière et al., 2012).

Retrospective analysis of the original ES data revealed that the variant was probably missed due to its mosaic state, and hence having been assigned a low quality by the variant calling software. Additionally, this individual had been clinically suspected to have underdevelopment of the left side of the body, rather than overgrowth of the right side of the body, which meant that an overgrowth syndrome had not previously been considered. Furthermore, likely due to the mosaicism, the proband presents with a relatively mild phenotype when considering the spectrum of *PIK3CA*-related overgrowth, which makes accurate clinical diagnosis challenging. Inclusion within Solve-RD ended a decade-long diagnostic odyssey for this individual and his family.

P0013065

This 22 year old male was born following an uncomplicated pregnancy and delivery at 39 weeks of gestation. Initial concerns regarding his development arose around six months of age and he was subsequently affected by severe developmental delay, with delayed motor, communicative and social milestones: crawling at 15 months, walking at two years and six months, first words at seven years of age, and speech characterised by severe verbal dyspraxia. Additional medical problems comprised divergent strabismus, muscle tone dysregulation with contractures, and inattentive and hyperactive behaviour with aggressive tantrums. The family history was unremarkable. Physical examination at 21 years and 11 months revealed a slender body and microcephaly (height 184 cm (SD=0); weight 51.5 kg, BMI 15.2; head circumference 54.5 cm, SD -2). He had a small asymmetric thorax of unusual shape (the mid-thoracic region being broader in the frontal plane and flattened in the sagittal plane compared to the high thoracic region), high thoracic kyphosis, and scapular winging. His hands and feet were slender with long fingers

and toes, camptodactyly of the 2nd, 3rd and 4th fingers of the right hand, and he exhibited elbow and knee contractures. Facial dysmorphisms included a long and narrow facial shape, full eyebrows with synophrys, downslant of the palpebral fissures, prominent eyelids with ptosis, divergent strabismus, low-set ears with a square-shaped and flattened upper helix, a short nose, an open mouth with full lip vermillion, a high and narrow palate with gum hypertrophy and irregular dentition. General investigations included brain MRI (at seven and seventeen years of age), EEG, brainstem auditory evoked potential analysis, general metabolic screening, and chest, spine and skull X-rays at seven years of age, none of which showed any clinically meaningful abnormalities.

This individual has undergone a wide variety of previous genetic investigations including karyotyping, analysis of repeats in *FMR1*, 3.5K and 32K BAC arrays, Affymetrix 6.0 array, Array CGH, and trio-based ES (Illumina HiSeq at BGI-Europa, after enrichment with Agilent SureSelectXT Human All Exon 50Mb Kit, with a mean target coverage of 110-fold (de Ligt et al., 2012)). Upon negative ES results, and repeated negative reanalyses, trio-based GS was performed (DNBSEQTM technology BGI, mean coverage >40-fold), but no conclusive diagnosis was established.

Systematic reanalysis of ES and GS data in Solve-RD with both the *de novo* variant calling and SNV/InDel low-hanging fruit analysis workflows led to the identification of a heterozygous *de novo* nonsense variant in *MN1*: Chr22(GRCh37):g.28146963C>T; NM_002430.2:c.3903G>A; p.(Trp1301*). (Supplementary Figure S6 Panel B). This nonsense variant in the last exon of *MN1* has been reported to be pathogenic in the literature (CEBALID syndrome (Mak et al., 2020); MIM#618774) and associated with a very similar phenotype to that observed in this individual. However, brain abnormalities reported in individuals with variants in *MN1* from other studies had not previously been detected in the brain imaging of this individual. Therefore, we performed retrospective reanalysis of his brain MRI, which revealed dysplasia of the cerebellar vermis, rhombencephalosynapsis and mild bitemporal narrowing of the skull, highly similar to the MN1-associated brain phenotype described by Mak et al (Mak et al., 2020).

Retrospective analysis of the original ES and GS data revealed that this variant in MN1 was indeed noticed by the diagnostic laboratory, but discarded because *MN1* had not yet been described to cause CEBALID syndrome at the time of the original analysis, having only been associated with familial susceptibility to meningioma (MIM#607174, https://www.omim.org/entry/607174) at that point in time. Thus, it was reanalysis within Solve-RD that finally led to the end of this individual's twenty-year diagnostic odyssey.

EURO-NMD

P0005327

This 14 year-old boy was born to non-consanguineous parents and had normal developmental milestones overall but was not active in sports from early childhood. He started to experience recurrent falls at eight years of age and went on to develop progressive proximal lower limb weakness with prominent fatigability, and a waddling gait. There was no history of bulbar or ocular symptoms. On examination, bilateral asymmetric ptosis with fatigability was observed, as was polyminimyoclonus. Muscle strength was normal in all four limbs, but fatigue occurred upon sustained arm abduction. Deep tendon reflexes were normal, as were serum creatine kinase levels, while repetitive nerve stimulation was inconclusive. Due to a clinical suspicion of Congenital Myasthenic Syndrome (CMS), a trial of pyridostigmine was initiated, but the individual was non-compliant. However, his parents reported spontaneous improvement in baseline limb weakness and falls over the following six years with only episodic worsening due to fever and exertional myalgias. There was a strong family history of diabetes on the maternal side and the mother's fasting glucose levels were suggestive of borderline diabetes, and she also has a long history of migraines. Retrospective serum lactate testing in both proband and mother showed mildly elevated levels (>20 mg/dl).

Prior genetic testing for spinal muscular atrophy was negative, as was ES with respect to detection of any significant variants in known CMS and myopathy genes.

Systematic reanalysis of the affected family members' ES data within Solve-RD led to the identification of a mitochondrial variant, m.3243A>G, in *MT-TL1* with an observed heteroplasmy of 0.27 in the proband and 0.14 in his mother (Supplementary Figure S7). The difference in heteroplasmy likely correlates with the mild phenotype observed in the proband, and the absence of mitochondrial myopathy features in his mother. However, non-specific findings of early onset diabetes and migraine headaches were noted in the mother. While the initial clinical suspicion in the proband was CMS due to the significant fatigability, the fact that mitochondrial disease can be clinically highly variable means that mild forms of mitochondrial myopathy can be difficult to diagnose clinically.

Hence it was the dedicated reanalysis of mitochondrial genes in Solve-RD, in conjunction with the retrospective correlation of phenotype and biochemistry which helped to determine the correct diagnosis for this individual.

P0014714

This young girl was referred at seven years of age, presenting with microcephaly, face abnormality, muscle hypotonia, and neurodevelopmental delay, leading to a clinical suspicion of Cornelia de Lange syndrome (MIM#122470; https://www.omim.org/entry/122470). Array CGH and ES were performed, but only genes related to the clinical suspicion (i.e. a virtual panel) were evaluated, and no relevant variants were identified at this point.

Systematic reanalysis of the proband's ES data within Solve-RD led to the identification of a *de novo* frameshift variant in the histone acetyltransferase p300 gene: EP300(NM_001429.4):c.1152_1153del; p.(Gly385GlnfsTer25). The variant was discovered following a cross-ERN *de novo* mutation scan (see Supplementary Table S5), suggesting a clinical diagnosis of Rubinstein Taybi syndrome (MIM#180849; https://www.omim.org/entry/180849). This prompted clinical re-evaluation of the proband's phenotype, at which point the clinical diagnosis was confirmed. Retrospective analysis of the original ES results revealed that the variant had not been called by the corresponding in-house pipeline.

P0012248

This young male was initially referred for genetic testing at three years of age. He presented with severe neurodevelopmental delay, microcephaly, absent speech, generalised hypotonia, nystagmus, and inability to walk. However, array CGH and ES results proved negative.

Systematic reanalysis of the proband's ES data within Solve-RD led to the identification of a *de novo* missense variant in synaptotagmin 1, *SYT1*(NM_001135806.2):c.1103T>C; p.(Ile368Thr). This variant was discovered following a cross-ERN *de novo* mutation scan, leading to a molecular diagnosis of Baker-Gordon syndrome (MIM#618218; https://www.omim.org/entry/618218).

Retrospective analysis of the original ES data revealed that the variant had not been identified by the corresponding in-house pipeline.

ERN GENTURIS

P0009136

This male proband was diagnosed with colorectal adenomatous polyposis in 1991 at the age of 31. However, initial symptoms (hematochezia) had appeared as early as 20 years of age. Colonoscopy revealed that the entire colon and rectum were

covered with a mass of polyps, which were almost exclusively tubular adenomas, suggestive of familial adenomatous polyposis (FAP), the most common type of gastrointestinal polyposis. FAP is an autosomal-dominant precancerous condition, caused by heterozygous pathogenic germline variants in the APC regulator of WNT signalling pathway gene, *APC*. Early detection and removal of adenomas is crucial, as otherwise they invariably result in colorectal cancer (CRC). Adenomas were also found to be present in the duodenum, and there were multiple fundic gland cysts in the stomach. To date the proband has not developed CRC due to prophylactic surgery (proctocolectomy) and frequent surveillance measures. The family history suggests an autosomal dominant inheritance pattern, consistent with FAP: the proband's sister and her son, as well as his own two children have all been affected by early-onset colorectal polyposis. Furthermore, the proband's father died of gastric cancer at 47 years of age, but it is unknown whether he had a polyposis in the upper or lower gastrointestinal tract.

During routine diagnostics, no pathogenic germline SNV or large deletion or duplication was identified in either *APC* or *MUTYH* by Sanger sequencing nor MLPA (kit P043-APC, MRC Holland). Neither did cDNA analysis of the whole transcript reveal any aberrant splicing that may result from deep intronic variants in *APC* (Spier et al., 2012). However, haplotype analysis demonstrated that all affected family members carry the same alleles at the *APC* locus, inherited from the paternal branch of the family. No pathogenic variant was identified in APC or any other potentially relevant gene following ES in a local research project.

Initial reanalysis of the ES data within Solve-RD did not identify any potentially causative germline SNV/InDels either. However, subsequent comprehensive CNV analysis uncovered a heterozygous, approximately 200bp germline deletion, at the beginning of coding exon 15 of the *APC* gene (Supplementary Figure S8), subsequently confirmed by qPCR. It is very likely that this results in an out-of-frame deletion resulting in a premature stop codon, and the variant segregates with the phenotype in the family. As a result of these findings the clinical diagnosis of FAP was confirmed, and predictive genetic testing can be offered to all at-risk members of the family.

Although the clinical course, family history, and haplotype analysis had already pointed to an underlying *APC* variant, the diagnostic deletion was not detected in routine diagnostics due to there being a lack of MLPA probes covering the specific region affected.

MT-AP6 Recurrent Variant Analysis

The mitochondrial variant m.9185T>C affecting the *ATP6* gene has been associated with phenotypic presentations ranging from Leigh syndrome, NARP (neuropathy, ataxia, and retinitis pigmentosa) with variable onset and severity, to mild isolated neuropathy, ataxia and intellectual disability (Castagna et al., 2007; Childs et al., 2007; Ganetzky et al., 2019). The level of heteroplasmy often correlates with the severity of the clinical presentation with >90% heteroplasmy associated with more severe disease, while homoplasmy has also been reported across tissues in some affected individuals. In the present cohort, the mitochondrial chromosome reanalysis identified m.9185T>C in five affected individuals from three unrelated families submitted by ERNs RND, ITHACA and EURO-NMD. The phenotypic presentations varied from severe early-onset Leigh-like syndrome in the ERN ITHACA proband (homoplasmy) to late-onset isolated axonal neuropathy in the ERN EURO-NMD proband (77% heteroplasmy). Intrafamilial variability was observed in the ERN ITHACA family with the mother and sister having relatively late onset and a milder disease course, while the proband presented with severe disease (Supplementary Table S7). Interestingly, near homoplasmic variant levels resulted in very different phenotypes in the three affected individuals from the ERN ITHACA family. Identification of this recurrent variant highlights the importance of the cross-ERN analysis in the present cohort, and the inclusion of mitochondrial variant analysis irrespective of ERN specific gene lists, since mitochondrial diseases often result in multisystemic involvement with variable onset and presentation.

Example of Actionability

Patient P0020778 is a 22-year-old male born in India to non-consanguineous parents who presented with progressive limb girdle weakness with easy fatigability from 2 years of age. Ptosis and swallowing difficulty was observed from 5 years of age. He also has voice fatigability and generalised wasting of limb girdle muscles. Serum creatine kinase (CK) was elevated (>1000 IU) and repetitive nerve stimulation (RNS) showed >10% decrement in quadriceps. Initial clinical suspicion was of limb-girdle myasthenic syndrome.

Patient P0020953 is a 16-year-old male born in India to non-consanguineous parents and presented with ptosis from birth and progressive limb girdle weakness with easy fatigability from 2 years of age. He also reported mild swallowing difficulty and had ankle contractures with thoracolumbar scoliosis. Serum CK was elevated (>2000 IU) and a significant decrement of >10% was observed in RNS of quadriceps. Calf muscle biopsy was reported as muscular dystrophy. He was clinically suspected to have limb-girdle muscular dystrophy / limb-girdle myasthenic syndrome.

Initial whole exome sequencing (WES) analysis in both patients was negative for any significant variants in coding or canonical splice site regions of known neuromuscular disease genes.

Re-analysis of WES data from both patients in Solve-RD (P0020778 – patient and unaffected sibling) and P0020953 -patient and parents), which included analysis of non-canonical splice proximal regions led to the identification of a homozygous intronic variant c.1023+5G>A near exon 5 / intron 5 junction of *DES* (Desmin) in both patients. The variant segregated as heterozygous in unaffected sibling of P0020778 and the parents of P0020953. In silico predictions (spliceAI) indicated a possible donor splicing defect. While undertaking reverse phenotype correlation in both patients and other unsolved CMS patients from the same Indian hospital, another female patient with a similar clinical phenotype was also found to have the same c.1023+5G>A homozygous variant in *DES* which had previously been classified as VUS in a diagnostic genetic laboratory. In view of the significant phenotype correlation in these three unrelated limb-girdle CMS patients, we considered the *DES* homozygous intronic variant as likely causative.

To further validate the pathogenicity, muscle biopsy of left tibialis anterior was performed in patient P0020778 retrospectively, which showed myopathic features and partial loss of Desmin staining on immunohistochemistry. Electron microscopy was suggestive of disorganised myofibrillar architecture with aggregates. RT-PCR of muscle RNA showed the presence of two transcripts: a reduced normal desmin transcript and a longer transcript with intron 5 inclusion suggesting a leaky splice site caused by the +5 variant. Western blotting showed reduced normal desmin protein. Neuromuscular junction involvement with clinically CMS like features have been previously reported in only one family with a homozygous frameshift *DES* mutation resulting in complete loss of desmin and a severe phenotype of childhood onset myopathy, cardiomyopathy and CMS (ref: PMID: 27440146). The identification of this novel intronic *DES* variant not only expands the clinicopathological spectrum of Desminopathies, but also establishes the c.1023+5G>A variant as a possible LG-CMS associated *DES* variant specific to the Indian sub-continent.

Individual P0020778 has reported significant improvement (50%) in their CMS symptoms following treatment of Pyridostigmine and Salbutamol after six months, while patient P0020953 reported no improvement or worsening, having taken the same medication for three months. However, a long-term course of Salbutamol is expected to have some impact on improving fatigable weakness. While Desminopathies currently have no known treatment, CMS drugs including

pyridostigmine and salbutamol can be useful in patients with associated neuromuscular junction involvement.

Supplementary Figures

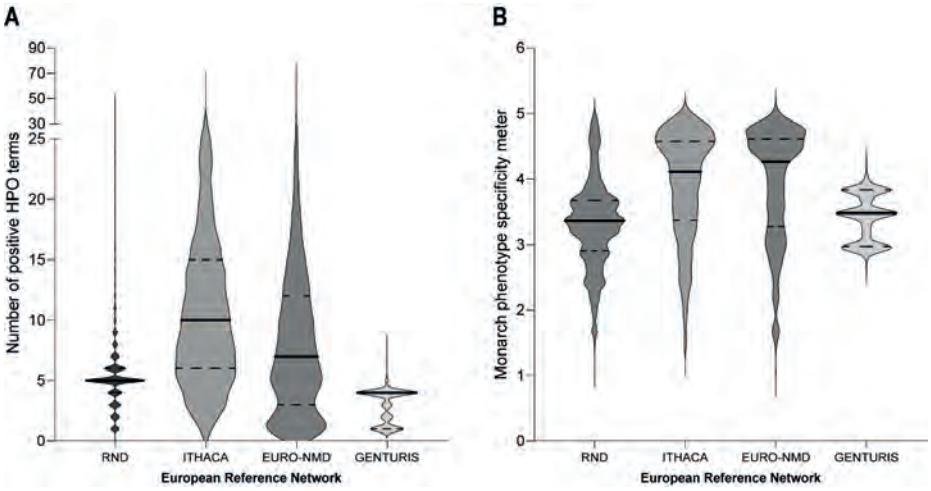


Figure S1: HPO terms (A) and Monarch phenotype specificity meter (B). Violin plots illustrating (A) the number of Human Phenotype Ontology terms associated to each proband across ERN and (B) the Monarch specificity score (range 0-5, higher better) which provides an indication of how comprehensive the phenotypic description of the affected individual is. The solid line indicates the median, and the dashed line the 25th and 75th centiles.

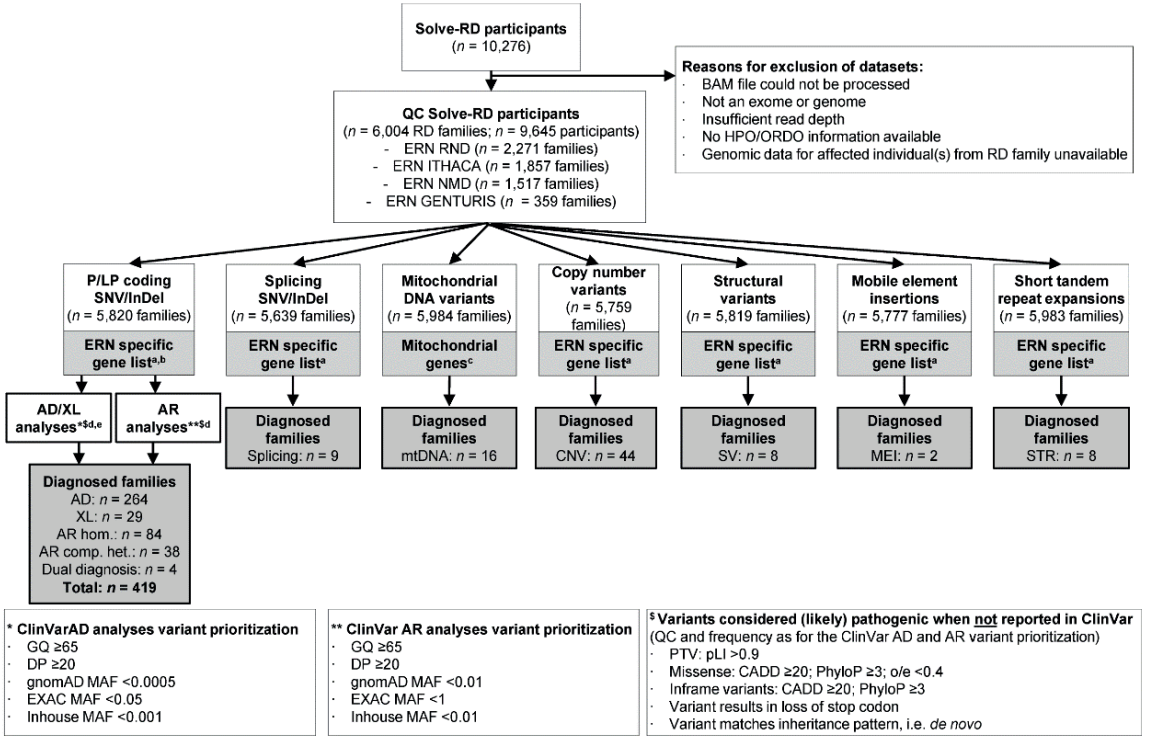


Figure S2: Flowgram of all analyses performed within the Solve-RD systematic reanalysis. a See Supplementary Table S2 for ERN specific gene lists; b *De novo* analysis was performed genome-wide, irrespective of previously identified disease genes; cSNV/InDels were investigated within the mitochondrial DNA; d Small exceptions in the prioritisation were made between ERNs for certain genes. See Online Methods, and Supplementary tables S14-S16 for further details.

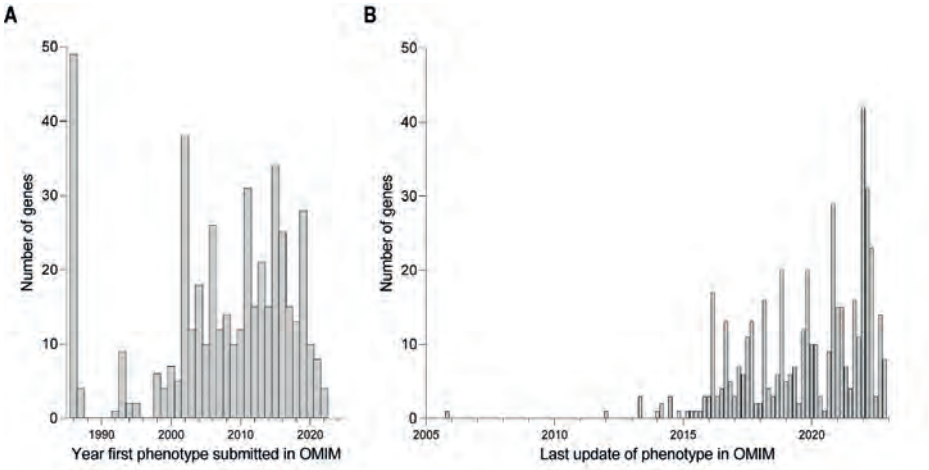


Figure S3: Date of initial creation, and of last update of OMIM records for genes shown to be disease-causing in this study. This figure shows (A) the date of creation of the first OMIM entry for a particular gene determined to be explanatory for the condition in a Solve-RD proband-phenotype association, and (B) the date of the last update of the relevant entry. The OMIM entry for 67 genes was only created after 01/01/2018, when Solve-RD started, and many genes of interest have had their records updated since then. This explains why a number of these genes were only confirmed as being disease-causing in affected individuals here as a result of reanalysis in Solve-RD.

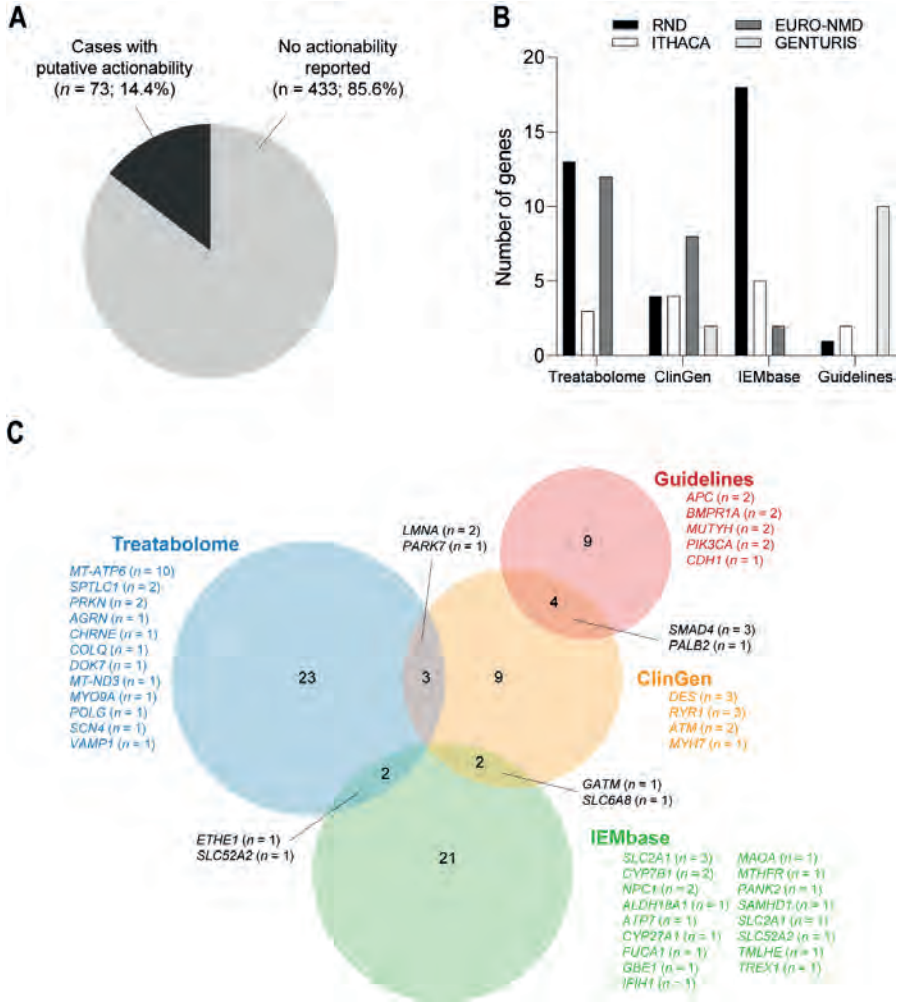


Figure S4: Clinical actionability. A) Percentage of solved cases for which the causative gene is reported in one of the three gene-treatment databases included in this study (ClinGen, IEMbase and Treatabolome) and guidelines for surveillance of genetic tumour risk syndromes. B) Gene-treatment databases and surveillance guidelines for genes in which (likely) disease-causing variants have been identified per ERN. C) List of genes with (likely) disease-causing variants, and number of RD probands/families diagnosed in this study in parentheses, identified in each of the three gene-treatment databases as well as surveillance guidelines included in this study.

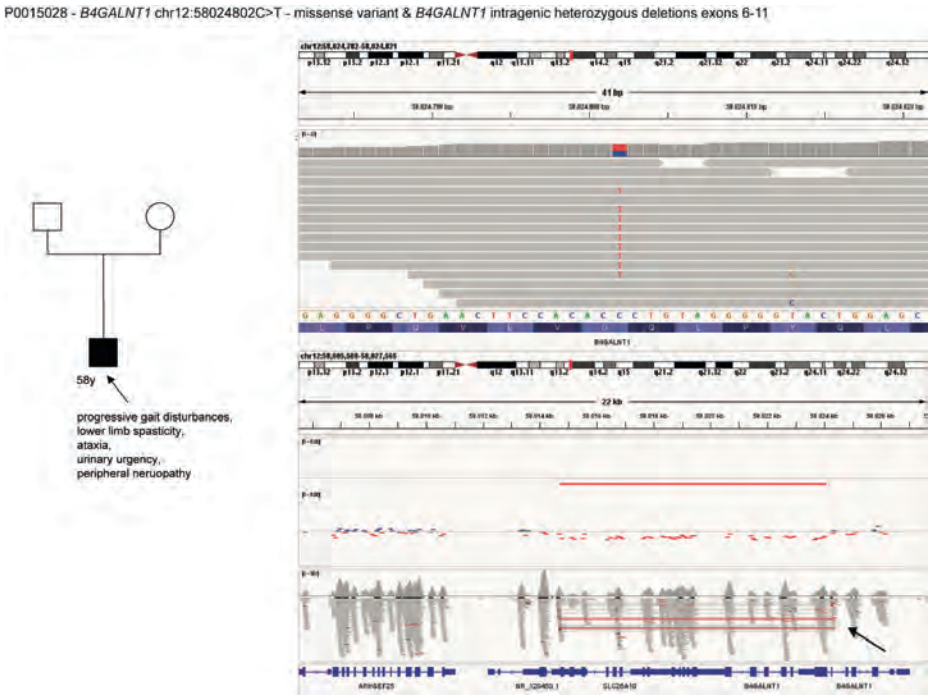


Figure S5: Example of an individual diagnosed with a RD from ERN RND. The left panel shows the pedigree of a 58-year old individual first diagnosed at 42 years of age with progressive gait disturbance and urinary urgency, in the absence of family history of these symptoms (P0015028). The right panel shows two IGV screenshots indicating a heterozygous missense SNV in *B4GALNT1* (top) and a heterozygous, approximately 10kb in length, deletion on the other allele (bottom), resulting in complete deletion of exons 6-11. Location of the deletion is indicated by the red line in the top track, supported by the reduced beta-allele frequency of variants in this region as shown in the centre track, and further supported by read pairs spanning the full 10kb (in red) observed in the lower track.

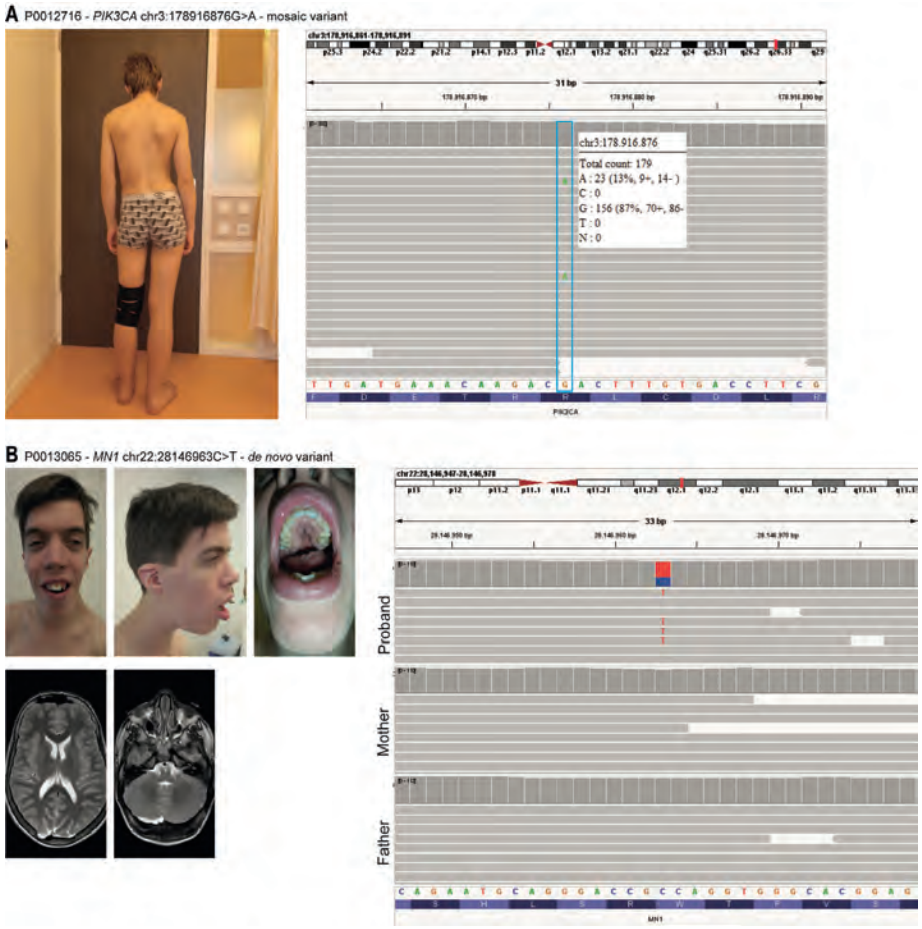


Figure S6: Examples of two individuals diagnosed with RDs from ERN ITHACA**.** A) The left panel shows the phenotypic presentation of a 24-year old male diagnosed at fifteen years of age with asymmetry of legs and face, described at that time as underdevelopment of the left side (P0012716, written consent that allows sharing of photographs was given). The IGV screenshot in the right panel confirms the presence of a rare *de novo* mosaic missense variant (observed in only 13% of reads) in *PIK3CA* (chr3:178916876G>A), validated by Sanger sequencing. This variant had previously been reported elsewhere to cause *PIK3CA*-related overgrowth, leading to a change in the clinical diagnosis for this young man, and the resolution of his diagnostic odyssey. B) The left panel shows the phenotypic presentation of an undiagnosed 22-year old male who had experienced severe developmental delay, and presented with a variety of physical anomalies, though brain MRI was initially reported to be uninformative (P0013065, written consent that allows sharing of photographs was given). The IGV screenshot in the right panel indicates the presence of a rare *de novo* nonsense variant in *MN1* (chr22:28146963C>T) unobserved in the parents. Retrospective reanalysis of the brain MRI revealed dysplasia of the cerebellar vermis, rhombencephalosynapsis and mild bitemporal narrowing of the skull, consistent with a diagnosis of CEBALID syndrome. The individuals described gave permission for their photos to be used in this publication, for which we thank them and their families.

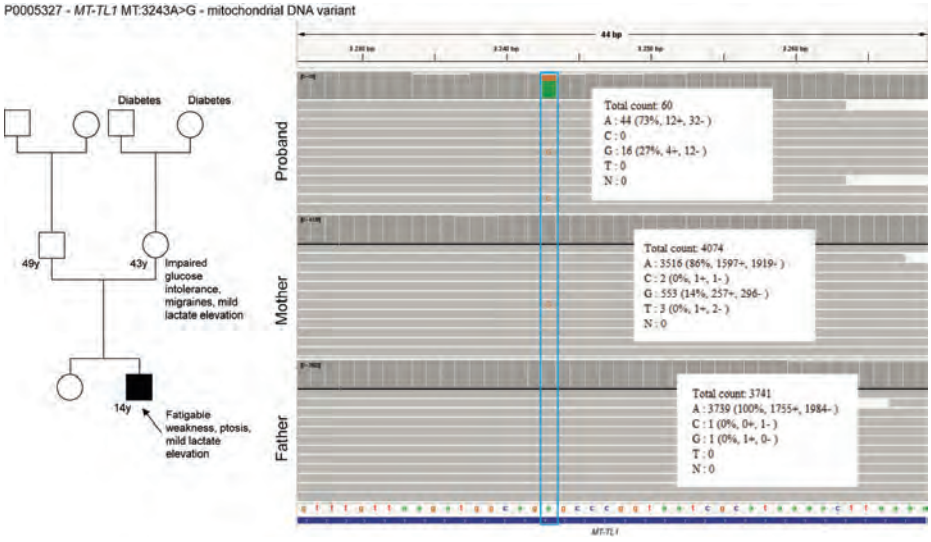


Figure S7: Example of an individual diagnosed with a RD from ERN EURO-NMD. The left panel shows the pedigree, and clinical history of proband P0005327 (indicated by the arrow). At eight years of age he began to develop progressive lower limb weakness and fatigability. The IGV screenshot in the right panel indicates the presence of a heteroplasmic mitochondrial variant (*MT-TL1*, MT:3243A>G)) observed with a frequency of 27% in the proband, and 14% in his mother. This difference may explain the divergence in symptoms between mother and child.

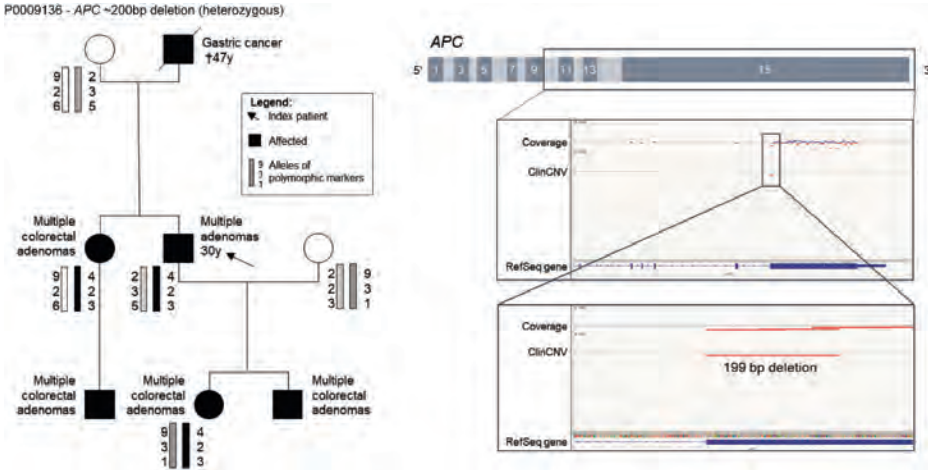


Figure S8: Example of a diagnosed individual with RD from ERN GENTURIS. Left panel: pedigree of proband P0009136 (indicated by the arrow). Haplotype analysis demonstrated that all affected individuals carry the same allele at the APC locus, inherited from the paternal branch of the family. Right panel: comprehensive CNV analysis uncovered a heterozygous germline deletion, approximately 200bp in length, at the beginning of coding exon 15 of the *APC* gene which could not be identified by routine diagnostics using just the sequencing and MLPA methods.

Additional Acknowledgements

The RD-Connect Genome-Phenome Analysis platform developed under FP7/2007–2013 funded project (grant agreement n° 305444) and ongoing funding from EJP-RD (grant numbers H2020 779257, H2020 825575), Instituto de Salud Carlos III (Grant numbers PT13/0001/0044, PT17/0009/0019; Instituto Nacional de Bioinformática, INB), ELIXIR-EXCELERATE (Grant number EU H2020 #676559) and ELIXIR Implementation Studies (Remote real-time visualisation of human rare disease genomics data (RD-Connect) stored at the EGA ELIXIR. 2017-2018; ELIXIR IT-2017-INTEGRATION, Rare Disease Infrastructure ELIXIR, 2019-2020 and the Beacon ELIXIR, 2019-2021). The RD-Connect GPAP has leveraged developments funded through project VEIS (001-P-001647 co-financed by the European Regional Development Fund of the European Union in the framework of the Operational Program FEDER of Catalonia 2014-2020 with the support of the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya) and URD-Cat (PERIS SLT002/16/00174, Departament de Salut, Generalitat de Catalunya).

The Spanish academic and research network RedIris (https://www.rediris.es/) provided the Aspera service used for uploading raw data for processing to the RD-Connect GPAP, and for transferring data between centres.

Netherlands Science Organisations (NWO VIDI 917.164.55 to CG).

The "Network for Italian Genomes - NIG" coordinated by AR and "Cell lines and DNA bank of Rett Syndrome, X-linked mental retardation and other genetic diseases", member of the Telethon Network of Genetic Biobanks (project no. GTB12001), and EuroBioBank network.

HL receives support from the Canadian Institutes of Health Research (Foundation Grant FDN-167281), the Canadian Institutes of Health Research and Muscular Dystrophy Canada (Network Catalyst Grant for NMD4C), the Canada Foundation for Innovation (CFI-JELF 38412), and the Canada Research Chairs program (Canada Research Chair in Neuromuscular Genomics and Health, 950-232279).

This work was furthermore supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) No 441409627, as part of the PROSPAX consortium under the frame of EJP RD, the European Joint Programme on Rare Diseases, under the EJP RD COFUND-EJP N° 825575 (to M.S., R.S, and R.H,) and the Clinician Scientist programme "PRECISE.net" funded by the Else Kröner-Fresenius-Stiftung (to C.W., M.W, R.S. and M.S.)

JPS was financed by Programme EXCELES, (ID Project No. LX22NPO5107) - Funded by the European Union – Next Generation EU.

Chapter 4

Unravelling undiagnosed rare disease cases by HiFi long-read genome sequencing

Wouter Steyaert*, Lydia Sagath*, German Demidov, Vicente A. Yépez, Anna Esteve-Codina, Julien Gagneur, Kornelia Ellwanger, Ronny Derks, Marjan Weiss, Amber den Ouden, Simone van den Heuvel, Hilde Swinkels, Nick Zomer, Marloes Steehouwer, Luke O'Gorman, Galuh Astuti, Kornelia Neveling, Rebecca Schüle, Jishu Xu, Matthis Synofzik, Danique Beijer, Holger Hengel, Ludger Schöls, Kristl G Claeys, Jonathan Baets, Liedewei Van de Vondel, Alessandra Ferlini, Rita Selvatici, Heba Morsy, Marwa Saeed Abd Elmaksoud, Volker Straub, Juliane Müller, Veronica Pini, Luke Perry, Anna Sarkozy, Irina Zaharieva, Francesco Muntoni, Enrico Bugiardini, Kiran Polavarapu, Rita Horvath, Evan Reid, Hanns Lochmüller, Marco Spinazzi, Marco Savarese, Solve-RD DITF-ITHACA, Solve-RD DITF-Euro-NMD, Solve-RD DITF-RND, Solve-RD DITF-EpiCARE, Leslie Matalonga, Steven Laurie, Han G. Brunner, Holm Graessner, Sergi Beltran, Stephan Ossowski, Lisenka E.L.M. Vissers, Christian Gilissen†, Alexander Hoischen† on behalf of the Solve-RD consortium

* These authors contributed equally to this work

†These authors jointly supervised this work

Genome Research. Under review.

Abstract

Solve-RD is a pan-European rare disease (RD) research program that aims to identify disease-causing genetic variants in previously undiagnosed RD families. We utilized 10-fold coverage HiFi long-read sequencing (LRS) for detecting causative structural variants (SVs), single nucleotide variants (SNVs), insertion-deletions (InDels), and short tandem repeat (STR) expansions in extensively studied RD families without a clear molecular diagnosis. Our cohort includes 293 individuals from 114 genetically undiagnosed RD families selected by European Rare Disease network (ERN) experts. Of these, 21 families were affected by so-called 'unsolvable' syndromes for which genetic causes remain unknown, and 93 families with at least one individual affected by a rare neurological, neuromuscular, or epilepsy disorder without genetic diagnosis despite extensive prior testing.

Clinical interpretation and orthogonal validation of variants in known disease genes yielded thirteen novel genetic diagnoses due to *de novo* and rare inherited SNVs, InDels, SVs, and STR expansions. In an additional four families, we identified a candidate disease-causing SV affecting several genes including a *MCF2*/*FGF13* fusion and *PSMA3* deletion. However, no common genetic cause was identified in any of the 'unsolvable' syndromes. Taken together, we found (likely) disease-causing genetic variants in 13% of previously unsolved families and additional candidate disease-causing SVs in another 4.3% of these families.

In conclusion, our results demonstrate the added value of HiFi long-read genome sequencing in undiagnosed rare diseases.

Keywords

LRS Special Issue, long-read sequencing (LRS), rare disease (RD), Solve-RD, undiagnosed RD, structural variants (SVs), short tandem repeats (STRs).

Introduction

Rare diseases (RD) affect 400 million people worldwide (Nguengang Wakap et al., 2020). It is estimated that 80% of these diseases have a genetic origin (Sernadela et al., 2017). Pinpointing the disease-causing genetic variant is important for RD families, because it ends an often time-consuming, stressful, and costly diagnostic odyssey (Biesecker & Green, 2014). In addition, several disease management strategies and treatment options depend on the specific disease gene or variant (Pogue et al., 2018).

With routinely used short-read sequencing (SRS) technologies, such as exome and genome sequencing, diagnostic yields vary between 8% and 70%, depending on the diseases studied and inclusion criteria used (Wright, FitzPatrick, et al., 2018). Recently, a large-scale reanalysis effort of exomes and genomes from undiagnosed disease families has been conducted within the Solve-RD consortium; this study illustrates that updated knowledge and improved variant identification and interpretation substantially increase diagnostic yield (S. Laurie, W. Steyaert, E. de Boer et al., Nat Medicine in press). Other re-analysis efforts have resulted in similar increases in diagnostic yield (Bullich et al., 2022; Liu et al., 2019; Wright, McRae, et al., 2018). Despite this, the majority of tested RD patients remain without a genetic diagnosis. Besides incomplete knowledge of the functional and phenotypic consequence of genetic variation, shortcomings at the variant identification level substantially contribute to the fact that many RD patients remain genetically undiagnosed. Indeed, SRS technologies result in an almost complete characterization of short genetic variants (single and multi-nucleotide substitutions and small insertions and deletions) in the unique portions of an individual's genome, but the analysis of duplicated and repetitive genomic regions and particularly the identification of structural variants (SVs) and short tandem repeat (STR) expansions remain far from complete (Chaisson et al., 2019; Chintalaphani et al., 2021; Porubsky et al., 2023). Several recent studies demonstrate that long-read sequencing (LRS) technologies uncover a whole new reservoir of (structural) genetic variation (Beyter et al., 2021; Chaisson et al., 2019; Kucuk et al., 2023; Pauper et al., 2021; Zook et al., 2020). This is especially true for SVs of intermediate size (50 to a couple of thousands base pairs), which are not identified in SRS data, nor with molecular cytogenetic technologies such as multiplex-ligation dependent probe amplification (MLPA) or comparative genomic hybridization arrays (aCGH). Now that these LRS technologies produce high-quality sequencing reads at steadily dropping costs, researchers are able to evaluate the hypothesis that part of the genetically undiagnosed RDs are caused by variants that remain hidden from previously used technologies. The exploration

and interpretation of SVs in undiagnosed RD families has indeed shown to be successful in the past couple of years for several disease phenotypes (Fadaie et al., 2021; Merker et al., 2018; Mizuguchi, Suzuki, et al., 2019; Mizuguchi, Toyota, et al., 2019; Sanchis-Juan et al., 2018; Zeng et al., 2019). Here, as part of the Solve-RD consortium effort, we applied long-read genome sequencing to two unique patient cohorts. Firstly, a cohort of 21 families (including 16 trios) with clinically well-recognized, so-called 'unsolvable' syndromes, including Aicardi (MIM ID %304050), Hallermann-Streiff (%234100), Gomez-Lopez-Hernandez (%601853), Pai (%155145), and syndromes belonging to the oculoauriculovertebral spectrum, all of which remain genetically elusive despite huge global efforts to identify the disease cause. These patients had not necessarily undergone previous testing. The second cohort consisted of 232 individuals from 93 families with rare neurological, neuromuscular, or epilepsy disorders. While most of these patients are affected by conditions for which several genetic causes are known, these particular families remained 'unsolved'; as extensive diagnostic and/or research testing, including prior exome or genome sequencing had failed to yield a diagnosis.

Results

We analyzed the genomes of 293 individuals from 114 previously undiagnosed RD families using HiFi long-read sequencing (Fig. 1; Supplemental Table S1). Part of the cohort consists of families selected by experts from ERN-ITHACA - mostly parent-offspring trios - with a clinically well-recognizable syndrome termed 'unsolvable' syndromes (n=21). The other part of the cohort (n = 93) are RD families with a rare neurological, neuromuscular, or epilepsy disorder, selected by experts from ERN-EURO-NMD, ERN-EpiCARE, ERN-RND, and ERN-ITHACA.

All of these families and/or syndromes were genetically well-studied in the past using SRS and other applicable approaches, but without diagnostic success (Supplemental Table S2). Consequently, we hypothesized that part of these syndromes are caused by genetic variants, mostly SVs or STRs, that cannot not be identified using SRS or probe-based technologies. Within known disease relevant genes (ERN specific gene lists; Methods) we assessed all types of SVs. Outside these gene panels we focused our analysis on putative de novo events in parent-offspring trios and, due to the lack of effective population databases for SVs, on large inherited SVs (> 100 kb (corresponding to breakend calls)) since these events are more likely to affect the phenotype (Coe et al., 2014; Methods). We also genotyped 56 known disease associated STR loci since these loci are highly relevant

for neurological disease, yet difficult to characterize using SRS techniques. To fully exploit our sequencing data we also identified SNVs. We assessed all rare SNVs within known disease relevant genes and, in parent-offspring trios, on putative de novo events in the complete genome (Methods).

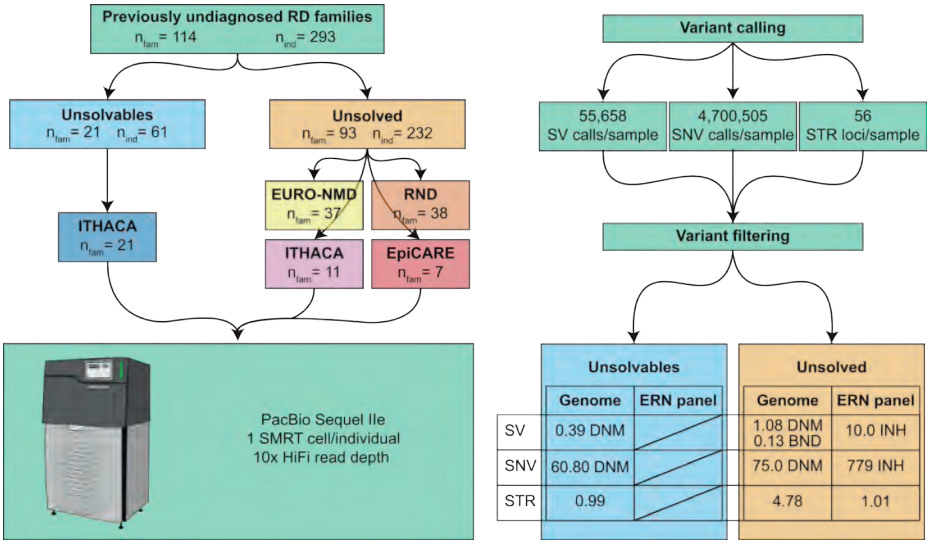


Figure 1: HiFi long-read sequencing in a unique cohort of 293 individuals from 114 RD families. The study cohort consists of two subcohorts: the 'unsolvables' (families affected by clinically well-recognizable syndromes for which the cause is yet unknown) and the 'unsolved' (families affected by a rare neurological, neuromuscular or epilepsy disease). All patients were recruited via four European Reference Networks and subsequently sequenced using a single SMRT cell of sequencing data per individual. Genome-wide calling of SVs and SNVs was conducted and STRs were genotyped at 56 known disease associated loci. Abbreviations: ERN = European Reference Network, BND = Breakend call, INH = Inherited variant, DNM = De novo mutation.

On average, we identified 55,658 SVs (≥ 20 bp; 23,385 SVs ≥ 50 bp) and 4,700,505 SNVs per individual (Methods; Supplemental Table S3). Of these, 13,481 SVs and 43,172 SNVs are private to one family. From the 18 visually curated putative de novo SVs for which flanking sequencing primers could be designed, four were confirmed as de novo variants in the child. In turn, five calls were false positives, six of the variants turned out to be true but inherited from a parent and 3 other variants were true positive too but the parental sequences failed (Methods; Supplemental Fig. S1; Supplemental Table S4,S5).

Identification of (likely) pathogenic variants in previously undiagnosed RD

Unsolvable syndromes

In the subcohort consisting of 21 families with 'unsolvable' syndromes, we could not identify a gene or locus in which rare (de novo) variants were present in multiple families with the same syndrome. However, in a sporadic female patient (P0185637) initially diagnosed with Aicardi syndrome, presenting with global developmental delay, partial agenesis of the corpus callosum, and abnormalities with the vasculature and innervation of the eye, we identified a de novo missense variant in TUBA1A (Tubulin alpha 1a, MIM ID *602529, NM_006009.4, Chr12:g.49,185,725C>T, c.641G>A, p.(Arg214His); Figure 2; Table 1). The variant has previously been described as a cause of lissencephaly 3 (LIS3, MIM ID #611603; (Bahi-Buisson et al., 2014)). Clinical reassessment of the patient's phenotype confirmed the new diagnosis.

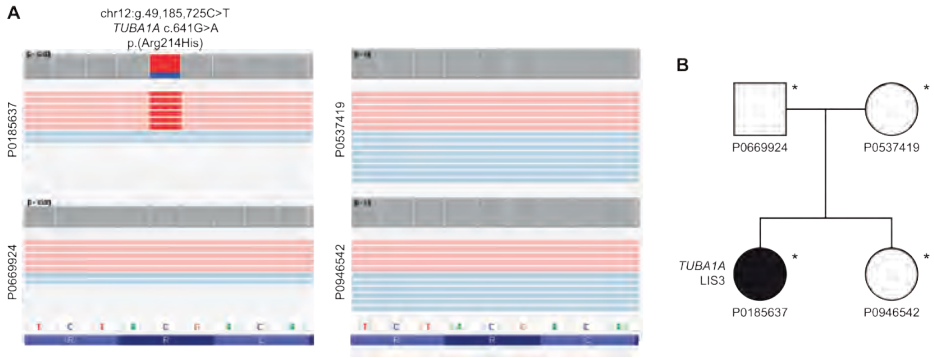


Figure 2: Visualization of the TUBA1A de novo missense variant in P0185637 using IGV, and a pedigree of the family. The variant has earlier been described as a cause of lissencephaly. The healthy family members do not carry the variant. Sequenced individuals are marked with an asterisk (*) in the pedigrees.

Disease-causing variants identified in rare neurological, neuromuscular and epilepsies diseases

After prioritizing and clinically interpreting genetic variants in the 93 families from the 'unsolved' cohort, we established a genetic diagnosis in 12 of them (Methods; Table 1).

Structural variants

In two unrelated male patients (P0078963 and P0695060; Fig. 3A-D) with muscular dystrophy we identified disease-explanatory inversions breaking *DMD* (Dystrophin,

MIM ID *300377, NM_004006.3; Fig. 3A, 3C). The breakpoints in patient P0078963 are ChrX:23,308,848 and ChrX:32,004,110 (hg38) resulting in an inversion of 8.7 Mb which breaks *DMD* in intron 44, resulting in a truncated transcript (Fig. 3A). This event was initially discovered by optical genome mapping (OGM) and LRS detected the exact breakpoints of the event. In patient P0695060 an inversion of ChrX:17,398,320-32,130,845 was identified (Fig. 3C). This event breaks *NHS* (NHS actin remodeling regulator, MIM ID *300457, NM_0129186.2) in intron 1 and *DMD* in intron 44. This event was confirmed by Sanger sequencing, which also highlighted the insertion of a short ATAAT sequence in the first intron of NHS, and of a 38 nucleotide sequence in the intron 44 of *DMD*, which likely favored the inversion. As these genes have opposite orientations on the chromosome, the inversion results in two theoretical fusion genes in which the exon orientation is conserved. However, neither theoretical gene product is in frame much past the fusion breakpoint. The likely disruption of both genes is also in line with the patient's phenotype, who in hindsight presents not only with a dystrophy but also with a cataract which is a characteristic feature of Nance-Horan syndrome (MIM ID #302350) caused by loss-of-function variants of *NHS*.

In a duo consisting of an affected father and affected son (P0011782 and P0011781, respectively; Fig. 3E-H) presenting with hereditary spastic paraplegia, we detected a 1.2 kb deletion encompassing the entire exon 6 of *REEP1* (Receptor expression-enhancing protein 1, MIM ID *609139, NM_001371279.1), Chr2:g.86,232,216-86,233,399del, eventually leading to a frameshift mutation p.(Gly140Cysfs*18) that removes exons 6-9. Both the father and the son are heterozygous carriers of this deletion (Fig. 3F-G). *REEP1* is expressed in larger degrees in all brain tissues, the gastrointestinal tract, and arterial tissues, but has some expression in most other tissue types as well. It regulates lipid droplet formation and the morphology of the endoplasmic reticulum (Renvoisé et al., 2016). Variants in *REEP1* have been described in autosomal recessive distal hereditary neuronopathy (MIM ID #620011) and autosomal dominant spastic paraplegia 31 (SPG31, MIM ID #610250). Of SNVs, frameshift variants are the most common causative variant type in SPG31 cases (Beetz et al., 2008). In addition, single-exon deletions in *REEP1* of exons 2 and 3 have been described as pathogenic (Goizet et al., 2011). Additionally, two cases with deletions encompassing more than one exon have been described (Ishiura et al., 2014), neither affecting exon 6.

In a singleton patient with adult onset distal myopathy (P0657753, Figure 3I-K), a 65kb duplication involving *MYOT* (myotilin, MIM ID *604103, Chr5:137,832,296-137,897,203) had earlier been identified by a gene panel for myofibrillar myopathy.

Table 1: Overview of disease-causing (DC) and candidate disease-causing (cDC) genetic variants in the complete study cohort. Allele frequency databases used: gnomAD v.4.1.0 (SNVs), gnomAD SVs v.4.1.0 (SVs), and gnomAD v.3.1.2 (STRs)

Participant ID	ERN	Cohort	DC/cDC	Gene name(s)	Variant type
P0185637	ITHACA	Unsolvables	DC	*TUBA1A* (NM_006009.4)	(*de novo*) SNV
P0695060	EURO-NMD	Unsolved	DC	*DMD,NHS (*NM_004006.3, NM_001291867.2)	SV (inversion)
P0078963	EURO-NMD	Unsolved	DC	*DMD (NM_004006.3)*	SV (inversion)
P0016368	RND	Unsolved	DC	*NOP56* (NM_006392.4)	STR
P0018996	RND	Unsolved	DC	*NOP56* (NM_006392.4)	STR
P0016356	RND	Unsolved	DC	*DAB1* (NM_001365792.1)	STR
P0019022	RND	Unsolved	DC	*RFC1* (NM_002913.5)	STR
P0008178	EURO-NMD	Unsolved	DC	*DMD* (NM_004006.3)	(deep intronic) SNV
P0016160	RND	Unsolved	DC	*SPAST* (NM_041946.4)	(intronic) SNV
P0631224	RND	Unsolved	DC	*TTN* (NM_001267550.2)	(*de novo*) SNV
P0657753	EURO-NMD	Unsolved	DC	*MYOT* (NM_006790.3)	SV (tandem duplication)
P0237528	EURO-NMD	Unsolved	DC	*REEP1* (NM_001371279.1)	(deep intronic) SNV
P0011781	RND	Unsolved	DC	*REEP1* (NM_001371279.1)	SV (deletion)
P0936700	EURO-NMD	Unsolved	cDC	*FGF13, MCF2 and F9* (NM_004114.5, NM_001171876.2, NM_000133.4)	(*de novo*) SV (duplication)
P0021581	EURO-NMD	Unsolved	cDC	*PSMA3* (NM_002788.4)	(*de novo*) SV (deletion)
P0537031	ITHACA	Unsolved	cDC	*CPE, TLL1, NEK1, CLCN3, ...*	SV (5Mb duplication)
P0016165	RND	Unsolved	cDC	*ARMC9, NCL*	SV (300 kb duplication)

	Inheritance	HGVS	GnomAD allele frequency	Orthogonal validation
	De novo AD	Chr12:g.49185725C>T, c.641G>A, p.(Arg214His)	N/A	-
	XLR;XLD	ChrX:g.17398320_32130845inv	N/A	Sanger
	XLR	ChrX:g.23308848 _32004110inv	N/A	OGM
	AD	Chr20:g.2652734_2652756GGCCTG[1200]	N/A	RNA-Seq
	AD	Chr20:g.2652734_2652756GGCCTG[34]	N/A	RNA-Seq
	AD	Chr1:g.57367044_57367118AAA AT[29]GAAAT[117]AAAAT[615]	N/A	-
	AR	Chr4:g.39348427_39348476delinsAAGGG[1181]; Chr4:g.39348427_39348476delinsAAGGG[271]	N/A	RNA-Seq
	XLR	ChrX:g.33174335C>T, c.31+36947G>A	N/A	Sanger
	AD	Chr2:g.32115840G>A, c.1004+5G>A, p.(spl)	6,24E-04	ES, exon-skipping, Sanger
	AR, Maternally inherited, *de novo* on paternal allele	Chr2:g.178530761dup, c.105854dup, p.(Pro35286Thrfs*13); Chr2:g.178640613del, c.40652del, p.(Pro13551Glnfs*47)	N/A; N/A	SRS
	AD	Chr5:g.137832296_137897203dup	N/A	SRS
	AD	Chr2:g.86327804T>C, c.32+9675A>G	6,57E-03	-
	AD	Chr2:g.86232216_86233399del, c.418-597_595+409del, p.(Gly140Cysfs*18)	N/A	PCR + LRS
	De novo AD/XLR	ChrX:g.139164887_139679311dup	N/A	PCR + LRS + cDNA + RNA-seq
	De novo AD	Chr14:g.58268649_58283944del	N/A	PCR + Sanger
	N/A	Chr4:g.165447976_170473344dup	N/A	Array CGH, ES
	AD	Chr2:g.231348004_231684006dup	N/A	-

Because we did not observe this variant in our filtered variant calls, we reverted to the raw sequencing data for this specific case, which allowed us to determine that the duplication was in tandem. The variant also segregates with the probands similarly affected sibling. Heterozygous variants in *MYOT* are a known cause of myofibrillar myopathy 3 (MIM ID #609200), a slowly progressive muscle disorder with an adult onset.

Repeat expansions

In two families with autosomal dominant (AD) ataxia, we identified disease-explanatory heterozygous expansions of the GGCCTG motif in intron 1 of the *NOP56* gene (Table 1; NOP56 ribonuclear protein, NM_006392.4, MIM ID *614154). Repeat expansions in *NOP56* are a known cause of AD spinocerebellar ataxia 36 (SCA36, MIM ID #614153; (Kobayashi et al., 2011)). The hexanucleotide motif count in a duo consisting of two affected siblings (P0016368 and P0018504; Fig. 4A) was estimated at >1200. In the other family, consisting of two affected family members in two generations and one unaffected family member (P0018996, P0019023, and P0019024, respectively, Fig. 4B), the motif count was >34 in the affected mother and >45 in the affected child. The pathogenic repeat threshold of *NOP56* is generally regarded to be 650 hexanucleotide repeats, however, shorter repeats are also known to be causative (Obayashi et al., 2015). The repeat expansion in the latter family was also discovered by reduced expression through RNA-sequencing and whole genome sequencing by parallel efforts in Solve-RD (Supplemental Fig S2).

In a family presenting with autosomal dominant ataxia (P0016356, P0019033), we found a repeat expansion in *DAB1* (DAB adaptor protein, NM_001365792.1, MIM ID *603448), a known causal gene for Spinocerebellar ataxia 37 (SCA37, MIM ID #615945). Age-dependent penetrant alleles have been reported to have an insertion of 31-75 ATTTC repeats, while the normal motifs are usually uninterrupted and consisting of 7-400 units of ATTTT (Matilla-Dueñas & Volpini, 1993). The analysis of LRS data indicated the presence of two alleles in the index case P0016356, one with 7 ATTTT repeats and another with a complex structure of (estimated) 615 ATTTT motifs, followed by 117 ATTTC repeats and then again by 29 ATTTT repeats (Fig. 4C), supported by 5 high quality spanning PacBio reads.

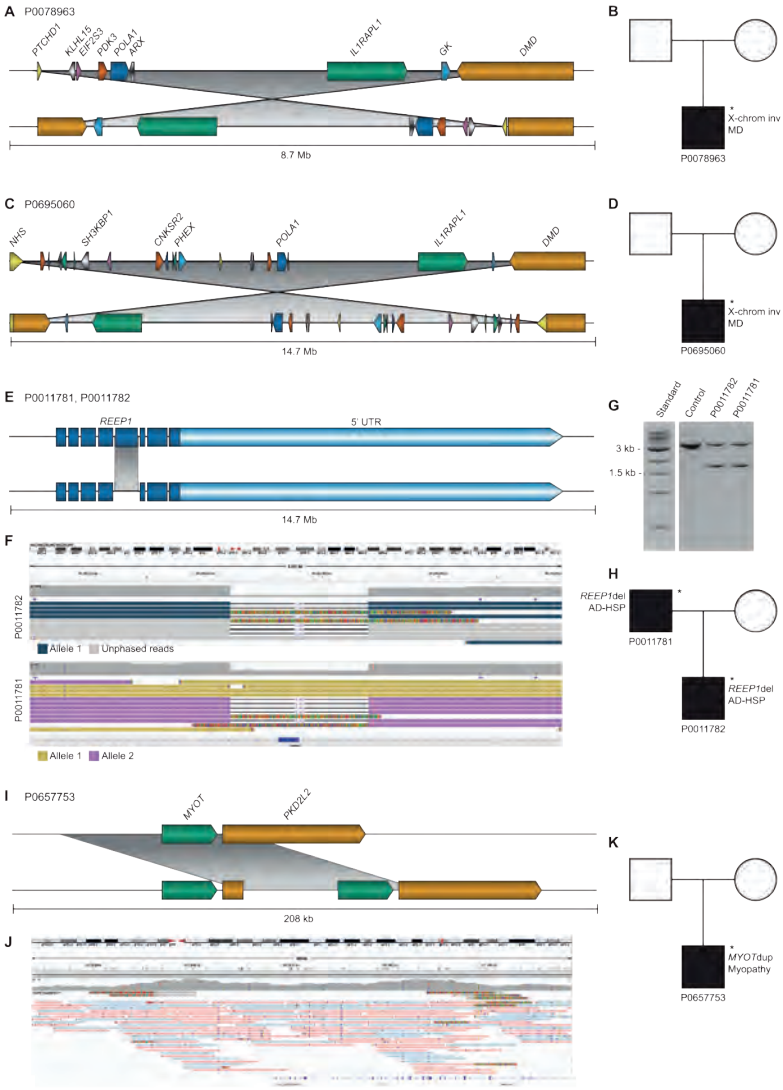


Figure 3: Visualization of disease-causing SVs in the 'unsolved' cohort in the form of cartoons and/or IGV screenshots, along with corresponding pedigrees. In two unrelated male patients (P0078963 in A and B, P0695060 in C and D) with muscular dystrophy, we found X-Chromosomal inversions (A-D). In both cases, *DMD* is disrupted (A and C), in one a second gene disruption adds to the phenotype (C). In a father and son with hereditary spastic paraplegia, we detected a deletion of *REEP1* exon 6 (E-H). The deletion in P001782 and P0011781 is shown here as a cartoon (E) and as a screenshot in IGV (F). The deletion was also visualized by agarose gel electrophoresis, which confirms that both patients are heterozygous for the deletion (G). The pedigree of the family is shown in (H). In a patient with adult onset distal myopathy, a 65 kb duplication involving *MYOT* (I) was confirmed to be in tandem by LRS (J). The pedigree of the family is shown in K. Sequenced individuals are marked with an asterisk (*) in the pedigrees (B, D, G, H, K). Abbreviations: MD = muscular dystrophy, AD-HSP = autosomal dominant hereditary spastic paraplegia.

Finally, a homozygous repeat of the pathogenic AAGGG motif in the *RFC1* gene (Replication factor C, NM_002913.5, MIM ID *102579) was found in a patient with ataxia (P0019027). Repeat expansions in RFC1 are known to cause CANVAS-spectrum disorder ("Cerebellar ataxia, neuropathy, and vestibular areflexia syndrome", MIM ID #614575), being very consistent with the observed phenotype. The number of AAGGG motifs was estimated by the tool to be 271 on one allele and 1181 on the other allele (Fig. 4D). However, it is possible that the first allele is longer than 271 pathogenic repeats, since it was inferred based on soft clipped reads, not reads spanning the full repeat. The visualization of LRS data from this patient in Integrative Genome Viewer (IGV) indicated that no normal alleles were present. The further validation of this likely causative repeat is prepared to be described elsewhere.

Single nucleotide variants

In a sporadic male patient with suspected titinopathy (P0008178), presenting with progressive proximal muscle weakness, and myopathic features in his muscle biopsy, we identified a deep intronic SNV in *DMD* (ChrX:33,174,335C>T) (Fig. 5A). This variant has previously been shown to be a cause of Becker muscular dystrophy (BMD, MIM ID #300376) through exonization of a 149 bp sequence within intron 1 of *DMD* (Okubo et al., 2020). Clinical reassessment of the patient's phenotype confirmed the BMD diagnosis.

In a duo consisting of an affected mother and daughter (P0859417, P0016160) with AD hereditary spastic paraplegia (HSP), we identified a variant in intron 6 of *SPAST* (Spastin, MIM ID *604277, NM_014946.4, Chr2:g.32,115,840G>A, c.1004+5G>A) (Fig. 5B). Variants in *SPAST* are known to cause HSP4 (MIM ID #182601; (Hazan et al., 1999)) and while the same variant has not been previously recorded, a variant affecting the same base has previously been evaluated as pathogenic in ClinVar (variation ID 989101). The variant was identified in parallel by the referring laboratory but was initially considered to be of uncertain significance. Subsequent RNA analysis eventually demonstrated skipping of exon 6 showing that the variant is likely pathogenic through loss-of-function exon-skipping.

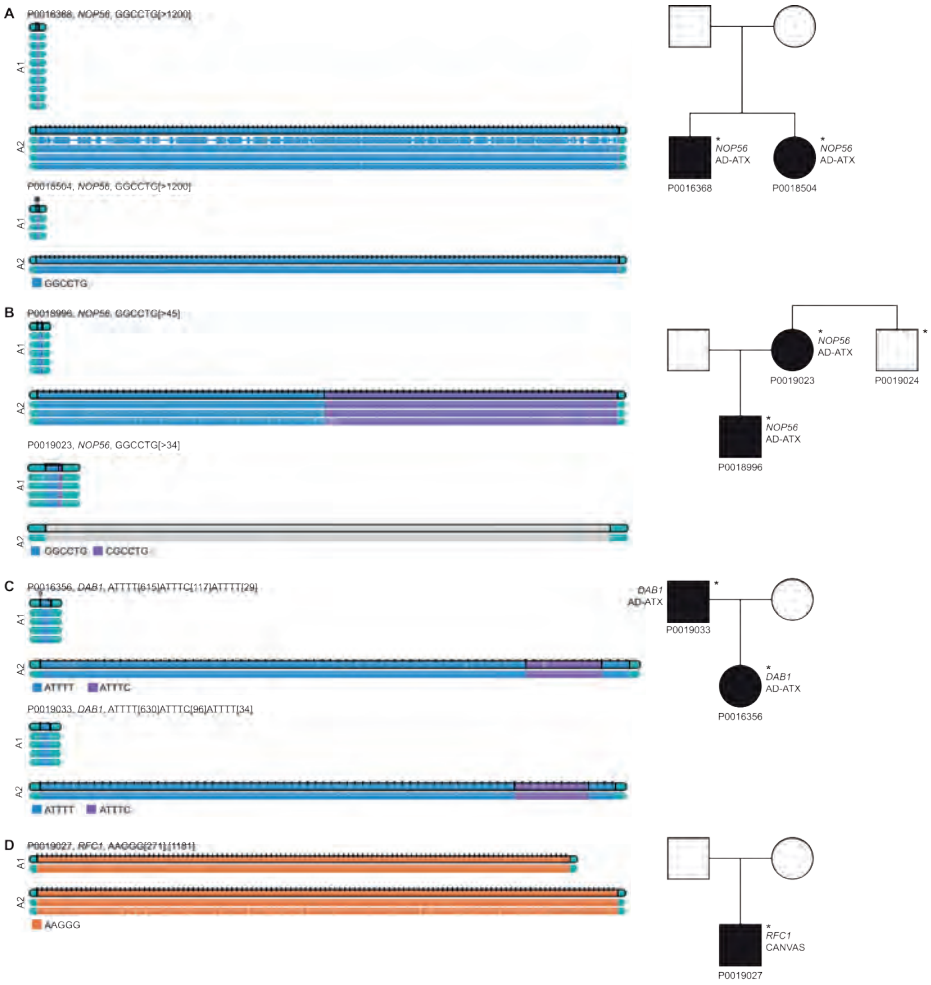


Figure 4: Visualizations produced using the PacBio TRGT tool and pedigrees for the families with pathogenic STR expansions. In siblings P0016368 and P0018504, a heterozygous GGCCTG expansion in NOP56 was detected (A). In another family, an expansion including the motifs GGCCTG and CGCCTG in NOP56 was detected in one generation (P0018996), and the STR expansion was subsequently also identified in the mother (B). In patient P0016356 and their father, we identified heterozygous STR expansion DAB1 including both ATTTT and ATTTC motifs (C). In another patient, we identified homozygous STR expansions in *RFC1*. Alleles are denoted by "A1" and A2". Sequenced individuals are marked with an asterisk (*) in the pedigrees. Abbreviations: AD-ATX = autosomal dominant ataxia.

In a sporadic patient with suspected titinopathy (P0631224), two pathogenic frameshift variants in *TTN* (titin, MIM ID *188840, NM_001267550.2) had been previously identified before submission to the Solve-RD collection. Of these, one was maternally inherited, Chr2:g.178530761dup, c.105854dup, p.(Pro35286Thrfs*13), and one de novo, Chr2:g.178640613del, c.40652del, p.(Pro13551Glnfs*47) (Fig. 5C). Both variants are located in ubiquitously expressed exons; the maternally inherited variant affects the constitutional exon 308, and the de novo variant affects exon 221, which is expressed in 99% of *TTN* transcripts (Savarese et al., 2018). Previous SRS efforts had not been successful in identifying on which allele the de novo event had occurred. Using our approach, we were able to successfully differentiate between the alleles and confirmed the two frameshift variants to be in trans, thus explanatory for the patient's phenotype (Fig. 5D).

In a family with suspected autosomal dominant hereditary spastic paraplegia (AD HSP), we identified a deep intronic substitution in the first intron of *REEP1* (Chr2:g.86327804T>C, NM_001371279.1:c.32+9675A>G), segregating in the affected mother and son. Previous genetic analysis with HSP and hereditary neuropathy panels was negative. The variant is predicted to alter splicing by activation of a cryptic donor site by Human Splicing Finder and MaxEntScan (Desmet et al., 2009; Yeo & Burge, 2003). Loss-of-function variants including splice-altering intronic variants in *REEP1* have previously been reported as causative in AD-HSP families (Züchner et al., 2006).

Candidate disease-causing variants identified in rare neurological, neuromuscular and epilepsy diseases

In addition to the pathogenic variants identified above, in which the disease gene is well established and fits the patient's phenotype according to clinical experts, our analyses revealed novel, likely pathogenic aberrations in four additional families (Table 1).

De novo duplication on Chromosome X

In a female patient (P0936700) presenting with arthrogryposis multiplex congenita, thoracolumbar scoliosis, and restrictive ventilatory defect, we discovered a 500 kb X-chromosomal tandem duplication (ChrX:g.139,164,887_139,679,311dup), which was confirmed *de novo* in the patient by gel electrophoresis and sequencing (Fig. 6A-D). The breakpoints of the duplication disrupt two genes*; FGF13* (fibroblast growth factor 13, MIM ID *300070, NM_004114.5), and *MCF2* (Cell line-derived transforming sequence, *311030, NM_001171876.2).

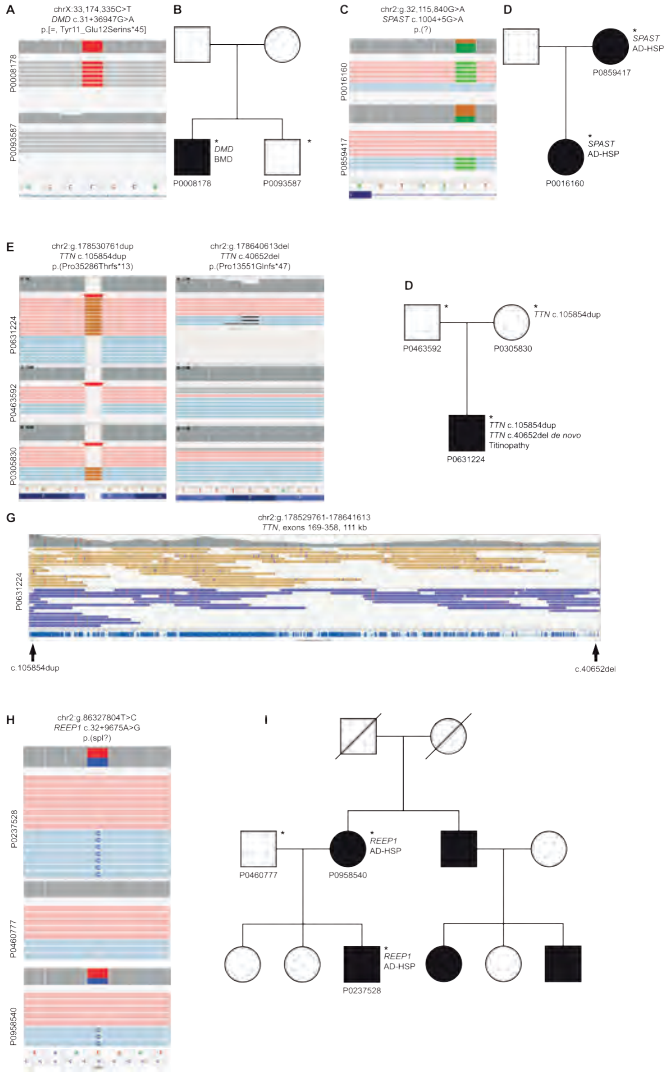


Figure 5: Visualization of disease-causing SNVs and InDels in the 'unsolved' cohort in the form of IGV screenshots, along with corresponding pedigrees. In a sporadic patient with suspected titinopathy, we identified a deep-intronic variant in *DMD* (A). The nonaffected sibling did not carry the variant (A and B). In a duo consisting of an affected mother and affected daughter with hereditary spastic paraplegia, we identified a non-canonical splice site variant in *SPAST* (C and D). In a patient with titinopathy (P063122), a maternally inherited and a *de novo* variant had been identified earlier (C-F). The two variants are located 109 kb apart, but the alleles were successfully phased through the entire region by LRS (G). In a family with an affected mother and son (P0847234 and P0958540), we detected an intronic variant in *SPAST* (H). The index patient also has an affected uncle, whose sample was not sequenced (I). The reads are colored by haplotag; pink and light blue, or yellow and purple represent different alleles in A, C, E, G, and H. Unphased reads, such as X-Chromosomal reads in males, are shown in grey (A and E). Sequenced individuals are marked with an asterisk (*) in the pedigrees (B, D, F, I). Abbreviations: BMD = Becker Muscular Dystrophy, AD-HSP = autosomal dominant hereditary spastic paraplegia.

FGF13 modulates the function and location of voltage-gated sodium channels in the brain (Fry et al., 2021). Mutations in the gene have been linked to developmental and epileptic encephalopathy 90 (MIM ID #301058) and intellectual developmental disorder (MIM ID #301095). *MCF2* is an oncogene belonging to a family of GDP-GRP exchange factors, the role of which is to modulate the activity of small GTPases in the Rho family. In addition to brain tissue, it has relatively high expression in the adrenal gland, the testes and the ovaries. Molinard-Chenu and colleagues reported a putative pathogenic missense mutation in *MCF2* in a patient presenting with complex perisylvian syndrome and demonstrated that murine Mcf2 controls the migration of cortical projection neurons in mice (Molinard-Chenu et al., 2020).

The duplication results in a hypothetical *FGF13-MCF2* fusion gene, in which the breakpoint resides within the second intron of *FGF13* and the first intron of *MCF2* (Fig. 6B). The second exon of *MCF2* is usually partially untranslated, but it is assumed that it is entirely translated in the fusion gene product; in this case, the entire fusion gene product is in-frame. The putative pathogenic mechanism of this fusion gene will be subject for another study.

PSMA3 C-terminal deletion

In a sporadic female patient (P0021581) we identified a *de novo* 15.3 kb deletion on Chromosome 14 affecting the three last exons (ex 9-11) of *PSMA3* (Proteasome subunit, Alpha type, 3, MIM ID *176843, NM_002788.4) (Fig. 6E-H), Chr14:58,268,649-58,283,944. The phenotype consists of marked delay of psychomotor development resulting in achievement of independent walking at the age of 3 years. The patient displays facial dysmorphism and marked intellectual disability. From the age of 21 there was progressive worsening of motor functioning. Nerve conduction studies revealed an axonal sensorimotor neuropathy. Unaffected siblings of the index patient did not carry the deletion, and haplotyping of *PSMA3* suggests that the deletion has arisen de novo in the patient. The deletion breakpoints were confirmed by Sanger sequencing, and its absence in the siblings was confirmed by PCR and gel electrophoresis (Fig. 6G).

PSMA3 is expressed in tissues throughout the body, including skeletal muscle and nerve tissues. As a proteasome subunit, the role of *PSMA3* is to contribute to the proteolytic pathway of aberrant proteins and/or proteins with high turnover rates in the ubiquitin-proteasome system (UPS). Variants in *PSMA3* have not previously been linked to disease and no structural variants only affecting *PSMA3* are described in any public databases. However, variants in genes contributing to the UPS have been linked to several neurodegenerative diseases caused by the aggregation of

neurotoxic proteins in the absence of a functioning ubiquitin-proteasome system. Biran and colleagues have proposed that the *PSMA3* C-terminal region targets intrinsically disordered proteins for degradation, and would thus play an important role in the ubiquitin-proteasome system (Biran et al., 2022). The loss-of-function observed/expected upper bound fraction (LOEUF) for *PSMA3* is 0.28 (gnomAD v.2.1.1), suggesting that the gene is likely important for normal function.

De novo duplication on Chromosome 4

In a singleton female patient (P0537031) with a congenital malformation syndrome, we identified a 5 Mb tandem *de novo* duplication on Chromosome 4 (Fig. 6I-J). The patient presented with a complex phenotype involving growth delay, facial syndromic features with optical and neurological involvement, cleft palate, and tonic-clonic seizures. The duplicated sequence is Chr4:165,447,976-170,473,341, and involves several known disease-causing genes, amongst which *NEK1* (MIM ID *604588), and *CLCN3* (MIM ID *600580).

While none of the known disease-causing genes within the duplicated region can be directly tied back to the phenotype of the patients, some overlap is present. Variants in *NEK1* are a known cause of a form of thoracic dysplasia (short-rib thoracic dysplasia 6 with or without polydactyly, MIM ID #263520). This syndrome involves cleft palate, and enlargement of the lateral ventricles; however, it is also characterized by several clinical manifestations not present in the patient. In turn, missense variants in *CLCN3* are a known cause of autosomal dominant neurodevelopmental disorder with seizures and brain abnormalities (MIM ID #619512). This disorder is characterized, among other symptoms, by dysmorphic facial features, hypertelorism, strabismus, abnormalities of the cerebellum and corpus callosum, and, in some patients, seizures.

300 kb duplication on Chromosome 2

In a family presenting with hereditary spastic paraplegia, we identified a 300 kb tandem duplication on Chromosome 2 in the affected father (P0016174) and an affected son (P0016165) (Fig. 6K-L). The non-affected brother of the son (P0018356) does not carry the duplication. The duplicated sequence is Chr2:231,348,004-231,684,006, with a breakpoint within *ARMC9* (armadillo repeat-containing protein 9, MIM ID *617612, NM_001352754.2) and containing *NCL* (nucleolin, MIM ID *164035, NM_005381.3), among other genes. RNA-seq confirmed upregulation *NCL* within the duplicated sequence.

Nucleolin is an ubiquitously expressed, major nucleolar protein in growing eukaryotic cells, and plays a role in the regulation of ribosomal RNA transcription, ribosome maturation and assembly, and transportation of ribosomal components between the nucleus and cytoplasm. It is predicted to be intolerant to loss-of-function variants (pLI 1.00) and dosage sensitive (LOEUF 0.18). In addition, variants in *ARMC9* are a known cause of Joubert syndrome 30 (MIM ID #213300), a genetically heterogeneous neurodevelopmental ciliopathy (Van De Weghe et al., 2017). Individuals with Joubert syndrome present with ataxia, along with hypotonia, abnormal eye movements, and cognitive impairment (Latour et al., 2020).

Discussion

The Solve-RD consortium has set itself the goal to genetically diagnose previously undiagnosed rare disease families. In our current study, HiFi long-read genome sequencing was conducted for 293 carefully selected patients and healthy relatives from 114 rare disease families. The cohort was further divided into two subcohorts, consisting of 'unsolvable' syndromes in which we expect the molecular cause to be currently unknown, and 'unsolved' families with RNDs or NMDs in which we expect to identify new variants in known or novel disease genes.

Whereas sequencing was performed at relatively modest coverage of ~10-fold, we identified and orthogonally validated pathogenic variants of all classes: SNVs, InDels, SVs and STRs. Although our approach is not ideally suited for obtaining highly comprehensive SNV call sets, strict filtering, interpretation and validation of calls shows that previously unidentified and/or misclassified SNVs also contribute to the diagnostic yield of our study. For example, we identified a previously unidentified deep-intronic hemizygous *DMD* variant leading to altered splicing in a male patient with muscular dystrophy, as well as an intronic *SPAST* variant segregating in a family with hereditary spastic paraplegia, which was not considered for clinical interpretation in the initial exome analysis.

In two cases, our study did not identify a novel variant but provided additional information about previously identified candidate variants. In the case of a titinopathy patient, a pathogenic maternally inherited mutation and a single base pair *de novo* deletion in *TTN* had already been identified. However, the diagnosis was inconclusive because the allele on which the *de novo* mutation had occurred could not be determined. The long reads of this study allowed for the phasing of this variant and could confirm that it occurred on the paternal allele thereby leading

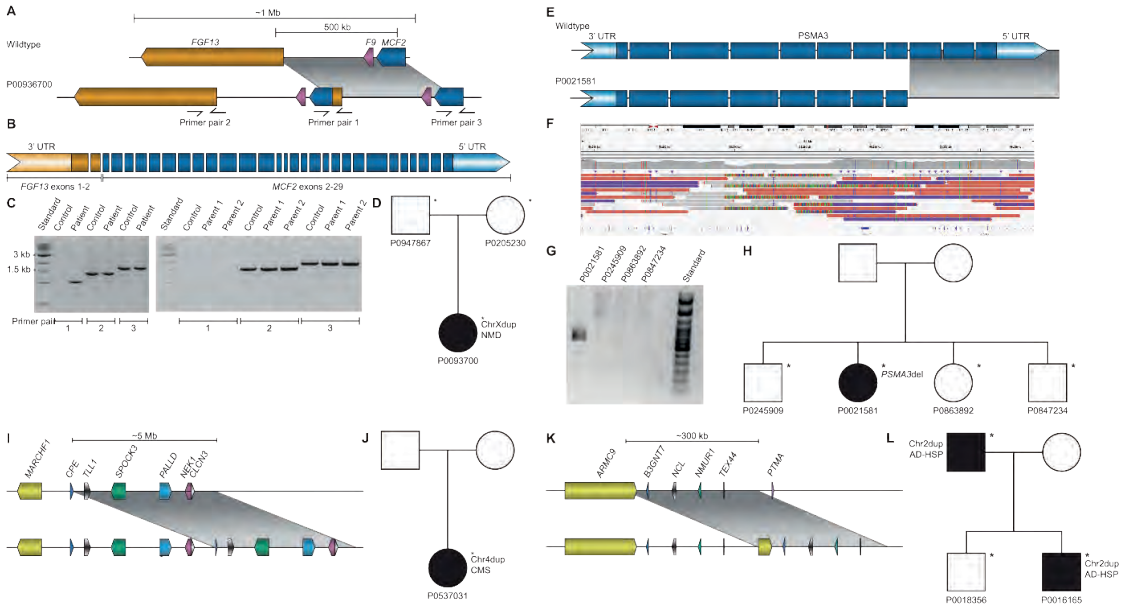


Figure 6: Visualization of candidate disease-causing SVs in the 'unsolved' cohort. In a sporadic patient (P0093700) with arthrogryposis multiplex congenita, we detected a X-Chromosomal tandem duplication (A-D). The duplication spans from intron 1 of MCF2 to intron 2 of FGF13 and also includes F9 (A). The result of the duplication is a hypothetical fusion gene including FGF13 exons 1-2 and MCF2 exons 2-29 (B). The duplication was validated by PCR and agarose gel electrophoresis (C) using primers targeting the breakpoints of the duplication, and a combination of the MCF2 forward and FGF13 reverse primers (Primer pair 1, A and C). In a sporadic patient with psychomotor development delay (P0021581) we detected a deletion of *PSMA3* exons 9 to 11, here shown as a cartoon (E) and as a screenshot using IGV (F). The deletion was validated by agarose gel electrophoresis (G) and Sanger sequencing. The index P0021581 has three healthy siblings who do not carry the deletion (H). In a sporadic female patient (P0537031) with congenital malformation syndrome, we detected a 5 Mb tandem duplication on Chromosome 4, visualized here as a cartoon (I). In a sporadic male patient (P0016165) with autosomal dominant spastic par aplegia, with a similarly affected father, we detected a 300 kb tandem duplication on Chromosome 2, visualized here as a cartoon (K). Sequenced individuals are marked with an asterisk (*). Abbreviations: NMD = neuromuscular disorder, CMS = congenital malformation syndrome, AD-HSP = autosomal dominant hereditary spastic paraplegia.

to a definite diagnosis. Similarly, a previously detected gain involving *MYOT*, was shown here to be a tandem-duplication.

In total, we identified disease-causing variants in 12 out of 93 families with unsolved disorders (13.0%), but only one in the 21 families (4.8%) with 'unsolvable' disorders. The variant found in the 'unsolvable' trio is a *de novo* SNV in *TUBA1A* in a patient initially suspected of Aicardi syndrome. The identified *TUBA1A* variant has previously been associated with LIS3 (Bahi-Buisson et al., 2014). In hindsight, and by reverse phenotyping, the clinical experts in this project also confirmed this as the disease causing genetic variant in this specific case. Individuals with LIS have severe neurological problems, including intellectual disability and epilepsy, and may appear phenotypically very similar to patients with Aicardi syndrome. The difference in the number of resolved cases between the two cohorts suggests that 'unsolvable' syndromes indeed are a special class of syndromes. In such cohorts other explanations for the disorder should be considered, such as methylation defects, somatic mutations, polygenic origin, larger heterogeneity than expected, or even non-genetic causes of disease (Boycott et al., 2018).

Next to the 13 diagnoses we also identified four candidate disease-causing SVs: one intragenic deletions in *PSMA3*, two large duplications (a 5Mb event breaking and involving multiple coding genes, and a 300kb event affecting the *ARMC9* and *NCL* genes) and an X-Chromosomal duplication likely leading to the production of a *FGF13-MCF2* fusion protein.

Of these, two events, the intragenic deletion in *PSMA3* and the X-Chromosomal duplication were de novo events. Although for *PSMA3*, parental samples were not available, we were able to infer the de novo status of the deletion by using the long-reads to reconstruct the haplotype on which the deletion occurred and observing that the deletion was not present in other siblings with the same haplotype. The low SV mutation rate in humans (strong evolutionary constraint against SVs) implies that such genic *de novo* SV events are good candidates for pathogenicity (Porubsky & Eichler, 2024).

The *de novo* deletion affects the last three exons of *PSMA3* in a sporadic patient with a phenotype similar to Charcot-Marie-Tooth disease type 2. The deletion likely results in truncated mRNA and subsequent nonsense mediated decay. Mutations in *PSMA3* have not been described in literature, however, the gene encodes for a subunit of the ubiquitin proteasome system (UPS). Variants in other genes encoding for UPS subunits have been described as causative in several neurodegenerative

disorders; in these, the absence of a functioning ubiquitin-proteasome system leads to the aggregation of neurotoxic proteins. Moreover, the gene is highly intolerant to loss-of-function variants, making it a likely candidate for a dominant disease gene. *PSMA3* is (lowly) expressed in whole blood and future gene expression analysis in patient material would likely provide further support for the disease-causing nature of the deletion.

In the patient with the 300 kb tandem duplication, RNA-Seq showed significant upregulation of *NCL* expression. While *NCL* is not a characterized disease gene, its role in RNA biology coupled with its intolerance score predictions for both loss-of-function variant and dosage sensitivities renders it a good candidate as the pathogenetic underlier of the phenotype in the family.

The de novo duplication on the X-Chromosome of a patient with arthrogryposis multiplex congenita, thoracolumbar scoliosis, and restrictive ventilatory defect likely results in a fusion *FGF13-MCF2* gene. *FGF13* is a X-linked dominant disease gene associated with neurodevelopmental disorder phenotypes, different from the phenotype observed in this female patient. *F9* is not known to be associated with a disease once duplicated, and also *MCF2* has not been associated with disease yet. We can hypothesize that the creation of a *FGF13-MCF2* fusion gene and especially the simultaneous loss of an *FGF13* wildtype allele may have a phenotypic consequence for this patient. However, further investigations, including X-Chromosome inactivation and functional studies, will be needed to understand the relationship between this event and the patient's phenotype.

Out of the four candidate variants, three concern variants affecting already established disease genes, although *ARMC9* is an unlikely causative gene in the family harboring the 300 kb duplication on Chromosome 2. Confirmation of the pathogenic nature of these variants may broaden the known phenotypic spectrum of the affected genes or establish new inheritance patterns. The *de novo* deletion affecting *PSMA3* suggests a potential novel neurodevelopmental disease gene, and the significant upregulation of *NCL* warrants more studies in its putative pathogenicity in the family with hereditary spastic paraplegia.

While it is challenging to compare diagnostic rates across different studies due to varying factors such as cohort sizes, patient inclusion criteria, sequencing methods, depth of sequencing, and analysis techniques, it can be cautiously suggested that diagnostic rates in LRS (long-read sequencing) reanalysis studies such as ours tend to be higher than those in SR-WGS (short-read whole genome sequencing)

reanalysis studies. This is likely because a subset of variants identified in LRS cannot be detected or are incompletely characterized with SR-WGS, for example inversions and STRs.

One factor currently limiting the diagnostic yield in LRS studies is the clinical interpretation of the large number of identified "rare" SVs. Large catalogs of identified variants from long-read sequencing of both affected and unaffected individuals will therefore be of critical importance to improve variant interpretation in such cases. Control population efforts are underway, for example in the All of Us Research Program (Mahmoud et al., 2024) or initiatives such as colors-db (https://colorsdb.org/). Such control cohorts may, in the future, help to diagnose additional patients in our cohort. Here, Solve-RD shares the full dataset, including expert curated pedigree and phenotype information (EGA: EGAD00001008602). In addition, we also share a frequency call set of high-quality SVs of the unrelated individuals as a resource for other researchers (Methods; EGA: EGAD00001008602). This resource shall prove valuable in particular with the increase in novel variant types for which LRS has higher sensitivity.

Recent studies suggest similar benefits of long-read sequencing for diagnosing previously undiagnosed rare disease cases. In similarly heterogeneous groups of rare diseases the authors demonstrated disease-causing or candidate disease-causing variants in 16/96 cases (16.7% (Hiatt et al., 2024)). In more homogenous disorders, such as sensorineural hearing loss, a study has shown diagnostic yield in 4/19 (21%) of included cases (Redfield et al., 2024). Further advancements in sequencing technology and bioinformatic algorithms will eventually make LRS a comprehensive test capable of accurately detecting all genetic variants (Ebert et al., 2021; Höps et al., 2024).

In conclusion, HiFi long-read genome sequencing was conducted for a unique cohort of 293 individuals from 114 previously studied rare disease families. While we did not identify a common genetic cause in any of the 'unsolvable' syndromes, we identified causal genetic variants in 13.0% of families from the 'unsolved' cohort, and candidate variants in an additional 4.3%. Our study shows the potential and effectiveness of even modest-coverage LRS in rare disease studies.

Methods

Study cohort

HiFi long-read genome sequencing was conducted for 293 individuals from 114 genetically undiagnosed rare disease families (Supplemental Table S1). Patient samples came from two sub-cohorts: the 'unsolvables' (n=61) for which genetic causes remain unknown, and the 'unsolved' (n=232) for which a previously hidden genetic variant in a known or yet unknown disease gene is expected to be the cause of disease. All of the patients and healthy relatives were carefully selected by experts from four European Reference Networks: RND (38 families; 95 individuals), EURO-NMD (37 families; 89 individuals), ITHACA (32 families; 88 individuals; including all 'unsolvables'), and EpiCARE (7 families; 21 individuals). Depending on the research hypothesis and sample availability 1 to 7 (un)affected individuals were selected per family for sequencing on a PacBio Sequel IIe instrument. The most represented family structure is the parent-offspring trio (nfamilies= 42; nsamples = 126; 43.0% of cohort). We have used a single SMRT cell of sequencing data per individual which, after read alignment (onto hg38) and read filtering, resulted in a mean HiFi read depth of 9.8 (Supplemental Table S6).

DNA sequencing

Genomic DNA was isolated from peripheral blood according to standard protocol and long-read genome HiFi sequenced using SMRT sequencing technology (Pacific Biosciences, Menlo Park, CA, USA). For every sample, 7-15 µg of DNA was sheared on Megaruptor 2 or 3 (Diagenode, Liège, Belgium) to a target size of 15-18 kb. Libraries were prepared with SMRTbell Prep Kit 2.0 or 3.0. Size selection was performed using a BluePippin DNA size selection system (Sage Science, Beverly, MA, USA) targeting fragments equal to or longer than 10 kb in length. Sequence primer and polymerase were bound to the complex using the Sequel II binding kit 3.2 (PacBio), and sequencing was performed on the Sequel IIe system with 2.0 Chemistry and 30h movie time per SMRTcell using a single flow cell per sample.

Primary data analysis

All samples were processed in the same fashion using a custom workflow based on standard methods from the Pacific Biosciences analysis pipeline (https://github.com/PacificBiosciences/pb-human-wgs-workflow-snakemake) (Supplemental Fig. S3). Sequencing reads were aligned to the GRCh38/Hg38 genome with pbmm2 (version 1.4.0 (Li, 2018b, 2021)) using default parameters. HiFi reads (>QV20) were extracted for all downstream analyzes. Small variant (substitution and indel) calling was performed using DeepVariant (version 1.1.0) with default settings

(Poplin, Chang, et al., 2018). No threshold for maximum size of the indels was applied, and all indel calls were used for further analyses. For parent-offspring trios, GLNexus (version 1.3.1) was used to conduct SNV joint genotyping (Yun et al., 2021).

Small variants were phased using Whatshap (version 1.1.0) and variants were annotated using an in-house developed pipeline (M. Martin et al., 2016). This variant annotation was based on the Variant Effect Predictor (VEP V.91) and Gencode 34 basic gene annotations. STR calling was performed using Tandem Repeat Genotyper (TRGT; version 0.3.3) at 56 known disease associated STR loci (Supplemental Table S7 (Dolzhenko et al., 2024)). SV calling was performed using PBSV (version 2.4.0) using default settings with a minimum SV size of 20 bp (https://github.com/PacificBiosciences/pbbioconda). SVs were annotated using AnnotSV (version 3.1.1 (Geoffroy et al., 2018)).

In each of the 114 RD families that comprise our study cohort we selected the maximum number of unrelated individuals resulting in a subcohort of 166 unrelated individuals. SVs merging using Jasmine resulted in a call set of 251,672 unique SVs (corresponding to 11,290,783 variant alleles in the subcohort) of which 59,876 are private to one individual (Kirsche et al., 2023). Only 1,971 unique SVs (0.78%; 51,433 alleles) in the complete call set affect a coding exon. An additional 2,965 unique SVs (1.18%; 111,231 alleles) alter the non-coding sequence of an exon and 95,197 unique SVs (37.8%; 4,445,817 alleles) reside in an intron of a protein coding gene. Lastly, 35,525 unique SVs (14.1%; 1,638,574 alleles) affect a non-coding gene.

The distribution of sequence gains and losses in our study cohort is characterized, as expected, by the inverse relation between SV length and frequency. Exceptions to this smooth decrease in density are the characteristic peaks for short interspersed nuclear elements (SINE) and long interspersed nuclear elements (LINE) peaks which are respectively present at medians +/- 323 bp and +/- 6,050 bp (Supplemental Fig. S4).

Variant filtering

Structural Variants

In parent-offspring trios we focused on putative *de novo* variants. For this, we selected sites which are covered by at least 8 HiFi reads in each of the members of the trio. Furthermore, at least 3 HiFi reads should support the variant allele in the child, and only SVs which are unique in the study cohort were retained (Supplemental Table S3). Because of the modest sequencing depth we subjected all

of the resulting SVs to visual inspection using IGV. In this step, we removed SV calls that were unclear in the child (despite the variant call), SVs for which one of the parents had a trace in their sequencing data (for example supported by one read) and SVs for which both or one of the parents only had 1 allele sequenced (based on the phased alignments). All of the remaining sites were subjected to primer design for further validation (cf. wet-lab validation).

In all of the other family structures we focused on rare inherited high-quality SVs that co-segregate with disease. To do so, we selected family-unique SV calls which were observed in all affected members of a given family and absent from all unaffected family members. Furthermore, in at least one of the affected family members the SVs needs to be covered by at least 8 HiFi reads of which 3 support the variant allele. In contrast to SV calls corresponding to well-characterized deletions, inversions, duplications and insertions we visually inspected all breakend-calls. We evaluated, based on coverage and on the complexity of the sequence context, whether or not a breakend-call could, together with the linked breakend-call, be a signature of a genetic event which is too large to be characterized as a deletion, inversion, duplication or insertion by pbsv. In this step, we required that all clipped reads support the same regional split. Since these calls support relatively large genetic events (> 100 kb) we clinically assessed them in the complete human genome. In contrast, clinical interpretation of SV calls corresponding to characterized deletions, inversions, duplications and insertions (size < 100 kb) was restricted to events that reside in genes within recently curated ERN-specific gene lists (S. Laurie, W. Steyaert, E. de Boer et al., Nat Med in press).

Single nucleotide variants

In parent-offspring trios we focused on putative *de novo* events. These were selected from the joint calls generated by GLnexus. We considered a variant to be putative de novo when the child is heterozygous (genotype '0/1') and both parents are homozygous for the reference allele (genotype '0/0'). In addition, we require that both parents and the child have ≥ 8 HiFi reads covering the site of which 3 reads support the variant allele in the child.

In all other families we selected for rare inherited SNVs that co-segregate with disease. For this, we selected SNVs which are unique to a single family that are present in all affected family members and absent from all unaffected family members. In contrast to the *de novo* variant interpretation we restricted variant interpretation for inherited variants to variants that reside in genes incorporated in the recently curated ERN-specific gene lists.

STR genotypes

STR genotypes were visualized in R per submitter group in comparison with the rest of the cohort in order to facilitate the evaluation of quality of calling per locus and the detection of pathogenically expanded alleles. These results were sent to the groups for clinical interpretation.

Wet-lab variant validation

Altogether 35 variants called as *de novo* were selected for validation using targeted LRS (Supplemental Fig. S1). Primers for the validations were designed using the online Primer3 design tool as per the manufacturer's suggestions. Primers were selected to be 18-21 nucleotides in length with a GC-content ranging from 40-60%. While an annealing temperature of 60°C was proposed to be optimal, annealing temperatures between 57-61°C were considered to be acceptable as well. Sizes of the products ranged between 1000-4000 nucleotides, to ensure capture of the full region and compatibility with PacBio LRS.

In three cases, two adjacent SVs could be covered by one primer pair, and for the large X-Chromosomal duplication, altogether three primer pairs were designed (Fig. 6). For 9 variants, primer design was not possible, resulting in a total number of 23 primer pairs designed for de novo variants (Supplemental Table S5). In addition to these, primers were also designed for the inherited candidate exon 6 deletion in *REEP1* and a 50 bp deletion in *MAPK8IP3* segregating with disease in P0016368 and P0018504.

The running conditions were optimized in gradient runs using melting temperatures of 60, 61, and 62°C and extension times appropriate to product length using LongAmp HotStart Taq 2x MasterMix (New England Biolabs, Ipswich, MA, USA). All successfully optimized primer pairs (n = 18) were run on patient, control, and parental DNA. The samples were cycled as follows: 94°C 30 seconds; 27 cycles of 94°C 30 seconds, 60-62°C 1 minute, 65°C 1 minute 40 seconds (short), 3 minutes 20 seconds (long) or 5 minutes (ultra); 65°C 10 minutes; 4°C hold. The short program was used for amplicons under 2500 bp, the long program for amplicons between 2500 and 3500 bp, and the ultra-long program for amplicons over 6000 bp. The PCR products were verified by agarose gel electrophoresis.

All successfully amplified patient samples were validated by targeted LRS. Subsequent sequencing of parental samples was performed as per the workflow above for samples in which the variant call was confirmed in the index.

Data access

All raw and processed sequencing data generated in this study have been submitted to the European Genome-phenome Archive (EGA; https://ega-archive.org/) under accession number EGAD00001008602.

Competing interest statement

The authors declare no competing interests

Acknowledgements

The Solve-RD consortium is grateful to all involved RD patients and their families as well as other contributors to Solve-RD.

The Solve-RD project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 779257. This research is supported (not financially) by several ERNs: ERN on Intellectual disability, TeleHealth, Autism and Congenital Anomalies (ERN ITHACA)—Project ID No 101085231; ERN on Rare Neurological Diseases (ERN RND)—Project ID No 101155994; ERN for Neuromuscular Diseases (ERN Euro-NMD)—Project ID No 101156434; and ERN for Rare and Complex Epilepsies (ERN EpiCARE) - Project ID No 101156811. The ERNs are co-funded by the European Union within the framework of the Third Health Program. We would also like to thank all other Solve-RD colleagues that were not mentioned by name in the author list, including members of the Solve-RD data interpretation task force (DATF), and other members of ERNs and DITFs.

We also thank The Radboud Technology Center Genomics for the library preparation and sequencing of all samples.

V.A.Y. and J.G. received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via the project NFDI 1/1 "GHGA - German Human Genome-Phenome Archive" (No 441914366). L.Sa. received funding from the Sigrid Jusélius Foundation (Fellowship No 220540).

R.S. received funding from the Bundesministerium für Bildung und Forschung (BMBF) through funding for the TreatHSP network (grant 01GM2209A) and the National Institute of Neurological Disorders and Stroke (NINDS) under Award Number R01NS072248.

H.H. was supported by the DFG (HE8803/1–1 to H.H.).

H.L. receives support from the Canadian Institutes of Health Research (CIHR) for Foundation Grant FDN-167281 (Precision Health for Neuromuscular Diseases), Transnational Team Grant ERT-174211 (ProDGNE) and Network Grant OR2-189333 (NMD4C), from the Canada Foundation for Innovation (CFI-JELF 38412), the Canada Research Chairs program (Canada Research Chair in Neuromuscular Genomics and Health, 950-232279), the European Commission (Grant # 101080249) and the Canada Research Coordinating Committee New Frontiers in Research Fund (NFRFG-2022-00033) for SIMPATHIC, and from the Government of Canada Canada First Research Excellence Fund (CFREF) for the Brain-Heart Interconnectome (CFREF-2022-00007).

Supplementary figures

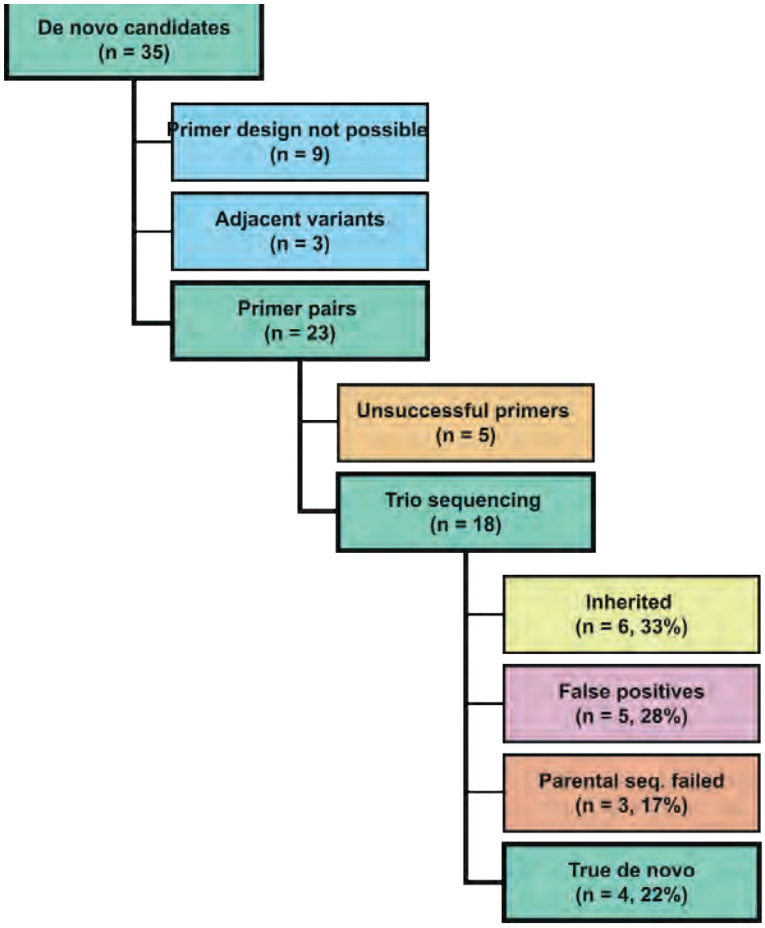


Figure S1: Flow of *de novo* variant validations by targeted LRS.

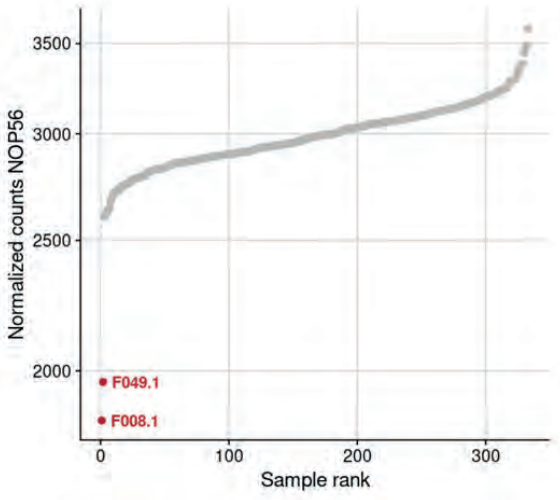


Figure S2: RNA-seq outlier analysis for *NOP56*. The vertical axis in the figure displays the normalized *NOP56* read count, the horizontal axis indicates the rank of the samples according to the sorted normalized read count. Two samples show aberrant mRNA expression: P0018996 and P0018504 (affected sister of the index P0016368). The reduction in expression with respect to the mean across the Solve-RD RNA-seq cohort is 35% for P0018996 and 39% for P0018504. In both cases this is indicative for the loss of one *NOP56* allele.

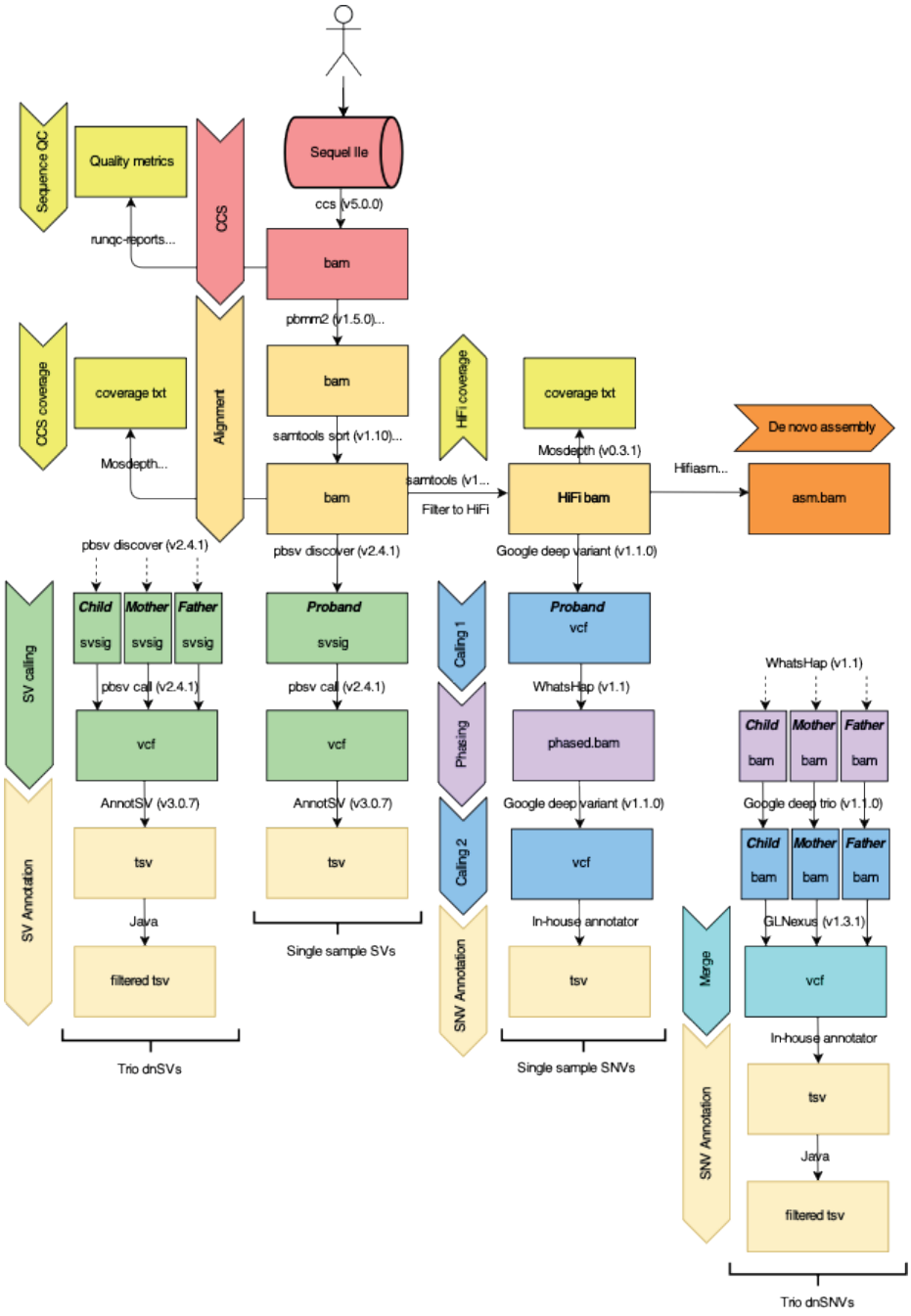


Figure S3: Schematic overview of the analysis workflow

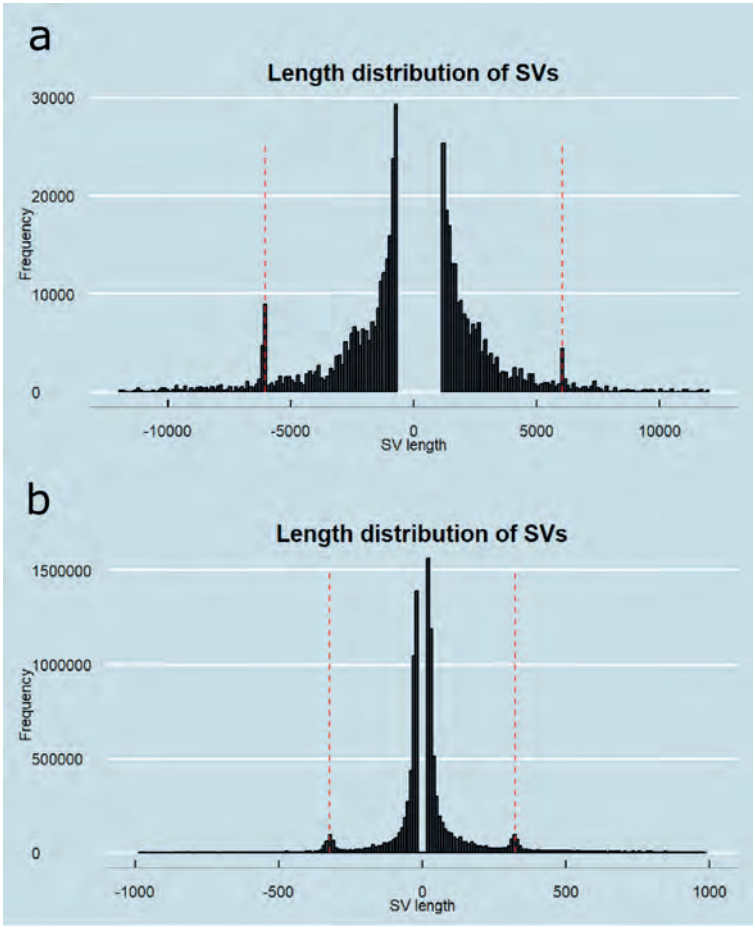
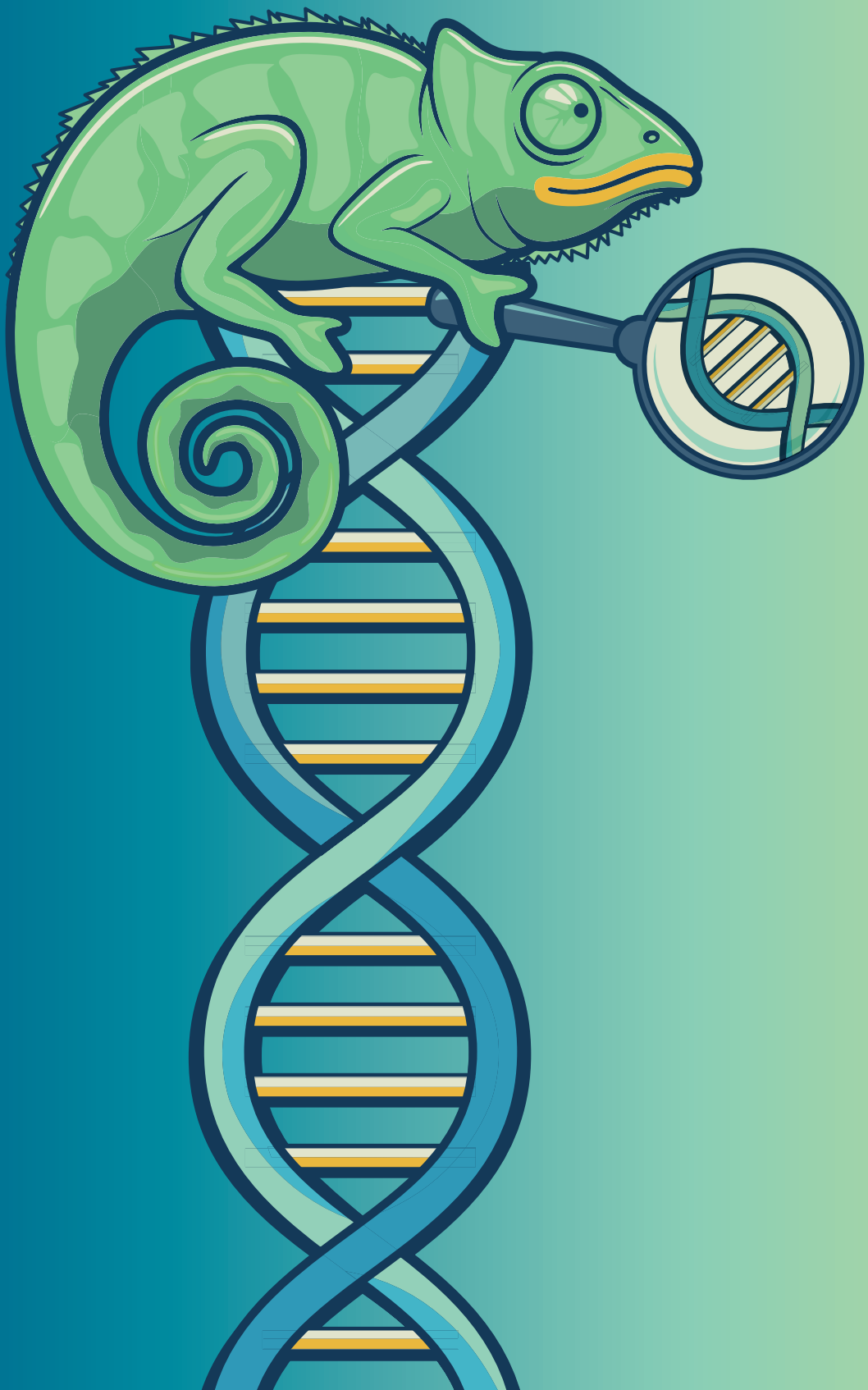


Figure S4: SV length distribution for sequence gains (insertions and tandem duplications) and sequence losses (deletions) visualized using different length and frequency ranges. In panel a the full length spectrum of sequence losses and sequence gains is visualized. In this plot, the peaks corresponding to the long interspersed nuclear elements become apparent at plus and minus 6,050 bp. Similarly in panel b where the horizontal axis ranges from -1,000 to +1,000 bp, the peaks corresponding to the short interspersed nuclear elements become apparent at 323 bp.



Chapter 5

General discussion

Due to a multitude of technological and biological factors the disease-causing genetic variant is identified in less than half of RD patients whose genome is sequenced (Nguengang Wakap et al., 2020; **chapter1**). The past five years I developed and applied approaches to tackle several of these limiting aspects. On the one hand we addressed challenges related to the identification of genetic variants in sequencing data (using novel bioinformatics methods in **chapters 2 and 3**, and by using the latest sequencing technologies in **chapter 4**) while on the other hand we collaborated with clinical experts from across Europe to improve clinical variant interpretation (**chapters 3 and 4**; Figure 1). While my work substantially contributes to an improved variant identification and interpretation, it is only one step forward on an ever-continuing road towards a complete understanding of our genes and phenotypes.

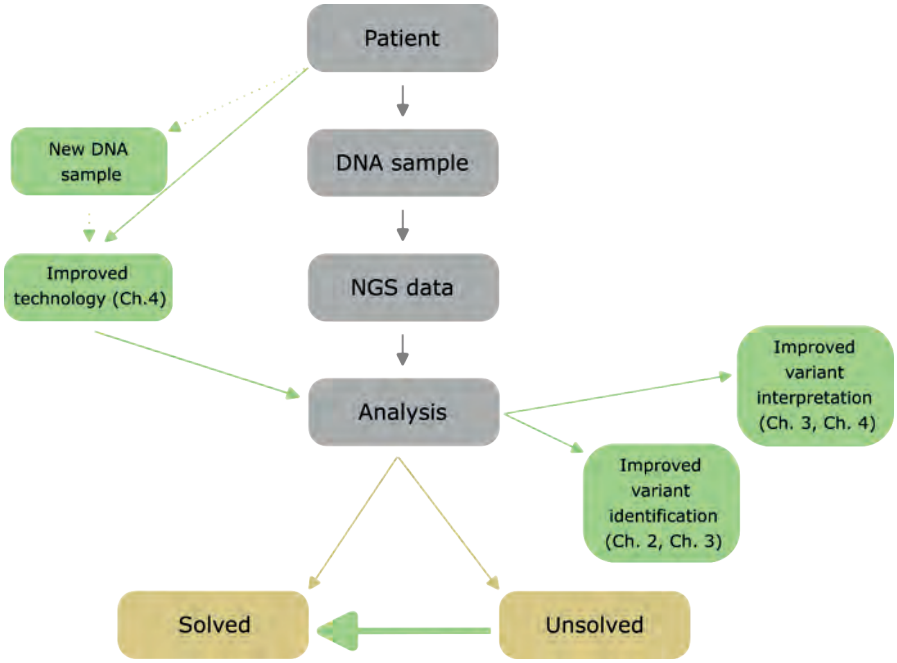


Figure 1: Schematic representation of the research conducted in this thesis. Rectangles with grey background colour represent a simplified flow chart of routine genetic testing. The work presented in this thesis contributes to improving genetic diagnostics with the ultimate goal of enhancing the diagnosis of rare diseases in more patients (rectangles with green background). Specifically, we have focused on improved variant identification in Chapters 2 and 3, and on enhanced variant interpretation in Chapters 3 and 4. Since not all genetic variants can be detected in previously generated data, we have also employed new technologies to generate novel data that are better suited for identifying variants that would otherwise remain hidden, particularly structural variants. The research in this thesis will contribute to the short- and medium-term integration of these new aspects into existing routine diagnostic practices, ultimately leading to an increased diagnostic yield.

Towards the identification of all genetic variants in an individual

In recent decades, major steps have been undertaken to determine the sequence of a complete human genome starting from DNA fragments of 50-150 base pairs. In **chapter 1** I introduced some of the main challenges associated with short-read sequencing. I also touched on long-read sequencing technologies which use fragments of several tens of kilobases, and strongly emerging now due to their greatly improved data quality and continuously decreased sequencing costs. Here, I further elaborate on some of these challenges and discuss the relevance and implications of the solutions developed during this PhD.

Paralogous genomic regions

Duplicated genomic regions cannot be analysed accurately with conventional next-generation sequencing analysis techniques. **Chapter 2** describes the design and application of a method (Chameleolyser) that enables the identification of genetic variants in the 3.5% of the exome that is characterized by sequence paralogy. Using Chameleolyser, an average of 60 previously hidden genetic variants are identified in a single exome. The vast majority of these variants reside in sets of identical paralogs and can therefore not be determined unambiguously using short-read sequencing data. In case one of the possible variants that could have caused the variant call (hence we call them variants with ambiguous positions) is of (clinical) interest, a downstream validation experiment needs to be conducted to verify in which of the paralogs the variant resides. For these repetitive genomic regions, the design of sequencing primers that result in the unique amplification of the region of interest is most often challenging, undoubtedly a major reason that prevented us from making more genetic diagnosis.

In the last couple of years, technologies have been developed and optimized that allow for the targeted enrichment of long DNA molecules without the need to design sequencing primers. Xdrop is a droplet-based sequence capture technology in which sequences of interest are discriminated from other DNA fragments by fluorescent labels (Madsen et al., 2020). Other strategies take advantage of the sequence specificity of CRISPR/Cas9 systems (Fiol et al., 2022; McDonald et al., 2021; Schultzhaus et al., 2021). Possibly in the future, these technologies can be used in tandem with algorithms such as Chameleolyser, further boosting diagnostic success in genetically undiagnosed patients.

In addition to the challenges that we faced to validate variants of interest, we figured out that the sensitivity of Chameleolyser for SNVs/Indels in identical paralogs is 43%, a major increase when compared to regular analysis techniques (where this sensitivity is exactly 0%), but far from perfection. Therefore, we can expect that many more genetic diagnoses hide within these difficult-to-analyse regions. An observation that ties in with this is the fact that the density of known disease genes among protein coding sequences within these regions is substantially lower when compared to the unique portions of the genome. Possibly, this reflects the technical challenges that are associated with the identification of genetic variants in these regions.

Besides the identification of genetic variants in identical paralogs, Chameleolyser also identifies homozygous deletions and homozygous ectopic gene conversions in nearly identical homologs. In case of an ectopic gene conversion event, a conventional read aligner will map the short sequence reads that originate from the acceptor site onto the reference sequence corresponding to the donor site. As a result, the acceptor site appears to be deleted while no deletion is present. Furthermore, the variants that are introduced into the acceptor site by means of this gene conversion are not identified. Chameleolyser conducts a coverage analysis of the uniquely aligned reads, jointly for all paralogs within a set of paralogs. Consequently, homozygous deletions and ectopic gene conversions are identified in a site-specific manner. To our knowledge, Chameleolyser is the first method that enables an exome-wide screening of ectopic gene conversions. Within the 41,755 individuals that comprise our study cohort, we identified a total of 20,432 rare homozygous deletions and 338,084 homozygous SNVs/Indels that are introduced by ectopic gene conversion events.

We filtered the complete output of Chameleolyser and clinically interpreted the rare variants that reside in disease-relevant genes. Among the previously undiagnosed patients in our study cohort (n=17,506) we established 25 new genetic diagnoses by either SNVs/Indels (some resulting from ectopic gene conversions) or deletions, or a combination thereof. All of the underlying pathogenic variants (n=29) were identified in one of three genes: *STRC*, *OTOA* and *SMN1*. These are genes for which orthogonal validation assays were already available in our laboratory.

At least as important as the identification of previously hidden genetic variants is the uncovering of ectopic gene conversions which tend to be falsely called as deletions by standard ES and GS analysis pipelines. Sometimes, the clinical consequence of the variant that is introduced by means of a gene conversion is

the same as a deletion, but this is most often not true. Within the *STRC* gene alone, we identified 47 ectopic gene conversions that do not introduce pLoF variation. Since these events are present in patients with all kinds of different phenotypes as well as in healthy parents of patients we can reasonably assume that these events are benign. This can only be true in case the alleles are indeed converted and not deleted. We note that using the standard CNV caller ExomeDepth, all of these events are called as homozygous deletions. This puts patients at risk for an erroneous molecular diagnosis. Even more problematic is the fact that probe-based validation techniques such as MLPA most often confirm these false events. The reason for this is that the typically used nucleic acid probes are too short to reliably derive the origin of the hybridizing DNA. For that reason, special care should be taken when interpreting potential deletions in the paralogous regions of the genome.

Long-read genome sequencing

Very promising for the analysis of these difficult genomic regions are long-read sequencing technologies such as single-molecule, real-time (SMRT) sequencing (Pacific Biosciences of California, Inc) and Nanopore sequencing (Oxford Nanopore technologies) which we used in **chapter 2** to validate the output of Chameleolyser in a subset of our study cohort (demonstrating > 88% validation success) and in **chapter 4** to genetically diagnose previously undiagnosed but well-studied RD families (X. Chen et al., 2023; Figure1). While short reads are typically 100 or 150 base pairs in length, with long-read sequencing technologies tens of thousands of bases can be sequenced consecutively. Consequently, reads that originate from genomic regions with a high sequence similarity to other regions in the genome can most often be uniquely and correctly aligned onto the reference genome because the anchoring sequence is long enough to include multiple sequence differences between the paralogous sequences.

Apart from offering the possibility to analyse the sequence of paralogous genomic regions, long-read sequencing technologies and analysis algorithms also enable a much more accurate identification of SVs as compared to short-read technologies while reaching similar sensitivity and specificity percentages for SNVs and small indels in the unique parts of the genome (Kucuk et al., 2023). The gain in accuracy that is observed for all types of variants in long-read sequencing technologies is primarily caused by the length of the sequencing reads, but also by the fact that these techniques are PCR-free and thus operate on native DNA molecules.

We established 62 genetic diagnoses based on the analysis of SVs in short-read exome data of 6,004 previously undiagnosed RD families (**chapter 3**). Nonetheless,

we know, based on comparisons with long-read sequencing data, that the vast majority of SVs remained unidentified. Furthermore, using short-read sequencing data it is hard or impossible to derive the exact breaking points, information which might be essential for clinically interpreting the variant. Also, to obtain a list of SVs with as few false positives as possible, most often consensus variant calling is performed with multiple variant callers per type of SV. In **chapter 3** for example we used three different variant callers to identify CNVs. For MEIs, we used two different callers after conducting a benchmark study with six callers and for STR expansions we used a single caller after a study in which the performance of three different callers was evaluated. The application of all of these different analysis algorithms requires specific expertise and the necessary computational resources, which is not available in all genetics centers. Because of the fact that the identification of SVs from short-read sequencing data will always be associated with larger numbers of false positive events (as compared to long-read sequencing technologies), expensive orthogonal validations will be needed. Of special difficulty is the identification of relatively small (the order of magnitude of a small exon) single copy number changes. In short-read sequencing data, as opposed to long-read sequencing, the identification of SVs is indirect based on a signature, most often the read depth. Because for small events that affect a single allele, this coverage profile is only distorted to a limited extent, specifically when considering the read depth variability which is intrinsically present in exome data and to a lesser extent also in genome data. Incidentally, this property is also the reason why I could not accurately identify heterozygous duplications, deletions and gene conversions in **chapter 2**.

We can expect that long-read sequencing technologies, which have been improved already in the past decade in terms of increased data quality and throughput and decreased sequencing cost, will eventually become the first-tier sequencing technology in routine genome sequencing. Nonetheless, the cost for sequencing a genome with long-read technologies is currently a multiple of the cost for generating a short-read genome and thus these technologies are yet not affordable for health-care systems. Pacific Biosciences announced that with their latest long-read sequencing device, the Revio, genomes can be sequenced with a median read depth of 30 for roughly 1000 Euros. Further technological advances will eventually decrease the cost even further and make the cost for generating these datasets comparable with the cost for generating short-read data.

Whole-genome *de novo* assembly

The length of the sequencing reads is not the only limiting factor for variant identification. The procedure of aligning reads to a reference genome also hinders

the identification of some genetic variants. In RD research and diagnostics but also in other genomics studies, sequencing reads derived from human genomes are all aligned to the same reference genome. Although the quality of the human reference genome has improved over the years, it is questionable whether a single representation of a human reference genome results in the highest quality variant list for each and everyone's sequenced genome (Ebler et al., 2022; Jarvis et al., 2022; Wang et al., 2022). Aligning reads to a reference genome as an intermediate step in the process of identifying genetic variants implies the assumption that all human genomes are sufficiently similar to this reference genome to allow for all reads aligning onto this reference. We know that this assumption is oftentimes violated. For example, sequencing reads that overlap a genomic region that harbours a complex genetic variant might be too different from the reference genome to be aligned onto it. To identify such variants, alternative variant discovery approaches exist. By these procedures, reads are *de novo* assembled to long sequence contigs which afterwards can be aligned onto a reference genome. Although the accuracy and efficiency of these algorithms are still insufficient for their wide application in routine diagnostics, they are extensively used by researchers for constructing the first reference genome of a certain species (Xu et al., 2018; Zimin et al., 2009, 2017) Also for improving the human reference sequence, long-read *de novo* assembly methods were used (Miga et al., 2020; Nurk et al., 2022; Rhie et al., 2023). A general idea in the field of long-read sequencing is to construct a collection of high quality human haplotype references onto which assembled long-reads can be aligned. As a result, the full spectrum of human genetic variation could be uncovered. Even though we do not use *de novo* assemblies for patient genomes yet, several researchers demonstrated that the use of the human T2T reference genome, which is an improved genome assembly as compared to GRCh38, helps for a more complete characterisation of genetic variants (Aganezov et al., 2024; Noyes et al., 2022; Olson et al., 2023). I expect that in the coming years with further decreasing long-read sequencing costs and with increasing quality of output and assembly algorithms, the *de novo* assembly of sequencing reads will be key in any routine genome sequencing analysis pipeline.

The epigenome

The information regarding a cell's functioning is not only stored in the order of nucleotides of which our genome is composed. Reversible chemical modifications to the DNA or to the proteins that are associated with DNA also play key roles in the regulation of our genes. It has been shown that these so-called epigenetic modifications which might or might not survive cell division are implicated in several human diseases (Bohacek & Mansuy, 2013; Levy et al., 2022; Portela &

Esteller, 2010) . Mendelian disorders of the epigenetic machinery are a group of disorders which are caused by genetic variants in genes encoding for proteins that regulate the epigenetic machinery (Fahrner & Bjornsson, 2014). In principle, to identify the genetic variants that are causative for these diseases, only the DNA of the respective patient needs to be sequenced and interpreted. Other diseases are caused or influenced by epigenetic changes which result from environmental changes. To decipher these diseases it is necessary to query the epigenomic layers in the cell which is a non-trivial task because each type or class of epigenetic change requires another experiment and analysis. The most widely studied epigenetic mark is 5-methylcystosine. To detect this modification in a high-throughput manner, typically, the DNA is treated with sodium bisulfite causing the unmethylated cytosine to convert to uracil which then is converted to thymine during PCR amplification (Frommer et al., 1992; Grunau et al., 2001). Although this protocol is still considered the gold standard for the determination of 5-methylcystosine, it also has quite some disadvantages such as the need for different experiments in case both the sequence of the DNA and its methylation status needs to be known. Also, bisulfite treatment of DNA introduces biases. A particular advantage of long-read sequencing techniques is the fact that the sequence of the DNA and its epigenetic marks can be determined in the same experiment. Because native unamplified molecules are sequenced with these methods, the epigenetic marks are still present on the sequenced DNA allowing for the direct identification by polymerase kinetics (SMRT sequencing, PacBio) or current changes (Nanopore sequencing, ONT). Not only do these technologies and methods enable to detect of 5-methylcystosine, but also a whole series of other epigenetic marks including 4-methylcytosine, 5-hydroxymethylcytosine, N6-methyladenine and 8-oxoguanine, probably making long-read sequencing technologies the method of choice in the near future (Logsdon et al., 2020).

Genetic mosaicism

Even if complete chromosomes can be sequenced accurately from telomere to telomere in the foreseeable future, the complete set of all genetic variants that an individual carries will still not be completely known. Most of human DNA that is sequenced nowadays is derived from white blood cells and because our DNA is not necessarily identical in all cells of the human body (as a result of mosaic and post-zygotic mutations), it might well be that a disease-causative genetic variant is not identifiable in sequence data derived from white blood cells, but only in a tissue which is difficult or impossible to sample. The last couple of years it has been shown extensively by multiple researchers that there is a relatively large genomic variability among different populations of cells in the brain and other tissues and

that these mosaic differences might lead to disease (Huisman et al., 2013; Kurek et al., 2012; McConnell et al., 2013). To identify or validate genetic variants which are only present in part of the sampled cells, deep sequencing followed by mosaic variant calling is typically applied. Because whole-genome deep sequencing is still costly, a targeted sequence capture of the regions of interest can be conducted followed by deep sequencing.

Future perspectives

Clearly, we still cannot determine all genetic variants in an individual which could possibly be of clinical interest, even not in a research setting. Looking back however at the progress that has been made in the past decades, we must recognize that huge steps have been taken to determine the perfect sequence of a complete human genome. In the next years and decades, further advances will eventually close the last gaps and probably introduce highly accurate long-read sequencing in clinical practice.

The phenotypic consequences of genetic variation

In addition to the difficult exercise of perfectly deriving the sequence of a patient's genome, there is an even bigger bottleneck which limits us from achieving higher diagnostic success rates and that is the clinical interpretation of genetic variants. The complex biological reality that lies between the sequence of the DNA and an individual's phenotype makes it impossible for the vast majority of variants identified in exome or genome experiments to immediately conclude whether and to what extent they contribute to the phenotype. Although *in silico* predictions based on current knowledge might be relatively accurate with respect to the effect of a variant on protein level, they fail in telling us with certainty if a particular genetic variant causes disease or not in a given patient. Even for the latest algorithms based on artificial intelligence it is impossible to capture the complete picture of an organism's or a cell's mechanics. For that reason, most often, years of functional studies are needed to prove the causal link between variants in a gene and a disease phenotype.

Exome and genome reanalysis

Every year the number of human genes that are linked to disease substantially increases. This observation suggests that the re-evaluation of existing genetic data using the latest biological and genetic information could potentially be of high value. **Chapter 3** describes such an analysis. In that work, which is part

of the Solve-RD project, we reassessed 9,874 exomes and genomes from 6,004 previously undiagnosed RD families. All of these datasets were carefully analysed by the contributing laboratory prior to 2018. From the 506 diagnoses that resulted from the systematic reanalysis of these data, a total of 63 (12.5%) were based on pathogenic variants in genes which were found to be a disease gene after 2018. The use of recent information at the variant level also resulted in a substantial number of diagnoses. In 165 RD families we identified disease-explanatory variants which were all VUSs prior to 2018. Their reclassification between 2018 and 2022 together with a careful judgement by the disease experts revealed that these variants are indeed the cause of the patient's disease. In addition to genetic diagnoses which result from knowledge that is acquired by the field in recent years, the clinical interpretation of VUSs by top disease experts across Europe resulted in a genetic diagnosis in 191 RD families. Although a small subset of these diagnoses might be attributed to the use of the latest and thus higher quality variant callers, most of these disease-causing genetic variants are easy-to-identify but more difficult to clinically interpret. The systematic Solve-RD reanalysis also resulted in 87 diagnoses due to difficult-to-identify genetic variants, but the vast majority of diagnoses that we established (n=419) are the result of the gain in biologic and genetic knowledge between 2018 and 2022 or from the disease expertise that was brought together within the project. Our Pan-European study is not the only large-scale reanalysis effort that has been conducted in recent years. Other consortia and research groups performed similar studies resulting in comparable diagnostic yields, further illustrating the importance of analysing RD sequencing data at multiple points in time using the latest RD and bioinformatics knowledge (Bullich et al., 2022; Liu et al., 2019; Wright, McRae, et al., 2018).

Even though these research activities substantially contribute to the field and potentially improve a large number of patient lives, we need to concede that most RD families still remain genetically undiagnosed. Undoubtedly, for a large number of these yet undiagnosed RD families, functional studies will be needed to unambiguously proof the causal relationship between a genetic variant and the phenotype. Because these studies are time and resource intensive, they are typically only conducted in case there already is a certain degree of certainty about the disease-explanatory nature of a specific genetic variant. The judgment that a certain VUS is more likely causative for disease as compared to another VUS is, apart from co-segregation and statistical information, largely formed by various *in silico* parameters such as the population allele frequency of the variant, the conservation of the altered nucleotide in other species, *in silico* predictors of pathogenicity and the function of the respective gene. For the paralogous regions in the genome, all of

these values are either unavailable or inaccurate causing the (clinical) interpretation of the variants within these regions more difficult as compared to the rest of the genome. Despite this, linking genetic variants to phenotypes is also challenging in the unique regions of the genome as a result of the complex path between variant and phenotype. In case of ultra-rare diseases, this challenge often becomes even bigger because to prioritise candidate disease-causing variants and to proof their involvement in the phenotype, multiple families with candidate disease-causing in the same gene are most often desirable and needed. A similar issue with numerical power presents in families affected by di- or oligogenic diseases.

Perspectives on the future

The improved bioinformatics approaches and advanced sequencing technologies presented in this work have the potential to increase the detection rate of disease-causing variants, leading to more definitive diagnoses in a greater number of cases. This is crucial for both patients and their families, as it provides clarity and direction for managing health concerns. For clinicians, these advancements mean they can have greater confidence in the diagnostic outcomes of genetic tests and may be able to offer treatment options more quickly and with greater certainty.

Furthermore, the results of this work may prompt clinicians to place greater emphasis on the iterative nature of genetic diagnostics. Instead of viewing genetic testing as a one-time event, patients may need to undergo multiple tests as new variants and their clinical significance are discovered. This requires a shift in how genetic testing is discussed, highlighting the possibility of follow-up testing and the ongoing re-evaluation of genetic data.

Looking ahead, I anticipate that with the continuous evolution in these fields, it will become possible to query all types of genetic and epigenetic changes in a genome with a single test, at a cost that is affordable for most healthcare systems. Undoubtedly, the pace at which genomes are sequenced will accelerate in the coming years and decades, eventually leading to a deeper insight into the human genome. I expect that this enhanced understanding will result in the development of novel, personalized treatment options, ultimately improving patient care and outcomes.

Chapter 6

References

Aartsma-Rus, A., Ginjaar, I. B., & Bushby, K. (2016). The importance of genetic diagnosis for Duchenne muscular dystrophy. Journal of Medical Genetics, 53(3), 145. https://doi.org/10.1136/jmedgenet-2015-103387

Aganezov, S., Yan, S. M., Soto, D. C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D. J., Shafin, K., Shumate, A., Xiao, C., Wagner, J., McDaniel, J., Olson, N. D., Sauria, M. E. G., Vollger, M. R., Rhie, A., Meredith, M., Martin, S., Lee, J., … Schatz, M. C. (2024). A complete reference genome improves analysis of human genetic variation. Science, 376(6588), eabl3533. https://doi.org/10.1126/science.abl3533

Ajani, J. A., D'Amico, T. A., Bentrem, D. J., Chao, J., Cooke, D., Corvera, C., Das, P., Enzinger, P. C., Enzler, T., Fanta, P., Farjah, F., Gerdes, H., Gibson, M. K., Hochwald, S., Hofstetter, W. L., Ilson, D. H., Keswani, R. N., Kim, S., Kleinberg, L. R., … Pluchino, L. A. (2022). Gastric Cancer, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology. Journal of the National Comprehensive Cancer Network : JNCCN, 20(2), 167–192. https://doi.org/10.6004/JNCCN.2022.0008

Alfares, A., Aloraini, T., subaie, L. Al, Alissa, A., Qudsi, A. Al, Alahmad, A., Mutairi, F. Al, Alswaid, A., Alothaim, A., Eyaid, W., Albalwi, M., Alturki, S., & Alfadhel, M. (2018). Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. Genetics in Medicine, 20(11), 1328–1333. https://doi.org/https://doi.org/10.1038/gim.2018.41

Atalaia, A., Thompson, R., Corvo, A., Carmody, L., Piscia, D., Matalonga, L., Macaya, A., Lochmuller, A., Fontaine, B., Zurek, B., Hernandez-Ferrer, C., Rheinard, C., Gómez-Andrés, D., Desaphy, J. F., Schon, K., Lohmann, K., Jennings, M. J., Synofzik, M., Riess, O., … Bonne, G. (2020). A guide to writing systematic reviews of rare disease treatments to generate FAIR-compliant datasets: building a Treatabolome. Orphanet Journal of Rare Diseases, 15(1). https://doi.org/10.1186/S13023-020-01493-7

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Schloss, J. A. (2015). A global reference for human genetic variation. In Nature (Vol. 526, Issue 7571, pp. 68–74). https://doi.org/10.1038/nature15393

Baetens, D., Stoop, H., Peelman, F., Todeschini, A.-L., Rosseel, T., Coppieters, F., Veitia, R. A., Looijenga, L. H. J., De Baere, E., & Cools, M. (2017). NR5A1 is a novel disease gene for 46,XX testicular and ovotesticular disorders of sex development. Genetics in Medicine, 19(4), 367–376. https://doi.org/10.1038/gim.2016.118

Bahi-Buisson, N., Poirier, K., Fourniol, F., Saillour, Y., Valence, S., Lebrun, N., Hully, M., Fallet Bianco, C., Boddaert, N., Elie, C., Lascelles, K., Souville, I., Consortium, L.-T., Beldjord, C., & Chelly, J. (2014). The wide spectrum of tubulinopathies: what are the key features for the diagnosis? Brain, 137(6), 1676–1700. https://doi.org/10.1093/brain/awu082

Baker, S. W., Murrell, J. R., Nesbitt, A. I., Pechter, K. B., Balciuniene, J., Zhao, X., Yu, Z., Denenberg, E. H., DeChene, E. T., Wilkens, A. B., Bhoj, E. J., Guan, Q., Dulik, M. C., Conlin, L. K., Abou Tayoun, A. N., Luo, M., Wu, C., Cao, K., Sarmady, M., … Santani, A. B. (2019a). Automated Clinical Exome Reanalysis Reveals Novel Diagnoses. The Journal of Molecular Diagnostics, 21(1), 38–48. https://doi.org/10.1016/J.JMOLDX.2018.07.008

Bauché, S., O'Regan, S., Azuma, Y., Laffargue, F., McMacken, G., Sternberg, D., Brochier, G., Buon, C., Bouzidi, N., Topf, A., Lacène, E., Remerand, G., Beaufrere, A. M., Pebrel-Richard, C., Thevenon, J., El Chehadeh-Djebbar, S., Faivre, L., Duffourd, Y., Ricci, F., … Nicole, S. (2016). Impaired Presynaptic High-Affinity Choline Transporter Causes a Congenital Myasthenic Syndrome with Episodic Apnea. American Journal of Human Genetics, 99(3), 753–761. https://doi.org/10.1016/j.ajhg.2016.06.033

Bazykin, G. A., & Kochetov, A. V. (2011). Alternative translation start sites are conserved in eukaryotic genomes. Nucleic Acids Research, 39(2), 567–577. https://doi.org/10.1093/nar/gkq806

Bean, L. J. H., Funke, B., Carlston, C. M., Gannon, J. L., Kantarci, S., Krock, B. L., Zhang, S., Bayrak-Toydemir, P., & AssuranceCommittee, on behalf of the A. L. Q. (2020). Diagnostic gene sequencing panels: from design to report—atechnical standard of the American College of Medical Genetics and Genomics(ACMG). Genetics in Medicine, 22(3), 453–461. https://doi.org/10.1038/s41436-019-0666-z

Beck, T. F., Mullikin, J. C., & Biesecker, L. G. (2016). Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. Clinical Chemistry, 62(4), 647–654. https://doi.org/10.1373/clinchem.2015.249623

Beetz, C., Schüle, R., Deconinck, T., Tran-Viet, K.-N., Zhu, H., Kremer, B. P. H., Frints, S. G. M., van Zelst-Stams, W. A. G., Byrne, P., Otto, S., Nygren, A. O. H., Baets, J., Smets, K., Ceulemans, B., Dan, B., Nagan, N., Kassubek, J., Klimpe, S., Klopstock, T., … Züchner, S. (2008). REEP1 mutation spectrum and genotype/phenotype correlation in hereditary spastic paraplegia type 31. Brain, 131(4), 1078–1086. https://doi.org/10.1093/brain/awn026

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.-L., & Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proceedings of the National Academy of Sciences of the United States of America, 112(17), 5473–5478. https://doi.org/10.1073/pnas.1418631112

Benarroch, L., Bonne, G., Rivier, F., & Hamroun, D. (2023). The 2023 version of the gene table of neuromuscular disorders (nuclear genome). Neuromuscular Disorders : NMD, 33(1), 76–117. https://doi.org/10.1016/J.NMD.2022.12.002

Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., Atlason, B. A., Kristmundsdottir, S., Mehringer, S., Hardarson, M. T., Gudjonsson, S. A., Magnusdottir, D. N., Jonasdottir, A., Jonasdottir, A., Kristjansson, R. P., Sverrisson, S. T., Holley, G., Palsson, G., Stefansson, O. A., … Stefansson, K. (2021a). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nature Genetics, 53(6), 779–786. https://doi.org/10.1038/S41588-021-00865-4

Biesecker, L. G., & Green, R. C. (2014a). Diagnostic Clinical Genome and Exome Sequencing. New England Journal of Medicine, 370(25), 2418–2425. https://doi.org/10.1056/NEJMra1312543

Biran, A., Myers, N., Steinberger, S., Adler, J., Riutin, M., Broennimann, K., Reuven, N., & Shaul, Y. (2022). The C-Terminus of the PSMA3 Proteasome Subunit Preferentially Traps Intrinsically Disordered Proteins for Degradation. Cells, 11(20). https://doi.org/10.3390/cells11203231

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O., & Barillot, E. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics (Oxford, England), 28(3), 423–425. https://doi.org/10.1093/bioinformatics/btr670

Bohacek, J., & Mansuy, I. M. (2013). Epigenetic Inheritance of Disease and Disease Risk. Neuropsychopharmacology, 38(1), 220–236. https://doi.org/10.1038/npp.2012.110

Bonne, G. (2021). The Treatabolome, an emerging concept. Journal of Neuromuscular Diseases, 8(3), 337–339. https://doi.org/10.3233/JND-219003

Borràs, D. M., Vossen, R. H. A. M., Liem, M., Buermans, H. P. J., Dauwerse, H., van Heusden, D., Gansevoort, R. T., den Dunnen, J. T., Janssen, B., Peters, D. J. M., Losekoot, M., & Anvar, S. Y. (2017). Detecting PKD1 variants in polycystic kidney disease patients by single-molecule long-read sequencing. Human Mutation, 38(7), 870–879. https://doi.org/https://doi.org/10.1002/humu.23223

Boycott, K. M., Azzariti, D. R., Hamosh, A., & Rehm, H. L. (2022). Seven years since the launch of the Matchmaker Exchange: The evolution of genomic matchmaking. Human Mutation, 43(6), 659–667. https://doi.org/10.1002/HUMU.24373

Boycott, K. M., Dyment, D. A., & Innes, A. M. (2018). Unsolved recognizable patterns of human malformation: Challenges and opportunities. American Journal of Medical Genetics Part C: Seminars in Medical Genetics, 178(4), 382–386. https://doi.org/https://doi.org/10.1002/ajmg.c.31665

Boycott, K. M., Rath, A., Chong, J. X., Hartley, T., Alkuraya, F. S., Baynam, G., Brookes, A. J., Brudno, M., Carracedo, A., den Dunnen, J. T., Dyke, S. O. M., Estivill, X., Goldblatt, J., Gonthier, C., Groft, S. C., Gut, I., Hamosh, A., Hieter, P., Höhn, S., … Lochmüller, H. (2017). International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. The American Journal of Human Genetics, 100(5), 695–705. https://doi.org/10.1016/j.ajhg.2017.04.003

Brahe, C., Servidei, S., Zappata, S., Ricci, E., Tonali, P., & Neri, G. (1995). Genetic homogeneity between childhood-onset and adult-onset autosomal recessive spinal muscular atrophy. Lancet (London, England), 346(8977), 741–742. https://doi.org/10.1016/s0140-6736(95)91507-9

Brandes, N., Goldman, G., Wang, C. H., Ye, C. J., & Ntranos, V. (2023). Genome-wide prediction of disease variant effects with a deep protein language model. Nature Genetics. https://doi.org/10.1038/s41588-023-01465-0

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Ling, X. S., Mastrangelo, C. H., Meller, A., Oliver, J. S., Pershin, Y. V, Ramsey, J. M., … Schloss, J. A. (2008). The potential and challenges of nanopore sequencing. Nature Biotechnology, 26(10), 1146–1153. https://doi.org/10.1038/nbt.1495

Bullich, G., Matalonga, L., Pujadas, M., Papakonstantinou, A., Piscia, D., Tonda, R., Artuch, R., Gallano, P., Garrabou, G., González, J. R., Grinberg, D., Guitart, M., Laurie, S., Lázaro, C., Luengo, C., Martí, R., Milà, M., Ovelleiro, D., Parra, G., … Vendrell, T. (2022a). Systematic Collaborative Reanalysis of Genomic Data Improves Diagnostic Yield in Neurologic Rare Diseases. The Journal of Molecular Diagnostics : JMD, 24(5), 529–542. https://doi.org/10.1016/J.JMOLDX.2022.02.003

Calabrese, C., Simone, D., Diroma, M. A., Santorsola, M., Guttà, C., Gasparre, G., Picardi, E., Pesole, G., & Attimonelli, M. (2014). MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. Bioinformatics (Oxford, England), 30(21), 3115–3117. https://doi.org/10.1093/bioinformatics/btu483

Campbell, L., Potter, A., Ignatius, J., Dubowitz, V., & Davies, K. (1997). Genomic Variation and Gene Conversion in Spinal Muscular Atrophy: Implications for Disease Process and Clinical Phenotype. The American Journal of Human Genetics, 61(1), 40–50. https://doi.org/https://doi.org/10.1086/513886

Carmichael, N., Tsipis, J., Windmueller, G., Mandel, L., & Estrella, E. (2015). "Is it Going to Hurt?": The Impact of the Diagnostic Odyssey on Children and Their Families. Journal of Genetic Counseling, 24(2), 325–335. https://doi.org/10.1007/s10897-014-9773-9

Casola, C., Zekonyte, U., Phillips, A. D., Cooper, D. N., & Hahn, M. W. (2012). Interlocus gene conversion events introduce deleterious mutations into at least 1% of human genes associated with inherited disease. Genome Research, 22(3), 429–435. https://doi.org/10.1101/gr.127738.111

Casse, F., Courtin, T., Tesson, C., Ferrien, M., Noël, S., Fauret-Amsellem, A.-L., Gareau, T., Guegan, J., Anheim, M., Mariani, L.-L., Le Forestier, N., Tranchant, C., Corvol, J.-C., Lesage, S., Brice, A., & (PDG), for the F. P. disease genetics study group. (2023). Detection of ATXN2 Expansions in an Exome Dataset: An Underdiagnosed Cause of Parkinsonism. Movement Disorders Clinical Practice, 10(4), 664–669. https://doi.org/https://doi.org/10.1002/mdc3.13699

Castagna, A. E., Addis, J., McInnes, R. R., Clarke, J. T. R., Ashby, P., Blaser, S., & Robinson, B. H. (2007). Late onset Leigh syndrome and ataxia due to a T to C mutation at bp 9,185 of mitochondrial DNA. American Journal of Medical Genetics. Part A, 143A(8), 808–816. https://doi.org/10.1002/AJMG.A.31637

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., … Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nature Communications, 10(1), 1784. https://doi.org/10.1038/s41467-018-08148-z

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., & Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. Nature Reviews Genetics, 8(10), 762–775. https://doi.org/10.1038/nrg2193

Chen, X., Harting, J., Farrow, E., Thiffault, I., Kasperaviciute, D., Hoischen, A., Gilissen, C., Pastinen, T., & Eberle, M. A. (2023). Comprehensive SMN1 and SMN2 profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. The American Journal of Human Genetics, 110(2), 240–250. https://doi.org/10.1016/j.ajhg.2023.01.001

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics (Oxford, England), 32(8), 1220–1222. https://doi.org/10.1093/BIOINFORMATICS/BTV710

Chen, Y., Dawes, R., Kim, H. C., Stenton, S. L., Walker, S., Ljungdahl, A., Lord, J., Ganesh, V. S., Ma, J., Martin-Geary, A. C., Lemire, G., D'Souza, E. N., Dong, S., Ellingford, J. M., Adams, D. R., Allan, K., Bakshi, M., Baldwin, E. E., Berger, S. I., … Whiffin, N. (2024). De novo variants in the non-coding spliceosomal snRNA gene RNU4-2 are a frequent cause of syndromic neurodevelopmental disorders. MedRxiv, 17, 2024.04.07.24305438. https://doi.org/10.1101/2024.04.07.24305438

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science (New York, N.Y.), 381(6664), eadg7492. https://doi.org/10.1126/science.adg7492

Childs, A. M., Hutchin, T., Pysden, K., Highet, L., Bamford, J., Livingston, J., & Crow, Y. J. (2007). Variable phenotype including Leigh syndrome with a 9185T>C mutation in the MTATP6 gene. Neuropediatrics, 38(6), 313–316. https://doi.org/10.1055/S-2008-1065355

Chintalaphani, S. R., Pineda, S. S., Deveson, I. W., & Kumar, K. R. (2021). An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. Acta Neuropathologica Communications, 9(1), 98. https://doi.org/10.1186/s40478-021-01201-x

Chong, J. X., Yu, J.-H., Lorentzen, P., Park, K. M., Jamal, S. M., Tabor, H. K., Rauch, A., Saenz, M. S., Boltshauser, E., Patterson, K. E., Nickerson, D. A., Bamshad, M. J., & Genomics, ; University of Washington Center for Mendelian. (2016). Gene discovery for Mendelian conditions via social networking: de novo variants in KDM1A cause developmental delay and distinctive facial features. Genetics in Medicine, 18(8), 788–795. https://doi.org/10.1038/gim.2015.161

Coe, B. P., Witherspoon, K., Rosenfeld, J. A., van Bon, B. W. M., Vulto-van Silfhout, A. T., Bosco, P., Friend, K. L., Baker, C., Buono, S., Vissers, L. E. L. M., Schuurs-Hoeijmakers, J. H., Hoischen, A., Pfundt, R., Krumm, N., Carvill, G. L., Li, D., Amaral, D., Brown, N., Lockhart, P. J., … Eichler, E. E. (2014). Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nature Genetics, 46(10), 1063–1071. https://doi.org/10.1038/ng.3092

Cohen, A. S. A., Farrow, E. G., Abdelmoity, A. T., Alaimo, J. T., Amudhavalli, S. M., Anderson, J. T., Bansal, L., Bartik, L., Baybayan, P., Belden, B., Berrios, C. D., Biswell, R. L., Buczkowicz, P., Buske, O., Chakraborty, S., Cheung, W. A., Coffman, K. A., Cooper, A. M., Cross, L. A., … Pastinen, T. (2022). Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes. Genetics in Medicine : Official Journal of the American College of Medical Genetics, 24(6), 1336–1348. https://doi.org/10.1016/J.GIM.2022.02.007

Conrad, D. F., Keebler, J. E. M., DePristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V, Zilversmit, M., Cartwright, R., Rouleau, G. A., Daly, M., Stone, E. A., Hurles, M. E., Awadalla, P., & Project, the 1000 G. (2011). Variation in genome-wide mutation rates within and between human families. Nature Genetics, 43(7), 712–714. https://doi.org/10.1038/ng.862

Cummings, B. B., Marshall, J. L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A. R., Bolduc, V., Waddell, L. B., Sandaradura, S. A., O'Grady, G. L., Estrella, E., Reddy, H. M., Zhao, F., Weisburd, B., Karczewski, K. J., O'Donnell-Luria, A. H., Birnbaum, D., Sarkozy, A., Hu, Y., … Zaugg, J. B. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Science Translational Medicine, 9(386). https://doi.org/10.1126/SCITRANSLMED.AAL5209

Dai, P., Honda, A., Ewans, L., McGaughran, J., Burnett, L., Law, M., & Phan, T. G. (2022). Recommendations for next generation sequencing data reanalysis of unsolved cases with suspected Mendelian disorders: A systematic review and meta-analysis. Genetics in Medicine, 24(8), 1618–1629. https://doi.org/10.1016/J.GIM.2022.04.021

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience, 10(2), giab008. https://doi.org/10.1093/gigascience/giab008

Danis, D., Jacobsen, J. O. B., Carmody, L. C., Gargano, M. A., McMurry, J. A., Hegde, A., Haendel, M. A., Valentini, G., Smedley, D., & Robinson, P. N. (2021). Interpretable prioritization of splice variants in diagnostic next-generation sequencing. The American Journal of Human Genetics, 108(9), 1564–1577. https://doi.org/10.1016/J.AJHG.2021.06.014

de Boer, E., Ockeloen, C. W., Matalonga, L., Horvath, R., Rodenburg, R. J., Coenen, M. J. H., Janssen, M., Henssen, D., Gilissen, C., Steyaert, W., Paramonov, I., Trimouille, A., Kleefstra, T., Verloes, A., & Vissers, L. E. L. M. (2021). A MT-TL1 variant identified by whole exome sequencing in an individual with intellectual disability, epilepsy, and spastic tetraparesis. European Journal of Human Genetics : EJHG, 29(9). https://doi.org/10.1038/S41431-021-00900-2

de Ligt, J., Willemsen, M. H., van Bon, B. W. M., Kleefstra, T., Yntema, H. G., Kroes, T., Vulto-van Silfhout, A. T., Koolen, D. A., de Vries, P., Gilissen, C., del Rosario, M., Hoischen, A., Scheffer, H., de Vries, B. B. A., Brunner, H. G., Veltman, J. A., & Vissers, L. E. L. M. (2012b). Diagnostic exome sequencing in persons with severe intellectual disability. The New England Journal of Medicine, 367(20), 1921–1929. https://doi.org/10.1056/NEJMOA1206524

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. Nature Biotechnology, 34(5), 518–524. https://doi.org/10.1038/nbt.3423

Demidov, G., Laurie, S., Torella, A., Piluso, G., Scala, M., Morleo, M., Nigro, V., Graessner, H., Banka, S., Macaya, A., Pérez-Dueñas, B., Jackson, A., Stevanin, G., de Sainte Agathe, J.-M., Havlovicová, M., Horvath, R., Pinelli, M., van Os, N. J. H., van de Warrenburg, B. P. C., … Ossowski, S. (2024). Structural variant calling and clinical interpretation in 6224 unsolved rare disease exomes. European Journal of Human Genetics : EJHG. https://doi.org/10.1038/S41431-024-01637-4

Demidov, G., Park, J., Armeanu-Ebinger, S., Roggia, C., Faust, U., Cordts, I., Blandfort, M., Haack, T. B., Schroeder, C., & Ossowski, S. (2021). Detection of mobile elements insertions for routine clinical diagnostics in targeted sequencing data. Molecular Genetics & Genomic Medicine, 9(12), e1807. https://doi.org/10.1002/mgg3.1807

Demidov, G., Sturm, M., & Ossowski, S. (2022). ClinCNV: multi-sample germline CNV detection in NGS data. BioRxiv, 2022.06.10.495642. https://doi.org/10.1101/2022.06.10.495642

Demidov, G., Yaldiz, B., Garcia-Pelaez, J., Boer, E. de, Schuermans, N., Vondel, L. Van de, Paramonov, I., Johansson, L. F., Musacchia, F., Benetti, E., Bullich, G., Sablauskas, K., Beltran, S., Gilissen, C., Hoischen, A., Ossowski, S., Voer, R. de, Lohmann, K., Oliveira, C., … Laurie, S. (2023). Comprehensive reanalysis for CNVs in ES data from unsolved rare disease cases results in new diagnoses. MedRxiv, 2023.10.22.23296993. https://doi.org/10.1101/2023.10.22.23296993

Denommé-Pichon, A. S., Bruel, A. L., Duffourd, Y., Safraou, H., Thauvin-Robinet, C., Tran Mau-Them, F., Philippe, C., Vitobello, A., Denommé-Pichon, A. S., Bruel, A. L., Duffourd, Y., Jean-Marçais, N., Moutton, S., Safraou, H., Thauvin-Robinet, C., Thevenon, J., Tran Mau-Them, F., Philippe, C., Vitobello, A., … Alembik, Y. (2023). A Solve-RD ClinVar-based reanalysis of 1522 index cases from ERN-ITHACA reveals common pitfalls and misinterpretations in exome sequencing. Genetics in Medicine : Official Journal of the American College of Medical Genetics, 25(4). https://doi.org/10.1016/J.GIM.2023.100018

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43(5), 491–498. https://doi.org/10.1038/ng.806

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., & Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Research, 37(9), e67–e67. https://doi.org/10.1093/nar/gkp215

Dolzhenko, E., Bennett, M. F., Richmond, P. A., Trost, B., Chen, S., van Vugt, J. J. F. A., Nguyen, C., Narzisi, G., Gainullin, V. G., Gross, A. M., Lajoie, B. R., Taft, R. J., Wasserman, W. W., Scherer, S. W., Veldink, J. H., Bentley, D. R., Yuen, R. K. C., Bahlo, M., & Eberle, M. A. (2020). ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. Genome Biology, 21(1), 102. https://doi.org/10.1186/s13059-020-02017-z

Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., Scheffler, K., van Vugt, J. J. F. A., French, C., Sanchis-Juan, A., Ibáñez, K., Tucci, A., Lajoie, B. R., Veldink, J. H., Raymond, F. L., … Eberle, M. A. (2019). ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics, 35(22), 4754–4756. https://doi.org/10.1093/bioinformatics/btz431

Dolzhenko, E., English, A., Dashnow, H., De Sena Brandine, G., Mokveld, T., Rowell, W. J., Karniski, C., Kronenberg, Z., Danzi, M. C., Cheung, W. A., Bi, C., Farrow, E., Wenger, A., Chua, K. P., Martínez-Cerdeño, V., Bartley, T. D., Jin, P., Nelson, D. L., Zuchner, S., … Eberle, M. A. (2024). Characterization and visualization of tandem repeats at genome scale. Nature Biotechnology. https://doi.org/10.1038/s41587-023-02057-3

Dolzhenko, E., van Vugt, J. J. F. A., Shaw, R. J., Bekritsky, M. A., Van Blitterswijk, M., Narzisi, G., Ajay, S. S., Rajan, V., Lajoie, B. R., Johnson, N. H., Kingsbury, Z., Humphray, S. J., Schellevis, R. D., Brands, W. J., Baker, M., Rademakers, R., Kooyman, M., Tazelaar, G. H. P., van Es, M. A., … Eberle, M. A. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Research, 27(11), 1895–1903. https://doi.org/10.1101/GR.225672.117/-/DC1

Dumont, B. L. (2015). Interlocus gene conversion explains at least 2.7 % of single nucleotide variants in human segmental duplications. BMC Genomics, 16(1), 456. https://doi.org/10.1186/s12864-015-1681-3

Ebbert, M. T. W., Jensen, T. D., Jansen-West, K., Sens, J. P., Reddy, J. S., Ridge, P. G., Kauwe, J. S. K., Belzil, V., Pregent, L., Carrasquillo, M. M., Keene, D., Larson, E., Crane, P., Asmann, Y. W., Ertekin-Taner, N., Younkin, S. G., Ross, O. A., Rademakers, R., Petrucelli, L., & Fryer, J. D. (2019). Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biology, 20(1), 97. https://doi.org/10.1186/s13059-019-1707-2

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., … Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science, 372(6537), eabf7117. https://doi.org/10.1126/science.abf7117

Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Mao, Y., Korbel, J. O., Eichler, E. E., Zody, M. C., Dilthey, A. T., & Marschall, T. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nature Genetics, 54(4), 518–525. https://doi.org/10.1038/s41588-022-01043-w

Eilbeck, K., Quinlan, A., & Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. Nature Reviews Genetics, 18(10), 599–612. https://doi.org/10.1038/nrg.2017.52

Fadaie, Z., Neveling, K., Mantere, T., Derks, R., Haer-Wigman, L., den Ouden, A., Kwint, M., O'Gorman, L., Valkenburg, D., Hoyng, C. B., Gilissen, C., Vissers, L. E. L. M., Nelen, M., Cremers, F. P. M., Hoischen, A., & Roosing, S. (2021). Long-read technologies identify a hidden inverted duplication in a family with choroideremia. HGG Advances, 2(4), 100046. https://doi.org/10.1016/j.xhgg.2021.100046

Fahrner, J. A., & Bjornsson, H. T. (2014). Mendelian disorders of the epigenetic machinery: tipping the balance of chromatin states. Annual Review of Genomics and Human Genetics, 15, 269–293. https://doi.org/10.1146/annurev-genom-090613-094245

Ferreira, C. R., van Karnebeek, C. D. M., Vockley, J., & Blau, N. (2019). A proposed nosology of inborn errors of metabolism. Genetics in Medicine, 21(1). https://doi.org/10.1038/s41436-018-0022-8

Fiol, A., Jurado-Ruiz, F., López Girona, E., & Aranzana, M. J. (2022). An efficient CRISPR-Cas9 enrichment sequencing strategy for characterizing complex and highly duplicated genomic regions. A case study in the Prunus salicina LG3-MYB10 genes cluster. Plant Methods, 18(1), 105. https://doi.org/10.1186/s13007-022-00937-4

Fischbach, G. D., & Lord, C. (2010). The simons simplex collection: A resource for identification of autism genetic risk factors. Neuron, 68(2), 192–195. https://doi.org/10.1016/j.neuron.2010.10.006

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., & Marks, D. S. (2021). Disease variant prediction with deep generative models of evolutionary data. Nature, 599(7883), 91–95. https://doi.org/10.1038/s41586-021-04043-8

Freeberg, M. A., Fromont, L. A., D'Altri, T., Romero, A. F., Ciges, J. I., Jene, A., Kerry, G., Moldes, M., Ariosa, R., Bahena, S., Barrowdale, D., Barbero, M. C., Fernandez-Orth, D., Garcia-Linares, C., Garcia-Rios, E., Haziza, F., Juhasz, B., Llobet, O. M., Milla, G., … Rambla, J. (2022). The European Genome-phenome Archive in 2021. Nucleic Acids Research, 50(D1), D980–D987. https://doi.org/10.1093/NAR/GKAB1059

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., Carter, N. P., Scherer, S. W., & Lee, C. (2006). Copy number variation: New insights in genome diversity. Genome Research, 16(8), 949–961. https://doi.org/10.1101/gr.3677206

French, J. D., & Edwards, S. L. (2020). The Role of Noncoding Variants in Heritable Disease. Trends in Genetics, 36(11), 880–891. https://doi.org/https://doi.org/10.1016/j.tig.2020.07.004

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., & Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proceedings of the National Academy of Sciences, 89(5), 1827–1831. https://doi.org/10.1073/pnas.89.5.1827

Fry, A. E., Marra, C., Derrick, A. V, Pickrell, W. O., Higgins, A. T., te Water Naude, J., McClatchey, M. A., Davies, S. J., Metcalfe, K. A., Tan, H. J., Mohanraj, R., Avula, S., Williams, D., Brady, L. I., Mesterman, R., Tarnopolsky, M. A., Zhang, Y., Yang, Y., Wang, X., … Chung, S.-K. (2021). Missense variants in the N-terminal domain of the A isoform of FHF2/FGF13 cause an X-linked developmental and epileptic encephalopathy. The American Journal of Human Genetics, 108(1), 176–185. https://doi.org/10.1016/j.ajhg.2020.10.017

Fu, J. M., Satterstrom, F. K., Peng, M., Brand, H., Collins, R. L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S. P., Stevens, C. R., Cusick, C., Babadi, M., Banks, E., Collins, B., Dodge, S., Gabriel, S. B., Gauthier, L., Lee, S. K., … Talkowski, M. E. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nature Genetics, 54(9). https://doi.org/10.1038/S41588-022-01104-0

Gambin, T., Akdemir, Z. C., Yuan, B., Gu, S., Chiang, T., Carvalho, C. M. B., Shaw, C., Jhangiani, S., Boone, P. M., Eldomery, M. K., Karaca, E., Bayram, Y., Stray-Pedersen, A., Muzny, D., Charng, W.-L., Bahrambeigi, V., Belmont, J. W., Boerwinkle, E., Beaudet, A. L., … Lupski, J. R. (2017). Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. Nucleic Acids Research, 45(4), 1633–1648. https://doi.org/10.1093/nar/gkw1237

Ganetzky, R. D., Stendel, C., McCormick, E. M., Zolkipli-Cunningham, Z., Goldstein, A. C., Klopstock, T., & Falk, M. J. (2019). MT-ATP6 mitochondrial disease variants: Phenotypic and biochemical features analysis in 218 published cases and cohort of 14 new cases. Human Mutation, 40(5), 499–515. https://doi.org/10.1002/HUMU.23723

Gangfuß, A., Lochmüller, H., Töpf, A., O'Heir, E., Horvath, R., Kölbel, H., Schweiger, B., Schara-Schmidt, U., & Roos, A. (2022). A de novo CSDE1 variant causing neurodevelopmental delay, intellectual disability, neurologic and psychiatric symptoms in a child of consanguineous parents. American Journal of Medical Genetics. Part A, 188(1), 283–291. https://doi.org/10.1002/AJMG.A.62494

Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Stephen Pittard, W., Mills, R. E., & Devine, S. E. (2017). The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. Genome Research, 27(11), 1916–1929. https://doi.org/10.1101/GR.218032.116/-/DC1

Gardner, E. J., Sifrim, A., Lindsay, S. J., Prigmore, E., Rajan, D., Danecek, P., Gallone, G., Eberhardt, R. Y., Martin, H. C., Wright, C. F., FitzPatrick, D. R., Firth, H. V, & Hurles, M. E. (2021). Detecting cryptic clinically relevant structural variation in exome-sequencing data increases diagnostic yield for developmental disorders. The American Journal of Human Genetics, 108(11), 2186–2194. https://doi.org/https://doi.org/10.1016/j.ajhg.2021.09.010

Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., & Muller, J. (2018). AnnotSV: an integrated tool for structural variations annotation. Bioinformatics (Oxford, England), 34(20), 3572–3574. https://doi.org/10.1093/BIOINFORMATICS/BTY304

Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., Van De Vorst, M., Van Bon, B. W. M., Willemsen, M. H., Kwint, M., Janssen, I. M., Hoischen, A., Schenck, A., Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H. G., De Vries, B. B. A., Kleefstra, T., Brunner, H. G., … Veltman, J. A. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature, 511(7509), 344–347. https://doi.org/10.1038/NATURE13394

Goizet, C., Depienne, C., Benard, G., Boukhris, A., Mundwiller, E., Solé, G., Coupry, I., Pilliod, J., Martin-Négrier, M.-L., Fedirko, E., Forlani, S., Cazeneuve, C., Hannequin, D., Charles, P., Feki, I., Pinel, J.-F., Ouvrard-Hernandez, A.-M., Lyonnet, S., Ollagnon-Roman, E., … Stevanin, G. (2011). REEP1 mutations in SPG31: Frequency, mutational spectrum, and potential association with mitochondrial morpho-functional dysfunction. Human Mutation, 32(10), 1118–1127. https://doi.org/https://doi.org/10.1002/humu.21542

Gordeeva, V., Sharova, E., Babalyan, K., Sultanov, R., Govorun, V. M., & Arapidi, G. (2021). Benchmarking germline CNV calling tools from exome sequencing data. Scientific Reports, 11(1), 14416. https://doi.org/10.1038/s41598-021-93878-2

Graessner, H., Zurek, B., Hoischen, A., & Beltran, S. (2021). Solving the unsolved rare diseases in Europe. European Journal of Human Genetics : EJHG, 29(9), 1319–1320. https://doi.org/10.1038/S41431-021-00924-8

Greene, D., Thys, C., Berry, I. R., Jarvis, J., Ortibus, E., Mumford, A. D., Freson, K., & Turro, E. (2024). Mutations in the U4 snRNA gene RNU4-2 cause one of the most prevalent monogenic neurodevelopmental disorders. Nature Medicine. https://doi.org/10.1038/S41591-024-03085-5

Griffin, H. R., Pyle, A., Blakely, E. L., Alston, C. L., Duff, J., Hudson, G., Horvath, R., Wilson, I. J., Santibanez-Koref, M., Taylor, R. W., & Chinnery, P. F. (2014). Accurate mitochondrial DNA sequencing using off-target reads provides a single test to identify pathogenic point mutations. Genetics in Medicine, 16(12), 962–971. https://doi.org/10.1038/gim.2014.66

Gross, A. M., Ajay, S. S., Rajan, V., Brown, C., Bluske, K., Burns, N. J., Chawla, A., Coffey, A. J., Malhotra, A., Scocchia, A., Thorpe, E., Dzidic, N., Hovanes, K., Sahoo, T., Dolzhenko, E., Lajoie, B., Khouzam, A., Chowdhury, S., Belmont, J., … Taft, R. J. (2019). Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. Genetics in Medicine, 21(5), 1121–1130. https://doi.org/https://doi.org/10.1038/s41436-018-0295-y

Grosz, B. R., Stevanovski, I., Negri, S., Ellis, M., Barnes, S., Reddel, S., Vucic, S., Nicholson, G. A., Cortese, A., Kumar, K. R., Deveson, I. W., & Kennerson, M. L. (2022). Long read sequencing overcomes challenges in the diagnosis of SORD neuropathy. Journal of the Peripheral Nervous System, 27(2), 120–126. https://doi.org/https://doi.org/10.1111/jns.12485

Grunau, C., Clark, S. J., & Rosenthal, A. (2001). Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. Nucleic Acids Research, 29(13), e65–e65. https://doi.org/10.1093/nar/29.13.e65

Guo, Y., Wei, X., Das, J., Grimson, A., Lipkin, S. M., Clark, A. G., & Yu, H. (2013). Dissecting Disease Inheritance Modes in a Three-Dimensional Protein Network Challenges the "Guilt-by-Association" Principle. The American Journal of Human Genetics, 93(1), 78–89. https://doi.org/https://doi.org/10.1016/j.ajhg.2013.05.022

H., S. P., O., K. J., Francesca, A., Can, A., Maika, M., Anya, T., Nick, S., Laurakay, B., Jay, S., null, null, & E., E. E. (2010). Diversity of Human Copy Number Variation and Multicopy Genes. Science, 330(6004), 641–646. https://doi.org/10.1126/science.1197005

Hancks, D. C., & Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. Mobile DNA, 7(1), 9. https://doi.org/10.1186/s13100-016-0065-9

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., & McCarroll, S. A. (2015). Large multiallelic copy number variations in humans. Nature Genetics, 47(3), 296–303. https://doi.org/10.1038/ng.3200

Hazan, J., Fonknechten, N., Mavel, D., Paternotte, C., Samson, D., Artiguenave, F., Davoine, C.-S., Cruaud, C., Dürr, A., Wincker, P., Brottier, P., Cattolico, L., Barbe, V., Burgunder, J.-M., Prud'homme, J.-F., Brice, A., Fontaine, B., Heilig, R., & Weissenbach, J. (1999). Spastin, a new AAA protein, is altered in the most frequent form of autosomal dominant spastic paraplegia. Nature Genetics, 23(3), 296–303. https://doi.org/10.1038/15472

Hehir-Kwa, J. Y., Pfundt, R., & Veltman, J. A. (2015). Exome sequencing and whole genome sequencing for the detection of copy number variation. Expert Review of Molecular Diagnostics, 15(8), 1023–1032. https://doi.org/10.1586/14737159.2015.1053467

Heinrich, V., Stange, J., Dickhaus, T., Imkeller, P., Krüger, U., Bauer, S., Mundlos, S., Robinson, P. N., Hecht, J., & Krawitz, P. M. (2012). The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. Nucleic Acids Research, 40(6), 2426–2431. https://doi.org/10.1093/nar/gkr1073

Hiatt, S. M., Lawlor, J. M. J., Handley, L. H., Latner, D. R., Bonnstetter, Z. T., Finnila, C. R., Thompson, M. L., Boston, L. B., Williams, M., Rodriguez-Nunez, I., Jenkins, J., Kelley, W. V, Bebin, E. M., Lopez, M. A., Hurst, A. C. E., Korf, B. R., Schmutz, J., Grimwood, J., & Cooper, G. M. (2024). Long-read genome sequencing and variant reanalysis increase diagnostic yield in neurodevelopmental disorders. Genome Research. https://doi.org/10.1101/gr.279227.124

Hiz Kurul, S., Oktay, Y., Töpf, A., Szabó, N. Z., Güngör, S., Yaramis, A., Sonmezler, E., Matalonga, L., Yis, U., Schon, K., Paramonov, I., Kalafatcilar, P., Gao, F., Rieger, A., Arslan, N., Yilmaz, E., Ekinci, B., Edem, P. P., Aslan, M., … Horvath, R. (2022). High diagnostic rate of trio exome sequencing in consanguineous families with neurogenetic diseases. Brain, 145(4), 1507–1518. https://doi.org/10.1093/BRAIN/AWAB395

Höps, W., Weiss, M. M., Derks, R., Galbany, J. C., den Ouden, A., van den Heuvel, S., Timmermans, R., Smits, J., Mokveld, T., Dolzhenko, E., Chen, X., van den Wijngaard, A., Eberle, M. A., Yntema, H. G., Hoischen, A., Gilissen, C., & Vissers, L. E. L. M. (2024). HiFi long-read genomes for difficult-to-detect clinically relevant variants. MedRxiv, 2024.09.17.24313798. https://doi.org/10.1101/2024.09.17.24313798

Hu, X., Feng, C., Zhou, Y., Harrison, A., & Chen, M. (2022). DeepTrio: a ternary prediction system for protein–protein interaction using mask multiple parallel convolutional neural networks. Bioinformatics, 38(3), 694–702. https://doi.org/10.1093/bioinformatics/btab737

Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N., Vives, L., Peluso, P., Boitano, M., Chin, C.-S., Korlach, J., Wilson, R. K., & Eichler, E. E. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Research, 27(5), 677–685. https://doi.org/10.1101/gr.214007.116

Huisman, S. A., Redeker, E. J. W., Maas, S. M., Mannens, M. M., & Hennekam, R. C. M. (2013). High rate of mosaicism in individuals with Cornelia de Lange syndrome. Journal of Medical Genetics, 50(5), 339. https://doi.org/10.1136/jmedgenet-2012-101477

Ishiura, H., Doi, K., Mitsui, J., Yoshimura, J., Matsukawa, M. K., Fujiyama, A., Toyoshima, Y., Kakita, A., Takahashi, H., Suzuki, Y., Sugano, S., Qu, W., Ichikawa, K., Yurino, H., Higasa, K., Shibata, S., Mitsue, A., Tanaka, M., Ichikawa, Y., … Tsuji, S. (2018). Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Nature Genetics, 50(4), 581–590. https://doi.org/10.1038/s41588-018-0067-2

Ishiura, H., Takahashi, Y., Hayashi, T., Saito, K., Furuya, H., Watanabe, M., Murata, M., Suzuki, M., Sugiura, A., Sawai, S., Shibuya, K., Ueda, N., Ichikawa, Y., Kanazawa, I., Goto, J., & Tsuji, S. (2014). Molecular epidemiology and clinical spectrum of hereditary spastic paraplegia in the Japanese population based on comprehensive mutational analyses. Journal of Human Genetics, 59(3), 163–172. https://doi.org/10.1038/jhg.2013.139

Jacobsen, J. O. B., Baudis, M., Baynam, G. S., Beckmann, J. S., Beltran, S., Buske, O. J., Callahan, T. J., Chute, C. G., Courtot, M., Danis, D., Elemento, O., Essenwanger, A., Freimuth, R. R., Gargano, M. A., Groza, T., Hamosh, A., Harris, N. L., Kaliyaperumal, R., Lloyd, K. C. K., … Robinson, P. N. (2022). The GA4GH Phenopacket schema defines a computable representation of clinical data. Nature Biotechnology, 40(6), 817–820. https://doi.org/10.1038/S41587-022-01357-4

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., & Farh, K. K. H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell, 176(3), 535-548.e24. https://doi.org/10.1016/J.CELL.2018.12.015

Jarvis, E. D., Formenti, G., Rhie, A., Guarracino, A., Yang, C., Wood, J., Tracey, A., Thibaud-Nissen, F., Vollger, M. R., Porubsky, D., Cheng, H., Asri, M., Logsdon, G. A., Carnevali, P., Chaisson, M. J. P., Chin, C.-S., Cody, S., Collins, J., Ebert, P., … Consortium, H. P. R. (2022). Semi-automated assembly of high-quality diploid human reference genomes. Nature, 611(7936), 519–531. https://doi.org/10.1038/s41586-022-05325-5

Kaplanis, J., Samocha, K. E., Wiel, L., Zhang, Z., Arvai, K. J., Eberhardt, R. Y., Gallone, G., Lelieveld, S. H., Martin, H. C., McRae, J. F., Short, P. J., Torene, R. I., de Boer, E., Danecek, P., Gardner, E. J., Huang, N., Lord, J., Martincorena, I., Pfundt, R., … Retterer, K. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature, 586(7831), 757–762. https://doi.org/10.1038/S41586-020-2832-5

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … Consortium, G. A. D. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Karolak, J. A., Vincent, M., Deutsch, G., Gambin, T., Cogné, B., Pichon, O., Vetrini, F., Mefford, H. C., Dines, J. N., Golden-Grant, K., Dipple, K., Freed, A. S., Leppig, K. A., Dishop, M., Mowat, D., Bennetts, B., Gifford, A. J., Weber, M. A., Lee, A. F., … Stankiewicz, P. (2019). Complex Compound Inheritance of Lethal Lung Developmental Disorders Due to Disruption of the TBX-FGF Pathway. The American Journal of Human Genetics, 104(2), 213–228. https://doi.org/https://doi.org/10.1016/j.ajhg.2018.12.010

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010

Khazeeva, G., Sablauskas, K., van der Sanden, B., Steyaert, W., Kwint, M., Rots, D., Hinne, M., van Gerven, M., Yntema, H., Vissers, L., & Gilissen, C. (2022). DeNovoCNN: a deep learning approach to de novo variant calling in next generation sequencing data. Nucleic Acids Research, 50(17), e97. https://doi.org/10.1093/NAR/GKAC511

Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Battle, A., Aganezov, S., & Schatz, M. C. (2023). Jasmine and Iris: population-scale structural variant comparison and analysis. Nature Methods, 20(3), 408–417. https://doi.org/10.1038/s41592-022-01753-3

Kobayashi, H., Abe, K., Matsuura, T., Ikeda, Y., Hitomi, T., Akechi, Y., Habu, T., Liu, W., Okuda, H., & Koizumi, A. (2011). Expansion of Intronic GGCCTG Hexanucleotide Repeat in NOP56 Causes SCA36, a Type of Spinocerebellar Ataxia Accompanied by Motor Neuron Involvement. The American Journal of Human Genetics, 89(1), 121–130. https://doi.org/10.1016/j.ajhg.2011.05.015

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., Callahan, T. J., Chute, C. G., Est, J. L., Galer, P. D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., … Robinson, P. N. (2021). The human phenotype ontology in 2021. Nucleic Acids Research, 49(D1). https://doi.org/10.1093/nar/gkaa1043

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biology, 20(1), 117. https://doi.org/10.1186/s13059-019-1720-5

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., Quinlan, A. R., Nickerson, D. A., & Eichler, E. E. (2012). Copy number variation detection and genotyping from exome sequence data. Genome Research, 22(8), 1525–1532. https://doi.org/10.1101/GR.138115.112

Kucuk, E., van der Sanden, B. P. G. H., O'Gorman, L., Kwint, M., Derks, R., Wenger, A. M., Lambert, C., Chakraborty, S., Baybayan, P., Rowell, W. J., Brunner, H. G., Vissers, L. E. L. M., Hoischen, A., & Gilissen, C. (2023). Comprehensive de novo mutation discovery with HiFi long-read sequencing. Genome Medicine, 15(1), 34. https://doi.org/10.1186/s13073-023-01183-6

Kurek, K. C., Luks, V. L., Ayturk, U. M., Alomari, A. I., Fishman, S. J., Spencer, S. A., Mulliken, J. B., Bowen, M. E., Yamamoto, G. L., Kozakewich, H. P. W., & Warman, M. L. (2012). Somatic Mosaic Activating Mutations in PIK3CA Cause CLOVES Syndrome. The American Journal of Human Genetics, 90(6), 1108–1115. https://doi.org/10.1016/j.ajhg.2012.05.006

La Spada, A. R., & Taylor, J. P. (2010). Repeat expansion disease: progress and puzzles in disease pathogenesis. Nature Reviews Genetics, 11(4), 247–258. https://doi.org/10.1038/nrg2748

Lagorce, D., Lebreton, E., Matalonga, L., Hongnat, O., Chahdil, M., Piscia, D., Paramonov, I., Ellwanger, K., Köhler, S., Robinson, P., Graessner, H., Beltran, S., Lucano, C., Hanauer, M., & Rath, A. (2024). Phenotypic similarity-based approach for variant prioritization for unsolved rare disease: a preliminary methodological report. European Journal of Human Genetics : EJHG, 32(2), 182–189. https://doi.org/10.1038/S41431-023-01486-7

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … Trust:, T. W. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860–921. https://doi.org/10.1038/35057062

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., Lyoshin, V., Maddipatla, Z., Maiti, R., Mitchell, J., O'Leary, N., Riley, G. R., Shi, W., Zhou, G., Schneider, V., … Kattman, B. L. (2020). ClinVar: improvements to accessing data. Nucleic Acids Research, 48(D1), D835–D844. https://doi.org/10.1093/nar/gkz972

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., & Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Research, 44(D1), D862–D868. https://doi.org/10.1093/nar/gkv1222

Latour, B. L., Van De Weghe, J. C., Rusterholz, T. D. S., Letteboer, S. J. F., Gomez, A., Shaheen, R., Gesemann, M., Karamzade, A., Asadollahi, M., Barroso-Gil, M., Chitre, M., Grout, M. E., van Reeuwijk, J., van Beersum, S. E. C., Miller, C. V, Dempsey, J. C., Morsy, H., Bamshad, M. J., Nickerson, D. A., … Doherty, D. (2020). Dysfunction of the ciliary ARMC9/TOGARAM1 protein module causes Joubert syndrome. The Journal of Clinical Investigation, 130(8), 4423–4439. https://doi.org/https://doi.org/10.1172/JCI131656

Laurent, S., Gehrig, C., Nouspikel, T., Amr, S. S., Oza, A., Murphy, E., Vannier, A., Béna, F. S., Carminho-Rodrigues, M. T., Blouin, J.-L., Cao Van, H., Abramowicz, M., Paoloni-Giacobino, A., & Guipponi, M. (2021). Molecular characterization of pathogenic OTOA gene conversions in hearing loss patients. Human Mutation, 42(4), 373–377. https://doi.org/https://doi.org/10.1002/humu.24167

Laurie, S., Fernandez-Callejo, M., Marco-Sola, S., Trotta, J. R., Camps, J., Chacón, A., Espinosa, A., Gut, M., Gut, I., Heath, S., & Beltran, S. (2016). From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing. Human Mutation. https://doi.org/10.1002/humu.23114

Laurie, S., Piscia, D., Matalonga, L., Corvo, A., Garcia, C., Fernandez-Callejo, M., Hernandez, C., Luengo, C., Ntalis, A. P., Protassio, J., Martinez, I., Pico, D., Thompson, R., Tonda, R., Bayes, M., Bullich, G., Camps, J., Paramonov, I., Trotta, J., … Beltran, S. (2022). The RD-Connect Genome-Phenome Analysis Platform: Accelerating diagnosis, research, and gene discovery for rare diseases. Human Mutation. https://doi.org/10.1002/HUMU.24353

Lee, H., Deignan, J. L., Dorrani, N., Strom, S. P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., Fox, M., Fogel, B. L., Martinez-Agosto, J. A., Wong, D. A., Chang, V. Y., Shieh, P. B., Palmer, C. G. S., Dipple, K. M., Grody, W. W., … Nelson, S. F. (2014). Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. JAMA, 312(18), 1880–1887. https://doi.org/10.1001/jama.2014.14604

Lee, J. J. Y., Wasserman, W. W., Hoffmann, G. F., Van Karnebeek, C. D. M., & Blau, N. (2018). Knowledge base and mini-expert platform for the diagnosis of inborn errors of metabolism. Genetics in Medicine : Official Journal of the American College of Medical Genetics, 20(1), 151–158. https://doi.org/10.1038/GIM.2017.108

Lefebvre, S., Bürglen, L., Reboullet, S., Clermont, O., Burlet, P., Viollet, L., Benichou, B., Cruaud, C., Millasseau, P., & Zeviani, M. (1995). Identification and characterization of a spinal muscular atrophy-determining gene. Cell, 80(1), 155–165. https://doi.org/10.1016/0092-8674(95)90460-3

Lelieveld, S. H., Reijnders, M. R. F., Pfundt, R., Yntema, H. G., Kamsteeg, E. J., De Vries, P., De Vries, B. B. A., Willemsen, M. H., Kleefstra, T., Löhner, K., Vreeburg, M., Stevens, S. J. C., Van Der Burgt, I., Bongers, E. M. H. F., Stegmann, A. P. A., Rump, P., Rinne, T., Nelen, M. R., Veltman, J. A., … Gilissen, C. (2016). Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nature Neuroscience, 19(9), 1194–1196. https://doi.org/10.1038/NN.4352

Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A., & Gilissen, C. (2015). Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. Human Mutation, 36(8), 815–822. https://doi.org/https://doi.org/10.1002/humu.22813

Lemire, G., Sanchis-Juan, A., Russell, K., Baxter, S., Chao, K. R., Singer-Berk, M., Groopman, E., Wong, I., England, E., Goodrich, J., Pais, L., Austin-Tse, C., DiTroia, S., O'Heir, E., Ganesh, V. S., Wojcik, M. H., Evangelista, E., Snow, H., Osei-Owusu, I., … O'Donnell-Luria, A. (2023). Exome copy number variant detection, analysis and classification in a large cohort of families with undiagnosed rare genetic disease. MedRxiv : The Preprint Server for Health Sciences. https://doi.org/10.1101/2023.10.05.23296595

Levy, M. A., Relator, R., McConkey, H., Pranckeviciene, E., Kerkhof, J., Barat-Houari, M., Bargiacchi, S., Biamino, E., Palomares Bralo, M., Cappuccio, G., Ciolfi, A., Clarke, A., DuPont, B. R., Elting, M. W., Faivre, L., Fee, T., Ferilli, M., Fletcher, R. S., Cherick, F., … Sadikovic, B. (2022). Functional correlation of genome-wide DNA methylation profiles in genetic neurodevelopmental disorders. Human Mutation, 43(11), 1609–1628. https://doi.org/https://doi.org/10.1002/humu.24446

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv Preprint ArXiv, 00(00), 3. https://doi.org/arXiv:1303.3997 [q-bio.GN]

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. Bioinformatics, 37(23), 4572–4574. https://doi.org/10.1093/bioinformatics/btab705

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lin, Y.-L., Chang, P.-C., Hsu, C., Hung, M.-Z., Chien, Y.-H., Hwu, W.-L., Lai, F., & Lee, N.-C. (2022). Comparison of GATK and DeepVariant by trio sequencing. Scientific Reports, 12(1), 1809. https://doi.org/10.1038/s41598-022-05833-4

Liu, P., Meng, L., Normand, E. A., Xia, F., Song, X., Ghazi, A., Rosenfeld, J., Magoulas, P. L., Braxton, A., Ward, P., Dai, H., Yuan, B., Bi, W., Xiao, R., Wang, X., Chiang, T., Vetrini, F., He, W., Cheng, H., … Yang, Y. (2019). Reanalysis of Clinical Exome Sequencing Data. The New England Journal of Medicine, 380(25), 2478–2480. https://doi.org/10.1056/NEJMC1812033

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. Nature Reviews Genetics, 21(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

López-Martín, E., Martínez-Delgado, B., Bermejo-Sánchez, E., & Alonso, J. (2018). SpainUDP: The Spanish Undiagnosed Rare Diseases Program. International Journal of Environmental Research and Public Health, 15(8). https://doi.org/10.3390/IJERPH15081746

Lorson, C. L., Hahnen, E., Androphy, E. J., & Wirth, B. (1999). A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. Proceedings of the National Academy of Sciences of the United States of America, 96(11), 6307–6311. https://doi.org/10.1073/pnas.96.11.6307

Lott, M. T., Leipzig, J. N., Derbeneva, O., Michael Xie, H., Chalkia, D., Sarmady, M., Procaccio, V., & Wallace, D. C. (2013). mtDNA Variation and Analysis Using Mitomap and Mitomaster. Current Protocols in Bioinformatics, 44(123). https://doi.org/10.1002/0471250953.BI0123S44

Madsen, E. B., Höijer, I., Kvist, T., Ameur, A., & Mikkelsen, M. J. (2020). Xdrop: Targeted sequencing of long DNA molecules from low input samples using droplet sorting. Human Mutation, 41(9), 1671–1679. https://doi.org/https://doi.org/10.1002/humu.24063

Mahmoud, M., Huang, Y., Garimella, K., Audano, P. A., Wan, W., Prasad, N., Handsaker, R. E., Hall, S., Pionzio, A., Schatz, M. C., Talkowski, M. E., Eichler, E. E., Levy, S. E., & Sedlazeck, F. J. (2024). Utility of long-read sequencing for All of Us. Nature Communications, 15(1), 837. https://doi.org/10.1038/s41467-024-44804-3

Mak, C. C. Y., Doherty, D., Lin, A. E., Vegas, N., Cho, M. T., Viot, G., Dimartino, C., Weisfeld-Adams, J. D., Lessel, D., Joss, S., Li, C., Gonzaga-Jauregui, C., Zarate, Y. A., Ehmke, N., Horn, D., Troyer, C., Kant, S. G., Lee, Y., Ishak, G. E., … Gordon, C. T. (2020). MN1 C-terminal truncation syndrome is a novel neurodevelopmental and craniofacial disorder with partial rhombencephalosynapsis. Brain : A Journal of Neurology, 143(1), 55–68. https://doi.org/10.1093/BRAIN/AWZ379

Mandelker, D., Schmidt, R. J., Ankala, A., McDonald Gibson, K., Bowser, M., Sharma, H., Duffy, E., Hegde, M., Santani, A., Lebo, M., & Funke, B. (2016). Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. Genetics in Medicine, 18(12), 1282–1289. https://doi.org/10.1038/gim.2016.58

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics (Oxford, England), 26(22), 2867–2873. https://doi.org/10.1093/BIOINFORMATICS/BTQ559

Manolio, T. A., Rowley, R., Williams, M. S., Roden, D., Ginsburg, G. S., Bult, C., Chisholm, R. L., Deverka, P. A., McLeod, H. L., Mensah, G. A., Relling, M. V, Rodriguez, L. L., Tamburro, C., & Green, E. D. (2019). Opportunities, resources, and techniques for implementing genomics in clinical care. The Lancet, 394(10197), 511–520. https://doi.org/https://doi.org/10.1016/S0140-6736(19)31140-7

Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. Frontiers in Genetics, 10(MAY). https://doi.org/10.3389/FGENE.2019.00426

Marshall, D. A., MacDonald, K. V., Heidenreich, S., Hartley, T., Bernier, F. P., Gillespie, M. K., McInnes, B., Innes, A. M., Armour, C. M., & Boycott, K. M. (2019). The value of diagnostic testing for parents of children with rare genetic diseases. Genetics in Medicine : Official Journal of the American College of Medical Genetics, 21(12), 2798–2806. https://doi.org/10.1038/S41436-019-0583-1

Martin, A. R., Williams, E., Foulger, R. E., Leigh, S., Daugherty, L. C., Niblock, O., Leong, I. U. S., Smith, K. R., Gerasimenko, O., Haraldsdottir, E., Thomas, E., Scott, R. H., Baple, E., Tucci, A., Brittain, H., de Burca, A., Ibañez, K., Kasperaviciute, D., Smedley, D., … McDonagh, E. M. (2019). PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nature Genetics 2019 51:11, 51(11), 1560–1565. https://doi.org/10.1038/s41588-019-0528-2

Martin, H. C., Kim, G. E., Pagnamenta, A. T., Murakami, Y., Carvill, G. L., Meyer, E., Copley, R. R., Rimmer, A., Barcia, G., Fleming, M. R., Kronengold, J., Brown, M. R., Hudspith, K. A., Broxholme, J., Kanapin, A., Cazier, J.-B., Kinoshita, T., Nabbout, R., Consortium, T. W., … Taylor, J. C. (2014). Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis. Human Molecular Genetics, 23(12), 3200–3211. https://doi.org/10.1093/hmg/ddu030

Martin, M., Patterson, M., Garg, S., Fischer, S. O., Pisanti, N., Klau, G. W., Schöenhuth, A., & Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. BioRxiv, 085050. https://doi.org/10.1101/085050

Matalonga, L., Hernandez-Ferrer, C., Piscia, D., Schüle, R., Synofzik, M., Töpf, A., Vissers, L. E. L. M., de Voer, R., Tonda, R., Laurie, S., Fernandez-Callejo, M., Picó, D., Garcia-Linares, C., Papakonstantinou, A., Corvò, A., Joshi, R., Diez, H., Gut, I., Hoischen, A., … Beltran, S. (2021). Solving patients with rare diseases through programmatic reanalysis of genome-phenome data. European Journal of Human Genetics. https://doi.org/10.1038/s41431-021-00852-7

Matilla-Dueñas, A., & Volpini, V. (1993). Spinocerebellar Ataxia Type 37. University of Washington, Seattle, Seattle (WA). http://europepmc.org/abstract/MED/31145571

McConnell, M. J., Lindberg, M. R., Brennand, K. J., Piper, J. C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R. S., Vermeesch, J. R., Hall, I. M., & Gage, F. H. (2013). Mosaic Copy Number Variation in Human Neurons. Science, 342(6158), 632–637. https://doi.org/10.1126/science.1243472

McDermott, J. H., Hickson, N., Banerjee, I., Murray, P. G., Ram, D., Metcalfe, K., Clayton-Smith, J., & Douzgou, S. (2018). Hypoglycaemia represents a clinically significant manifestation of PIK3CA- and CCND2-associated segmental overgrowth. Clinical Genetics, 93(3), 687–692. https://doi.org/10.1111/CGE.13145

McDonald, T. L., Zhou, W., Castro, C. P., Mumm, C., Switzenberg, J. A., Mills, R. E., & Boyle, A. P. (2021). Cas9 targeted enrichment of mobile elements using nanopore sequencing. Nature Communications, 12(1), 3586. https://doi.org/10.1038/s41467-021-23918-y

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biology, 17(1). https://doi.org/10.1186/s13059-016-0974-4

Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., Waggott, D., Utiramerur, S., Hou, Y., Smith, K. S., Montgomery, S. B., Wheeler, M., Buchan, J. G., Lambert, C. C., Eng, K. S., Hickey, L., Korlach, J., Ford, J., & Ashley, E. A. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genetics in Medicine : Official Journal of the American College of Medical Genetics, 20(1), 159–163. https://doi.org/10.1038/GIM.2017.86

Meynert, A. M., Ansari, M., FitzPatrick, D. R., & Taylor, M. S. (2014). Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics, 15(1), 247. https://doi.org/10.1186/1471-2105-15-247

Meynert, A. M., Bicknell, L. S., Hurles, M. E., Jackson, A. P., & Taylor, M. S. (2013). Quantifying single nucleotide variant detection sensitivity in exome sequencing. BMC Bioinformatics, 14(1), 195. https://doi.org/10.1186/1471-2105-14-195

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, V. A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., … Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature, 585(7823), 79–84. https://doi.org/10.1038/s41586-020-2547-7

Mizuguchi, T., Suzuki, T., Abe, C., Umemura, A., Tokunaga, K., Kawai, Y., Nakamura, M., Nagasaki, M., Kinoshita, K., Okamura, Y., Miyatake, S., Miyake, N., & Matsumoto, N. (2019). A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. Journal of Human Genetics, 64(5), 359–368. https://doi.org/10.1038/s10038-019-0569-5

Mizuguchi, T., Toyota, T., Adachi, H., Miyake, N., Matsumoto, N., & Miyatake, S. (2019). Detecting a long insertion variant in SAMD12 by SMRT sequencing: implications of long-read whole-genome sequencing for repeat expansion diseases. Journal of Human Genetics, 64(3), 191–197. https://doi.org/10.1038/s10038-018-0551-7

Molinard-Chenu, A., Fluss, J., Laurent, S., Laurent, M., Guipponi, M., & Dayer, A. G. (2020). MCF2 is linked to a complex perisylvian syndrome and affects cortical lamination. Annals of Clinical and Translational Neurology, 7(1), 121–125. https://doi.org/https://doi.org/10.1002/acn3.50949

Muona, M., Berkovic, S. F., Dibbens, L. M., Oliver, K. L., Maljevic, S., Bayly, M. A., Joensuu, T., Canafoglia, L., Franceschetti, S., Michelucci, R., Markkinen, S., Heron, S. E., Hildebrand, M. S., Andermann, E., Andermann, F., Gambardella, A., Tinuper, P., Licchetta, L., Scheffer, I. E., … Lehesjoki, A.-E. (2015). A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. Nature Genetics, 47(1), 39–46. https://doi.org/10.1038/ng.3144

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Ng, P. C., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. Genome Research, 11(5), 863–874. https://doi.org/10.1101/gr.176601

Nguengang Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y., & Rath, A. (2020). Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. European Journal of Human Genetics, 28(2). https://doi.org/10.1038/s41431-019-0508-0

Noyes, M. D., Harvey, W. T., Porubsky, D., Sulovari, A., Li, R., Rose, N. R., Audano, P. A., Munson, K. M., Lewis, A. P., Hoekzema, K., Mantere, T., Graves-Lindsay, T. A., Sanders, A. D., Goodwin, S., Kramer, M., Mokrab, Y., Zody, M. C., Hoischen, A., Korbel, J. O., … Eichler, E. E. (2022). Familial long-read sequencing increases yield of de novo mutations. The American Journal of Human Genetics, 109(4), 631–646. https://doi.org/10.1016/j.ajhg.2022.02.014

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V, Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. Science, 376(6588), 44–53. https://doi.org/10.1126/science.abj6987

Obayashi, M., Stevanin, G., Synofzik, M., Monin, M.-L., Duyckaerts, C., Sato, N., Streichenberger, N., Vighetto, A., Desestret, V., Tesson, C., Wichmann, H.-E., Illig, T., Huttenlocher, J., Kita, Y., Izumi, Y., Mizusawa, H., Schöls, L., Klopstock, T., Brice, A., … Dürr, A. (2015). Spinocerebellar ataxia type 36 exists in diverse populations and can be caused by a short hexanucleotide GGCCTG repeat expansion. Journal of Neurology, Neurosurgery &Amp; Psychiatry, 86(9), 986. https://doi.org/10.1136/jnnp-2014-309153

Okubo, M., Noguchi, S., Hayashi, S., Nakamura, H., Komaki, H., Matsuo, M., & Nishino, I. (2020). Exon skipping induced by nonsense/frameshift mutations in DMD gene results in Becker muscular dystrophy. Human Genetics, 139(2), 247–255. https://doi.org/10.1007/s00439-019-02107-4

Olson, N. D., Wagner, J., Dwarshuis, N., Miga, K. H., Sedlazeck, F. J., Salit, M., & Zook, J. M. (2023). Variant calling and benchmarking in an era of complete human genome sequences. Nature Reviews Genetics, 24(7), 464–483. https://doi.org/10.1038/s41576-023-00590-0

Palmer, E. E., Sachdev, R., Macintosh, R., Melo, U. S., Mundlos, S., Righetti, S., Kandula, T., Minoche, A. E., Puttick, C., Gayevskiy, V., Hesson, L., Idrisoglu, S., Shoubridge, C., Thai, M. H. N., Davis, R. L., Drew, A. P., Sampaio, H., Andrews, P. I., Lawson, J., … Kirk, E. (2021). Diagnostic Yield of Whole Genome Sequencing After Nondiagnostic Exome Sequencing or Gene Panel in Developmental and Epileptic Encephalopathies. Neurology, 96(13), e1770 LP-e1782. https://doi.org/10.1212/WNL.0000000000011655

Parrish, A., Caswell, R., Jones, G., Watson, C. M., Crinnion, L. A., & Ellard, S. (2017). An enhanced method for targeted next generation sequencing copy number variant detection using ExomeDepth [version 1; peer review: 1 approved, 1 approved with reservations]. Wellcome Open Research, 2(49). https://doi.org/10.12688/wellcomeopenres.11548.1

Pauly, M. G., Brüggemann, N., Efthymiou, S., Grözinger, A., Diaw, S. H., Chelban, V., Turchetti, V., Vona, B., Tadic, V., Houlden, H., Münchau, A., & Lohmann, K. (2023). Not to Miss: Intronic Variants, Treatment, and Review of the Phenotypic Spectrum in VPS13D-Related Disorder. International Journal of Molecular Sciences, 24(3). https://doi.org/10.3390/IJMS24031874

Pauly, M. G., Korenke, G. C., Diaw, S. H., Grözinger, A., Cazurro-Gutiérrez, A., Pérez-Dueñas, B., González, V., Macaya, A., Serrano Antón, A. T., Peterlin, B., Božović, I. B., Maver, A., Münchau, A., & Lohmann, K. (2023). The Expanding Phenotypical Spectrum of WARS2-Related Disorder: Four Novel Cases with a Common Recurrent Variant. Genes, 14(4). https://doi.org/10.3390/GENES14040822

Pauper, M., Kucuk, E., Wenger, A. M., Chakraborty, S., Baybayan, P., Kwint, M., van der Sanden, B., Nelen, M. R., Derks, R., Brunner, H. G., Hoischen, A., Vissers, L. E. L. M., & Gilissen, C. (2021). Long-read trio sequencing of individuals with unsolved intellectual disability. European Journal of Human Genetics, 29(4), 637–648. https://doi.org/10.1038/s41431-020-00770-0

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825–2830.

Philippakis, A. A., Azzariti, D. R., Beltran, S., Brookes, A. J., Brownstein, C. A., Brudno, M., Brunner, H. G., Buske, O. J., Carey, K., Doll, C., Dumitriu, S., Dyke, S. O. M., den Dunnen, J. T., Firth, H. V, Gibbs, R. A., Girdea, M., Gonzalez, M., Haendel, M. A., Hamosh, A., … Rehm, H. L. (2015). The Matchmaker Exchange: a platform for rare disease gene discovery. Human Mutation, 36(10), 915–921. https://doi.org/10.1002/HUMU.22858

Picardi, E., & Pesole, G. (2012). Mitochondrial genomes gleaned from human whole-exome sequencing. Nature Methods, 9(6), 523–524. https://doi.org/10.1038/nmeth.2029

Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., Wood, N. W., Hambleton, S., Burns, S. O., Thrasher, A. J., Kumararatne, D., Doffinger, R., & Nejentsev, S. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics (Oxford, England), 28(21), 2747–2754. https://doi.org/10.1093/bioinformatics/bts526

Pogue, R. E., Cavalcanti, D. P., Shanker, S., Andrade, R. V, Aguiar, L. R., de Carvalho, J. L., & Costa, F. F. (2018). Rare genetic diseases: update on diagnosis, treatment and online resources. Drug Discovery Today, 23(1), 187–195. https://doi.org/https://doi.org/10.1016/j.drudis.2017.11.002

Polla, D. L., Bhoj, E. J., Verheij, J. B. G. M., Wassink-Ruiter, J. S. K., Reis, A., Deshpande, C., Gregor, A., Hill-Karfe, K., Silfhout, A. T. V., Pfundt, R., Bongers, E. M. H. F., Hakonarson, H., Berland, S., Gradek, G., Banka, S., Chandler, K., Gompertz, L., Huffels, S. C., Stumpel, C. T. R. M., … de Brouwer, A. P. M. (2021). De novo variants in MED12 cause X-linked syndromic neurodevelopmental disorders in 18 females. Genetics in Medicine, 23(4), 645–652. https://doi.org/10.1038/s41436-020-01040-6

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 36(10), 983–987. https://doi.org/10.1038/nbt.4235

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Auwera, G. A. Van der, Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. BioRxiv, 201178. https://doi.org/10.1101/201178

Portela, A., & Esteller, M. (2010). Epigenetic modifications and human disease. Nature Biotechnology, 28(10), 1057–1068. https://doi.org/10.1038/nbt.1685

Porubsky, D., & Eichler, E. E. (2024). A 25-year odyssey of genomic technology advances and structural variant discovery. Cell, 187(5), 1024–1037. https://doi.org/https://doi.org/10.1016/j.cell.2024.01.002

Porubsky, D., Vollger, M. R., Harvey, W. T., Rozanski, A. N., Ebert, P., Hickey, G., Hasenfeld, P., Sanders, A. D., Stober, C., Consortium, H. P. R., Korbel, J. O., Paten, B., Marschall, T., & Eichler, E. E. (2023). Gaps and complex structurally variant loci in phased genome assemblies. Genome Research, 33(4), 496–510. https://doi.org/10.1101/gr.277334.122

Qian, Y., Mancini-DiNardo, D., Judkins, T., Cox, H. C., Brown, K., Elias, M., Singh, N., Daniels, C., Holladay, J., Coffee, B., Bowles, K. R., & Roa, B. B. (2017). Identification of pathogenic retrotransposon insertions in cancer predisposition genes. Cancer Genetics, 216–217, 159–169. https://doi.org/https://doi.org/10.1016/j.cancergen.2017.08.002

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramensky, V., Bork, P., & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Research, 30(17), 3894–3900. https://doi.org/10.1093/nar/gkf493

Ramoni, R. B., Mulvihill, J. J., Adams, D. R., Allard, P., Ashley, E. A., Bernstein, J. A., Gahl, W. A., Hamid, R., Loscalzo, J., McCray, A. T., Shashi, V., Tifft, C. J., Adams, D. R., Adams, C. J., Alejandro, M. E., Allard, P., Ashley, E. A., Azamian, M. S., Bacino, C. A., … Wise, A. L. (2017). The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. The American Journal of Human Genetics, 100(2), 185–192. https://doi.org/https://doi.org/10.1016/j.ajhg.2017.01.006

Ramu, A., Noordam, M. J., Schwartz, R. S., Wuster, A., Hurles, M. E., Cartwright, R. A., & Conrad, D. F. (2013). DeNovoGear: de novo indel and point mutation discovery and phasing. Nature Methods, 10(10), 985–987. https://doi.org/10.1038/nmeth.2611

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., Dufke, A., Cremer, K., Hempel, M., Horn, D., Hoyer, J., Joset, P., Röpke, A., Moog, U., Riess, A., … Strom, T. M. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet (London, England), 380(9854), 1674–1682. https://doi.org/10.1016/S0140-6736(12)61480-9

Redfield, S. E., Shao, W., Sun, T., Pastolero, A., Rowell, W. J., French, C. E., Nolan, C., Holt, J. M., Saunders, C. T., Fanslow, C., Lampraki, E. M., Lambert, C., Kenna, M., Eberle, M., Rockowitz, S., & Shearer, A. E. (2024). Long-Read Sequencing Increases Diagnostic Yield for Pediatric Sensorineural Hearing Loss. MedRxiv, 2024.09.30.24314377. https://doi.org/10.1101/2024.09.30.24314377

Rehm, H. L. (2022). Time to make rare disease diagnosis accessible to all. Nature Medicine 2022 28:2, 28(2), 241–242. https://doi.org/10.1038/s41591-021-01657-3

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., Ledbetter, D. H., Maglott, D. R., Martin, C. L., Nussbaum, R. L., Plon, S. E., Ramos, E. M., Sherry, S. T., & Watson, M. S. (2015). ClinGen--the Clinical Genome Resource. The New England Journal of Medicine, 372(23), 2235–2242. https://doi.org/10.1056/NEJMSR1406261

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research, 47(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Renvoisé, B., Malone, B., Falgairolle, M., Munasinghe, J., Stadler, J., Sibilla, C., Park, S. H., & Blackstone, C. (2016). Reep1 null mice reveal a converging role for hereditary spastic paraplegia proteins in lipid droplet regulation. Human Molecular Genetics, 25(23), 5111–5125. https://doi.org/10.1093/hmg/ddw315

Retterer, K., Scuffins, J., Schmidt, D., Lewis, R., Pineda-Alvarez, D., Stafford, A., Schmidt, L., Warren, S., Gibellini, F., Kondakova, A., Blair, A., Bale, S., Matyakhina, L., Meck, J., Aradhya, S., & Haverfield, E. (2015). Assessing copy number from exome sequencing and exome array CGH based on CNV spectrum in a large clinical cohort. Genetics in Medicine, 17(8), 623–629. https://doi.org/10.1038/gim.2014.160

Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., Hook, P. W., Koren, S., Rautiainen, M., Alexandrov, I. A., Allen, J., Asri, M., Bzikadze, A. V, Chen, N.-C., Chin, C.-S., Diekhans, M., Flicek, P., Formenti, G., Fungtammasan, A., … Phillippy, A. M. (2023). The complete sequence of a human Y chromosome. Nature, 621(7978), 344–354. https://doi.org/10.1038/s41586-023-06457-y

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine. https://doi.org/10.1038/gim.2015.30

Riess, O., Sturm, M., Menden, B., Liebmann, A., Demidov, G., Witt, D., Casadei, N., Admard, J., Schütz, L., Ossowski, S., Taylor, S., Schaffer, S., Schroeder, C., Dufke, A., & Haack, T. (2024). Genomes in clinical care. Npj Genomic Medicine, 9(1), 20. https://doi.org/10.1038/s41525-024-00402-2

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., McVean, G., Lunter, G., & Consortium, W. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics, 46(8), 912–918. https://doi.org/10.1038/ng.3036

Rishishwar, L., Mariño-Ramírez, L., & Jordan, I. K. (2017). Benchmarking computational tools for polymorphic transposable element detection. Briefings in Bioinformatics, 18(6), 908–918. https://doi.org/10.1093/bib/bbw072

Rivas, M. A., Pirinen, M., Conrad, D. F., Lek, M., Tsang, E. K., Karczewski, K. J., Maller, J. B., Kukurba, K. R., DeLuca, D. S., Fromer, M., Ferreira, P. G., Smith, K. S., Zhang, R., Zhao, F., Banks, E., Poplin, R., Ruderfer, D. M., Purcell, S. M., Tukiainen, T., … Estivill, X. (2015). Effect of predicted protein-truncating genetic variants on the human transcriptome. Science, 348(6235), 666–669. https://doi.org/10.1126/science.1261877

Rivière, J. B., Mirzaa, G. M., O'Roak, B. J., Beddaoui, M., Alcantara, D., Conway, R. L., St-Onge, J., Schwartzentruber, J. A., Gripp, K. W., Nikkel, S. M., Worthylake, T., Sullivan, C. T., Ward, T. R., Butler, H. E., Kramer, N. A., Albrecht, B., Armour, C. M., Armstrong, L., Caluseriu, O., … Dobyns, W. B. (2012). De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nature Genetics, 44(8), 934–940. https://doi.org/10.1038/NG.2331

Robin, N. H. (2006). It does matter: the importance of making the diagnosis of a genetic syndrome. Current Opinion in Pediatrics, 18(6). https://journals.lww.com/co-pediatrics/fulltext/2006/12000/it_does_matter__the_importance_of_making_the.2.aspx

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. Nature Biotechnology 2011 29:1, 29(1), 24–26. https://doi.org/10.1038/nbt.1754

Roller, E., Ivakhno, S., Lee, S., Royce, T., & Tanner, S. (2016). Canvas: versatile and scalable detection of copy number variants. Bioinformatics, 32(15), 2375–2377. https://doi.org/10.1093/bioinformatics/btw163

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. Genome Biology, 14(5), R51. https://doi.org/10.1186/gb-2013-14-5-r51

Sabatella, M., Mantere, T., Waanders, E., Neveling, K., Mensenkamp, A. R., van Dijk, F., Hehir-Kwa, J. Y., Derks, R., Kwint, M., O'Gorman, L., Tropa Martins, M., Gidding, C. E. M., Lequin, M. H., Küsters, B., Wesseling, P., Nelen, M., Biegel, J. A., Hoischen, A., Jongmans, M. C., & Kuiper, R. P. (2021). Optical genome mapping identifies a germline retrotransposon insertion in SMARCB1 in two siblings with atypical teratoid rhabdoid tumors. The Journal of Pathology, 255(2), 202–211. https://doi.org/10.1002/PATH.5755

Samuels, D. C., Han, L., Li, J., Quanghu, S., Clark, T. A., Shyr, Y., & Guo, Y. (2013). Finding the lost treasures in exome sequencing data. Trends in Genetics, 29(10), 593–599. https://doi.org/https://doi.org/10.1016/j.tig.2013.07.006

Sanchis-Juan, A., Stephens, J., French, C. E., Gleadall, N., Mégy, K., Penkett, C., Shamardina, O., Stirrups, K., Delon, I., Dewhurst, E., Dolling, H., Erwood, M., Grozeva, D., Stefanucci, L., Arno, G., Webster, A. R., Cole, T., Austin, T., Branco, R. G., … Carss, K. J. (2018). Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. Genome Medicine, 10(1), 95. https://doi.org/10.1186/s13073-018-0606-6

Santoyo, G., & Romero, D. (2005). Gene conversion and concerted evolution in bacterial genomes★. FEMS Microbiology Reviews, 29(2), 169–183. https://doi.org/10.1016/j.fmrre.2004.10.004

Savarese, M., Jonson, P. H., Huovinen, S., Paulin, L., Auvinen, P., Udd, B., & Hackman, P. (2018). The complexity of titin splicing pattern in human adult skeletal muscles. Skeletal Muscle, 8(1), 11. https://doi.org/10.1186/s13395-018-0156-z

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. Human Molecular Genetics, 19(R2), R227–R240. https://doi.org/10.1093/hmg/ddq416

Schaefer, A. M., McFarland, R., Blakely, E. L., He, L., Whittaker, R. G., Taylor, R. W., Chinnery, P. F., & Turnbull, D. M. (2008). Prevalence of mitochondrial DNA disease in adults. Annals of Neurology, 63(1), 35–39. https://doi.org/10.1002/ana.21217

Schüle, R., Timmann, D., Erasmus, C. E., Reichbauer, J., Wayand, M., van de Warrenburg, B., Schöls, L., Wilke, C., Bevot, A., Zuchner, S., Beltran, S., Laurie, S., Matalonga, L., Graessner, H., & Synofzik, M. (2021). Solving unsolved rare neurological diseases—a Solve-RD viewpoint. European Journal of Human Genetics. https://doi.org/10.1038/s41431-021-00901-1

Schultzhaus, Z., Wang, Z., & Stenger, D. (2021). CRISPR-based enrichment strategies for targeted sequencing. Biotechnology Advances, 46, 107672. https://doi.org/https://doi.org/10.1016/j.biotechadv.2020.107672

Scocchia, A., Wigby, K. M., Masser-Frye, D., Del Campo, M., Galarreta, C. I., Thorpe, E., McEachern, J., Robinson, K., Gross, A., Bennett, M., Bluske, K., Brown, C. M., Buchanan, A., Burns, B., Burns, N. J., Chandrasekhar, A., Chawla, A., Clause, A. R., Coffey, A. J., … Reporting. (2019). Clinical whole genome sequencing as a first-tier test at a resource-limited dysmorphology clinic in Mexico. Npj Genomic Medicine, 4(1), 5. https://doi.org/10.1038/s41525-018-0076-1

Sergey, N., Sergey, K., Arang, R., Mikko, R., V., B. A., Alla, M., R., V. M., Nicolas, A., Lev, U., Ariel, G., Sergey, A., J., H. S., Mark, D., A., L. G., Michael, A., E., A. S., Matthew, B., G., B. G., Y., B. S., … M., P. A. (2022). The complete sequence of a human genome. Science, 376(6588), 44–53. https://doi.org/10.1126/science.abj6987

Sernadela, P., González-Castro, L., Carta, C., van der Horst, E., Lopes, P., Kaliyaperumal, R., Thompson, M., Thompson, R., Queralt-Rosinach, N., Lopez, E., Wood, L., Robertson, A., Lamanna, C., Gilling, M., Orth, M., Merino-Martinez, R., Posada, M., Taruscio, D., Lochmüller, H., … Oliveira, J. L. (2017). Linked Registries: Connecting Rare Diseases Patient Registries through a Semantic Web Layer. BioMed Research International, 2017, 8327980. https://doi.org/10.1155/2017/8327980

Shashi, V., Schoch, K., Spillmann, R., Cope, H., Tan, Q. K.-G., Walley, N., Pena, L., McConkie-Rosell, A., Jiang, Y.-H., Stong, N., Need, A. C., Goldstein, D. B., Adams, D. R., Alejandro, M. E., Allard, P., Ashley, E. A., Azamian, M. S., Bacino, C. A., Balasubramanyam, A., … Zheng, A. (2019). A comprehensive iterative approach is highly effective in diagnosing individuals who are exome negative. Genetics in Medicine, 21(1), 161–172. https://doi.org/https://doi.org/10.1038/s41436-018-0044-2

Shearer, A. E., Kolbe, D. L., Azaiez, H., Sloan, C. M., Frees, K. L., Weaver, A. E., Clark, E. T., Nishimura, C. J., Black-Ziegelbein, E. A., & Smith, R. J. H. (2014). Copy number variants are a common cause of non-syndromic hearing loss. Genome Medicine, 6(5), 37. https://doi.org/10.1186/gm554

Shefchek, K. A., Harris, N. L., Gargano, M., Matentzoglu, N., Unni, D., Brush, M., Keith, D., Conlin, T., Vasilevsky, N., Zhang, X. A., Balhoff, J. P., Babb, L., Bello, S. M., Blau, H., Bradford, Y., Carbon, S., Carmody, L., Chan, L. E., Cipriani, V., … Osumi-Sutherland, D. (2020). The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Research, 48(D1), D704–D715. https://doi.org/10.1093/NAR/GKZ997

Smedley, D., Smith, K. R., Martin, A., Thomas, E. A., McDonagh, E. M., Cipriani, V., Ellingford, J. M., Arno, G., Tucci, A., Vandrovcova, J., Chan, G., Williams, H. J., Ratnaike, T., Wei, W., Stirrups, K., Ibanez, K., Moutsianas, L., Wielscher, M., Need, A., … Caulfield, M. (2021). 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. The New England Journal of Medicine, 385(20), 1868–1880. https://doi.org/10.1056/NEJMOA2035790

Sone, J., Mitsuhashi, S., Fujita, A., Mizuguchi, T., Hamanaka, K., Mori, K., Koike, H., Hashiguchi, A., Takashima, H., Sugiyama, H., Kohno, Y., Takiyama, Y., Maeda, K., Doi, H., Koyano, S., Takeuchi, H., Kawamoto, M., Kohara, N., Ando, T., … Sobue, G. (2019). Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nature Genetics, 51(8), 1215–1221. https://doi.org/10.1038/s41588-019-0459-y

Spier, I., Horpaopan, S., Vogt, S., Uhlhaas, S., Morak, M., Stienen, D., Draaken, M., Ludwig, M., Holinski-Feder, E., Nöthen, M. M., Hoffmann, P., & Aretz, S. (2012). Deep intronic APC mutations explain a substantial proportion of patients with familial or early-onset adenomatous polyposis. Human Mutation, 33(7), 1045–1050. https://doi.org/10.1002/humu.22082

Stenson, P. D., Mort, M., Ball, E. V, Chapman, M., Evans, K., Azevedo, L., Hayden, M., Heywood, S., Millar, D. S., Phillips, A. D., & Cooper, D. N. (2020). The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. Human Genetics, 139(10), 1197–1207. https://doi.org/10.1007/s00439-020-02199-3

Steyaert, W. (2023). Systematic analysis of paralogous regions in 41,755 exomes uncovers clinically relevant variation. https://doi.org/10.5281/zenodo.8172517

Steyaert, W., Verschuere, S., Coucke, P. J., & Vanakker, O. M. (2021). Comprehensive validation of a diagnostic strategy for sequencing genes with one or multiple pseudogenes using pseudoxanthoma elasticum as a model. Journal of Genetics and Genomics, 48(4), 289–299.

Stoler, N., & Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. NAR Genomics and Bioinformatics, 3(1), lqab019. https://doi.org/10.1093/nargab/lqab019

Stranneheim, H., Lagerstedt-Robinson, K., Magnusson, M., Kvarnung, M., Nilsson, D., Lesko, N., Engvall, M., Anderlid, B.-M., Arnell, H., Johansson, C. B., Barbaro, M., Björck, E., Bruhn, H., Eisfeldt, J., Freyer, C., Grigelioniene, G., Gustavsson, P., Hammarsjö, A., Hellström-Pigg, M., … Wedell, A. (2021). Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. Genome Medicine, 13(1), 40. https://doi.org/10.1186/s13073-021-00855-5

Supernat, A., Vidarsson, O. V., Steen, V. M., & Stokowy, T. (2018). Comparison of three variant callers for human whole genome sequencing. Scientific Reports, 8(1), 17851. https://doi.org/10.1038/s41598-018-36177-7

Tambuyzer, E., Vandendriessche, B., Austin, C. P., Brooks, P. J., Larsson, K., Miller Needleman, K. I., Valentine, J., Davies, K., Groft, S. C., Preti, R., Oprea, T. I., & Prunotto, M. (2020). Therapies for rare diseases: therapeutic modalities, progress and challenges ahead. Nature Reviews Drug Discovery, 19(2), 93–111. https://doi.org/10.1038/s41573-019-0049-9

Tan, R., Wang, Y., Kleinstein, S. E., Liu, Y., Zhu, X., Guo, H., Jiang, Q., Allen, A. S., & Zhu, M. (2014). An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. Human Mutation, 35(7), 899–907. https://doi.org/https://doi.org/10.1002/humu.22537

Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Frontiers in Bioengineering and Biotechnology, 3. https://doi.org/10.3389/fbioe.2015.00092

te Paske, I. B. A. W., Garcia-Pelaez, J., Sommer, A. K., Matalonga, L., Starzynska, T., Jakubowska, A., Valle, L., Capella, G., Aretz, S., Holinski-Feder, E., Steinke-Lange, V., Laner, A., Schröck, E., Rump, A., Ligtenberg, M., Hoischen, A., Geverink, N., Evans, D. G., Tischkowitz, M., … de Voer, R. M. (2021). A mosaic PIK3CA variant in a young adult with diffuse gastric cancer: case report. European Journal of Human Genetics. https://doi.org/10.1038/s41431-021-00853-6

Te Paske, I. B. A. W., Mensenkamp, A. R., Neveling, K., Baert-Desurmont, S., Claes, K. B. M., de Leeneer, K., Elze, L., van den Heuvel, S., van der Post, R. S., van Twuijver, Y., van Ham, T. J., Wagner, A., de Jong, M. M., Leter, E. M., Nielsen, M., Hoogerbrugge, N., Ligtenberg, M. J. L., & De Voer, R. M. (2022). Noncoding Aberrations in Mismatch Repair Genes Underlie a Substantial Part of the Missing Heritability in Lynch Syndrome. Gastroenterology, 163(6), 1691-1694.e7. https://doi.org/10.1053/J.GASTRO.2022.08.041

Teer, J. K., & Mullikin, J. C. (2010). Exome sequencing: the sweet spot before whole genomes. Human Molecular Genetics, 19(R2), R145–51. https://doi.org/10.1093/hmg/ddq333

Thiffault, I., Farrow, E., Zellmer, L., Berrios, C., Miller, N., Gibson, M., Caylor, R., Jenkins, J., Faller, D., Soden, S., & Saunders, C. (2019). Clinical genome sequencing in an unbiased pediatric cohort. Genetics in Medicine, 21(2), 303–310. https://doi.org/https://doi.org/10.1038/s41436-018-0075-8

Thung, D. T., de Ligt, J., Vissers, L. E. M., Steehouwer, M., Kroon, M., de Vries, P., Slagboom, E. P., Ye, K., Veltman, J. A., & Hehir-Kwa, J. Y. (2014). Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biology, 15(10), 488. https://doi.org/10.1186/s13059-014-0488-x

Töpf, A., Johnson, K., Bates, A., Phillips, L., Chao, K. R., England, E. M., Laricchia, K. M., Mullen, T., Valkanas, E., Xu, L., Bertoli, M., Blain, A., Casasús, A. B., Duff, J., Mroczek, M., Specht, S., Lek, M., Ensini, M., MacArthur, D. G., … Straub, V. (2020). Sequential targeted exome sequencing of 1001 patients affected by unexplained limb-girdle weakness. Genetics in Medicine : Official Journal of the American College of Medical Genetics, 22(9), 1478–1488. https://doi.org/10.1038/S41436-020-0840-3

Töpf, A., Pyle, A., Griffin, H., Matalonga, L., Schon, K., Sickmann, A., Schara-Schmidt, U., Hentschel, A., Chinnery, P. F., Kölbel, H., Roos, A., & Horvath, R. (2021). Exome reanalysis and proteomic profiling identified TRIP4 as a novel cause of cerebellar hypoplasia and spinal muscular atrophy (PCH1). European Journal of Human Genetics : EJHG, 29(9), 1348–1353. https://doi.org/10.1038/S41431-021-00851-8

Torene, R. I., Galens, K., Liu, S., Arvai, K., Borroto, C., Scuffins, J., Zhang, Z., Friedman, B., Sroka, H., Heeley, J., Beaver, E., Clarke, L., Neil, S., Walia, J., Hull, D., Juusola, J., & Retterer, K. (2020). Mobile element insertion detection in 89,874 clinical exomes. Genetics in Medicine : Official Journal of the American College of Medical Genetics, 22(5), 974–978. https://doi.org/10.1038/S41436-020-0749-X

Turro, E., Astle, W. J., Megy, K., Gräf, S., Greene, D., Shamardina, O., Allen, H. L., Sanchis-Juan, A., Frontini, M., Thys, C., Stephens, J., Mapeta, R., Burren, O. S., Downes, K., Haimel, M., Tuna, S., Deevi, S. V. V., Aitman, T. J., Bennett, D. L., … Raymond, F. L. (2020). Whole-genome sequencing of patients with rare diseases in a national health system. Nature, 583(7814), 96–102. https://doi.org/10.1038/S41586-020-2434-2

Van De Weghe, J. C., Rusterholz, T. D. S., Latour, B., Grout, M. E., Aldinger, K. A., Shaheen, R., Dempsey, J. C., Maddirevula, S., Cheng, Y.-H. H., Phelps, I. G., Gesemann, M., Goel, H., Birk, O. S., Alanzi, T., Rawashdeh, R., Khan, A. O., Bamshad, M. J., Nickerson, D. A., Neuhauss, S. C. F., … Doherty, D. (2017). Mutations in ARMC9, which Encodes a Basal Body Protein, Cause Joubert Syndrome in Humans and Ciliopathy Phenotypes in Zebrafish. The American Journal of Human Genetics, 101(1), 23–36. https://doi.org/https://doi.org/10.1016/j.ajhg.2017.05.010

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V, Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics, 43(1), 11.10.1-11.10.33. https://doi.org/https://doi.org/10.1002/0471250953.bi1110s43

van der Sanden, B. P. G. H., Corominas, J., de Groot, M., Pennings, M., Meijer, R. P. P., Verbeek, N., van de Warrenburg, B., Schouten, M., Yntema, H. G., Vissers, L. E. L. M., Kamsteeg, E. J., & Gilissen, C. (2021). Systematic analysis of short tandem repeats in 38,095 exomes provides an additional diagnostic yield. Genetics in Medicine : Official Journal of the American College of Medical Genetics, 23(8), 1569–1573. https://doi.org/10.1038/S41436-021-01174-1

Veltman, J. A., & Brunner, H. G. (2012). De novo mutations in human genetic disease. Nature Reviews Genetics, 13(8), 565–575. https://doi.org/10.1038/nrg3241

Vendrell-Mir, P., Barteri, F., Merenciano, M., González, J., Casacuberta, J. M., & Castanera, R. (2019). A benchmark of transposon insertion detection tools using real data. Mobile DNA, 10(1), 53. https://doi.org/10.1186/s13100-019-0197-9

W, R. B., Jane, D., Gerard, M. N., Elizabeth, T., C, B. S., Pavel, D., Matthias, G., F, M. E., E, W. C., W, K. M., Richard, M., Felix, R., Isabelle, S.-G., M, R. S., Qunming, D., Sally, R., Karl, Y., Claudia, O., & Stuart, E. J. (2024). A CFTR Potentiator in Patients with Cystic Fibrosis and the G551D Mutation. New England Journal of Medicine, 365(18), 1663–1672. https://doi.org/10.1056/NEJMoa1105185

Wagner, M., Berutti, R., Lorenz-Depiereux, B., Graf, E., Eckstein, G., Mayr, J. A., Meitinger, T., Ahting, U., Prokisch, H., Strom, T. M., & Wortmann, S. B. (2019). Mitochondrial DNA mutation analysis from exome sequencing—A more holistic approach in diagnostics of suspected mitochondrial disease. Journal of Inherited Metabolic Disease, 42(5), 909–917. https://doi.org/https://doi.org/10.1002/jimd.12109

Walsh, J. B. (1995). How often do duplicated genes evolve new functions? Genetics, 139(1), 421–428. https://doi.org/10.1093/genetics/139.1.421

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Chaisson, M. J. P., Chang, X., Cook-Deegan, R., Felsenfeld, A. L., Fulton, R. S., Garrison, E. P., Garrison, N. A., Graves-Lindsay, T. A., Ji, H., Kenny, E. E., … Consortium, the H. P. R. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature, 604(7906), 437–446. https://doi.org/10.1038/s41586-022-04601-8

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015). Exome Sequencing: Current and Future Perspectives. G3 Genes|Genomes|Genetics, 5(8), 1543–1550. https://doi.org/10.1534/g3.115.018564

Weihl, C. C., Töpf, A., Bengoechea, R., Duff, J., Charlton, R., Garcia, S. K., Domínguez-González, C., Alsaman, A., Hernández-Laín, A., Franco, L. V., Sanchez, M. E. P., Beecroft, S. J., Goullee, H., Daw, J., Bhadra, A., True, H., Inoue, M., Findlay, A. R., Laing, N., … Straub, V. (2023). Loss of function variants in DNAJB4 cause a myopathy with early respiratory failure. Acta Neuropathologica, 145(1). https://doi.org/10.1007/S00401-022-02510-8

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology, 37(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wijngaard, R., Demidov, G., O'Gorman, L., Corominas-Galbany, J., Yaldiz, B., Steyaert, W., de Boer, E., Vissers, L. E. L. M., Kamsteeg, E.-J., & Pfundt, R. (2024). Mobile element insertions in rare diseases: a comparative benchmark and reanalysis of 60,000 exome samples. European Journal of Human Genetics, 32(2), 200–208.

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., Khor, C. C., Petric, R., Hibberd, M. L., & Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Research, 40(22), 11189–11201. https://doi.org/10.1093/nar/gks918

Wortmann, S. B., Oud, M. M., Alders, M., Coene, K. L. M., van der Crabben, S. N., Feichtinger, R. G., Garanto, A., Hoischen, A., Langeveld, M., Lefeber, D., Mayr, J. A., Ockeloen, C. W., Prokisch, H., Rodenburg, R., Waterham, H. R., Wevers, R. A., van de Warrenburg, B. P. C., Willemsen, M. A. A. P., Wolf, N. I., … van Karnebeek, C. D. M. (2022). How to proceed after "negative" exome: A review on genetic diagnostics, limitations, challenges, and emerging new multiomics techniques. Journal of Inherited Metabolic Disease, 45(4), 663–681. https://doi.org/10.1002/JIMD.12507

Wright, C. F., Campbell, P., Eberhardt, R. Y., Aitken, S., Perrett, D., Brent, S., Danecek, P., Gardner, E. J., Chundru, V. K., Lindsay, S. J., Andrews, K., Hampstead, J., Kaplanis, J., Samocha, K. E., Middleton, A., Foreman, J., Hobson, R. J., Parker, M. J., Martin, H. C., … Firth, H. V. (2023). Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom and Ireland. New England Journal of Medicine, 388(17), 1559–1571. https://doi.org/10.1056/NEJMOA2209046/SUPPL_FILE/NEJMOA2209046_DATA-SHARING.PDF

Wright, C. F., FitzPatrick, D. R., & Firth, H. V. (2018). Paediatric genomics: diagnosing rare disease in children. Nature Reviews Genetics, 19(5), 253–268. https://doi.org/10.1038/nrg.2017.116

Wright, C. F., McRae, J. F., Clayton, S., Gallone, G., Aitken, S., FitzGerald, T. W., Jones, P., Prigmore, E., Rajan, D., Lord, J., Sifrim, A., Kelsell, R., Parker, M. J., Barrett, J. C., Hurles, M. E., FitzPatrick, D. R., & Firth, H. V. (2018). Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. Genetics in Medicine : Official Journal of the American College of Medical Genetics, 20(10), 1216–1223. https://doi.org/10.1038/GIM.2017.246

Wu, N., Ming, X., Xiao, J., Wu, Z., Chen, X., Shinawi, M., Shen, Y., Yu, G., Liu, J., Xie, H., Gucev, Z. S., Liu, S., Yang, N., Al-Kateb, H., Chen, J., Zhang, J., Hauser, N., Zhang, T., Tasic, V., … Zhang, F. (2015). TBX6 Null Variants and a Common Hypomorphic Allele in Congenital Scoliosis. New England Journal of Medicine, 372(4), 341–350. https://doi.org/10.1056/NEJMoa1406829

Xicola, R. M., Clark, J. R., Carroll, T., Alvikas, J., Marwaha, P., Regan, M. R., Lopez-Giraldez, F., Choi, J., Emmadi, R., Alagiozian-Angelova, V., Kupfer, S. S., Ellis, N. A., & Llor, X. (2019). Implication of DNA repair genes in Lynch-like syndrome. Familial Cancer, 18(3), 331–342. https://doi.org/10.1007/s10689-019-00128-6

Xu, S., Xiao, S., Zhu, S., Zeng, X., Luo, J., Liu, J., Gao, T., & Chen, N. (2018). A draft genome assembly of the Chinese sillago (Sillago sinica), the first reference genome for Sillaginidae fishes. GigaScience, 7(9), giy108. https://doi.org/10.1093/gigascience/giy108

Yaldiz, B., Kucuk, E., Hampstead, J., Hofste, T., Pfundt, R., Corominas Galbany, J., Rinne, T., Yntema, H. G., Hoischen, A., Nelen, M., Gilissen, C., Riess, O., Haack, T. B., Graessner, H., Zurek, B., Ellwanger, K., Ossowski, S., Demidov, G., Sturm, M., … Consortium, S.-R. (2023). Twist exome capture allows for lower average sequence coverage in clinical exome sequencing. Human Genomics, 17(1), 39. https://doi.org/10.1186/s40246-023-00485-5

Yao, R., Zhang, C., Yu, T., Li, N., Hu, X., Wang, X., Wang, J., & Shen, Y. (2017). Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. Molecular Cytogenetics, 10(1), 30. https://doi.org/10.1186/s13039-017-0333-5

Yauy, K., de Leeuw, N., Yntema, H. G., Pfundt, R., & Gilissen, C. (2020). Accurate detection of clinically relevant uniparental disomy from exome sequencing data. Genetics in Medicine, 22(4), 803–808. https://doi.org/10.1038/s41436-019-0704-x

Yeo, G., & Burge, C. B. (2003). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology, 322–331. https://doi.org/10.1145/640075.640118

Yépez, V. A., Gusic, M., Kopajtich, R., Mertes, C., Smith, N. H., Alston, C. L., Ban, R., Beblo, S., Berutti, R., Blessing, H., Ciara, E., Distelmaier, F., Freisinger, P., Häberle, J., Hayflick, S. J., Hempel, M., Itkis, Y. S., Kishita, Y., Klopstock, T., … Prokisch, H. (2022). Clinical implementation of RNA sequencing for Mendelian disease diagnostics. Genome Medicine, 14(1). https://doi.org/10.1186/S13073-022-01019-9

Yuan, Y., Zhang, J., Chang, Q., Zeng, J., Xin, F., Wang, J., Zhu, Q., Wu, J., Lu, J., Guo, W., Yan, X., Jiang, H., Zhou, B., Li, Q., Gao, X., Yuan, H., Yang, S., Han, D., Mao, Z., … Dai, P. (2014). De novo mutation in ATP6V1B2 impairs lysosome acidification and causes dominant deafness-onychodystrophy syndrome. Cell Research, 24(11), 1370–1373. https://doi.org/10.1038/cr.2014.77

Yun, T., Li, H., Chang, P.-C., Lin, M. F., Carroll, A., & McLean, C. Y. (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics, 36(24), 5582–5589. https://doi.org/10.1093/bioinformatics/btaa1081

Zeng, S., Zhang, M.-Y., Wang, X.-J., Hu, Z.-M., Li, J.-C., Li, N., Wang, J.-L., Liang, F., Yang, Q., Liu, Q., Fang, L., Hao, J.-W., Shi, F.-D., Ding, X.-B., Teng, J.-F., Yin, X.-M., Jiang, H., Liao, W.-P., Liu, J.-Y., … Tang, B.-S. (2019). Long-read sequencing identified intronic repeat expansions in SAMD12 from Chinese pedigrees affected with familial cortical myoclonic tremor with epilepsy. Journal of Medical Genetics, 56(4), 265–270. https://doi.org/10.1136/jmedgenet-2018-105484

Zhai, Y., Zhang, Z., Shi, P., Martin, D. M., & Kong, X. (2021). Incorporation of exome-based CNV analysis makes trio-WES a more powerful tool for clinical diagnosis in neurodevelopmental disorders: A retrospective study. Human Mutation, 42(8), 990–1004. https://doi.org/https://doi.org/10.1002/humu.24222

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., & Wang, L. (2014). CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics (Oxford, England), 30(7), 1006–1007. https://doi.org/10.1093/bioinformatics/btt730

Zhou, X., Feliciano, P., Shu, C., Wang, T., Astrovskaya, I., Hall, J. B., Obiajulu, J. U., Wright, J. R., Murali, S. C., Xu, S. X., Brueggeman, L., Thomas, T. R., Marchenko, O., Fleisch, C., Barns, S. D., Snyder, L. A. G., Han, B., Chang, T. S., Turner, T. N., … Chung, W. K. (2022). Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. Nature Genetics, 54(9), 1305–1319. https://doi.org/10.1038/S41588-022-01148-2

Zimin, A. V, Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C. P., Sonstegard, T. S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J. A., & Salzberg, S. L. (2009). A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology, 10(4), R42. https://doi.org/10.1186/gb-2009-10-4-r42

Zimin, A. V, Puiu, D., Hall, R., Kingan, S., Clavijo, B. J., & Salzberg, S. L. (2017). The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. GigaScience, 6(11), gix097. https://doi.org/10.1093/gigascience/gix097

Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., Sherry, S., Koren, S., Phillippy, A. M., Boutros, P. C., Sahraeian, S. M. E., Huang, V., Rouette, A., Alexander, N., Mason, C. E., Hajirasouliha, I., Ricketts, C., Lee, J., Tearle, R., … Salit, M. (2020). A robust benchmark for detection of germline large deletions and insertions. Nature Biotechnology, 38(11), 1347–1355. https://doi.org/10.1038/s41587-020-0538-8

Züchner, S., Wang, G., Tran-Viet, K.-N., Nance, M. A., Gaskell, P. C., Vance, J. M., Ashley-Koch, A. E., & Pericak-Vance, M. A. (2006). Mutations in the Novel Mitochondrial Protein REEP1 Cause Hereditary Spastic Paraplegia Type 31. The American Journal of Human Genetics, 79(2), 365–369. https://doi.org/10.1086/505361

Zurek, B., Ellwanger, K., Vissers, L. E. L. M., Schüle, R., Synofzik, M., Töpf, A., de Voer, R. M., Laurie, S., Matalonga, L., Gilissen, C., Ossowski, S., 't Hoen, P. A. C., Vitobello, A., Schulze-Hentrich, J. M., Riess, O., Brunner, H. G., Brookes, A. J., Rath, A., Bonne, G., … Graessner, H. (2021). Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. European Journal of Human Genetics. https://doi.org/10.1038/s41431-021-00859-0

Chapter 7

Summaries

English summary

Rare diseases (RD), the vast majority of which have a genetic origin, are individually rare but collectively affect more than 400 million people worldwide. As a result, these diseases pose a serious economical, societal and healthcare burden. Knowledge of the precise cause of the disease is, among other reasons, needed to optimize medical treatment and to allow for reproductive options such as pre-implantation and prenatal diagnosis. Currently, comprehensive analyses of genome-wide sequencing data identifies the disease-causing genetic variant only in about half of RD patients. Although complete genomes can be sequenced today at relatively low cost and high quality, there is a substantial number of genetic variants that cannot or only incompletely be identified using current technologies and methodologies. Examples of these variants include alterations of the mitochondrial DNA, variants in the paralogous regions of the genome and structural variants (an overview is given in **chapter 1**). Shortcomings at the variant identification level are however not the only reason why a genetic diagnosis cannot be established. The clinical interpretation of genetic variants is another challenge preventing us from achieving higher diagnostic yield rates. The work that has been conducted as part of this PhD is focused on the development and application of bioinformatic methods to both improve variant identification and variant interpretation, ultimately aiming for higher diagnostic success.

In **chapter 2** I describe the design and application of a method (Chameleolyser) that enables the accurate sequence analysis of paralogous regions in the genome. To identify genetic variants in identical paralogs, Chameleolyser extracts reads from a conventional read alignment and realigns them onto a reference genome in which all but one paralog in a set of paralogs is masked. Subsequently, a sensitive variant calling is applied. Apart from identifying variants in identical paralogs, Chameleolyser also enables the identification of homozygous deletions and homozygous ectopic gene conversions. The application of our method on the diagnostic exome cohort from the Radboudumc resulted in the identification of an average of 60 genetic variants in a single exome that cannot be identified with conventional analysis pipelines. Having demonstrated that these hidden genetic variants could be identified using our new method, we evaluated the fraction of genetic diagnoses that occur in RD patients as a result of a so-far hidden disease-causing variants. Clinical interpretation of these variants in yet undiagnosed RD patients resulted in 25 genetic diagnoses (0.14%) that could not be established using regular exome analysis techniques.

Chapter 3 describes a large-scale reanalysis of exome and genome sequencing data from 6,004 previously undiagnosed RD families. This study is part of the Solve-RD research project which is the largest multi-center re-analysis study worldwide. Sequence, phenotypic and pedigree data was collected from 37 institutes across Europe and uniformly reanalysed in five different data analysis working groups. The variants that were identified by these data analysis experts were then jointly interpreted with European disease experts. We demonstrated that this 2-level expert review system is highly efficient as it only took 4,253 expert review hours to interpret variants in 6,004 RD families. In total, a genetic diagnosis was established in 506 (8.4%) families by considering well-established disease genes. The parallel *ad-hoc* expert review analysis at local sites provided an additional 249 (4.1%) diagnosis. Furthermore, 378 variants in 333 RD families were identified as candidate disease-causing. We investigated why the genetic diagnoses were made as part of our systematic reanalysis effort but not before. First, 12.5% of diagnoses result from pathogenic variants in genes which were only found to be a disease gene in or after the year 2018, the starting year of the project. Second, pathogenic variants in 32.6% of diagnosed RD families were variants of unknown significance before 2018 but were reclassified in ClinVar between 2018 and 2021. Third, 37.7% of RD families were diagnosed a result of pathogenic variants which were still not present in ClinVar in 2021 (as a pathogenic variant), but were interpreted as disease-causing by collaborating clinical experts from across Europe. These three reasons together, sum up to 82.8% of diagnoses, related to variant interpretation. The remaining 17.2% of novel genetic diagnoses were generally easy-to-interpret but difficult-to-identify variants. These more exotic variant types include copy number variants, mobile element insertions, short tandem repeat expansions, other structural variants, non-canonical splice site variants and mitochondrial DNA variants. The complete set of sequencing, phenotypic and pedigree data together with the clinically interpreted variants were made available via the European Genome Archive (EGA) as well as Genome-Phenome Analysis Platform (GPAP) as a resource for the whole RD community. This ever growing resource will still allow for new diagnoses and discoveries.

The research project in **chapter 4** is also part of the Solve-RD research project. Here, HiFi long-read genome sequencing was applied to a unique cohort of 293 individuals from 114 previously undiagnosed RD families which was compiled by disease-experts from four European reference networks (ERN-ITHACA, ERN-RND, EURO-NMD and ERN EpiCARE). Included in this cohort are patients with a well-characterized clinical syndrome for which no disease gene is yet identified (the unsolvables; 61 individuals in 21 families). The remaining RD families are

exome or genome negative patients (and healthy relatives) that present with a rare neurological or neuromuscular disease (the unsolved; 232 affected individuals in 93 families). Because the analysis of short-read sequencing data did not reveal the disease-causing variant in any of these families, it was hypothesized that so-far hidden SVs (including short tandem repeat expansions) or SNVs might be responsible for the phenotype. To identify these, we selected all ultra-rare SVs and SNVs that co-segregate with disease and that reside in a well-established disease gene. In parent-offspring trios (healthy parents with an affected child) a dedicated *de novo* variant discovery was conducted. Clinical interpretation of these variants solved one parent-offspring trio, originally diagnosed with the 'unsolvable syndrome' Aicardi syndrome, in which a disease-explanatory *de novo* missense variant in the *TUBA1A* gene was identified. Careful inspection of the patient's clinical features indeed confirmed that she was initially clinically misdiagnosed, probably as result of the clinical overlap between Lissencephaly 3 (caused by pathogenic variants in *TUBA1A*) and Aicardi syndrome. In the subcohort of unsolved RD families, 12 (13%) genetic diagnosis were established, all in well-established disease genes: 3 SVs, 4 STR expansions and 5 SNVs. In addition to these firm diagnoses, candidate disease causing variants are presented partially in novel genes/loci. Our analysis demonstrates the added value of HiFi long-read genome sequencing of previously undiagnosed RD families.

In **chapter 5** I discuss the results and relevance of the work that I did over the past five years. I elaborate on our achievements and the challenges that still remain. Furthermore, I reflect on how the field might evolve in the foreseeable future as sequencing costs further decrease and quality of output and analysis algorithms further improve. All of the new technologies that are adopted by the research community, progressively provide answers for the numerous patients with clinical diagnostic questions.

Nederlandse samenvatting

Zeldzame ziekten (ZZ), waarvan de meerderheid een genetische oorsprong heeft, zijn op individueel niveau zeldzaam, maar treffen wereldwijd meer dan 400 miljoen mensen. Deze ziekten leggen een aanzienlijke last op de economie, samenleving en gezondheidszorg. Kennis over de exacte oorzaken is essentieel om medische behandelingen te optimaliseren en reproductieve opties zoals pre-implantatie- en prenatale diagnostiek mogelijk te maken. Momenteel wordt bij ongeveer de helft van de ZZ-patiënten een ziekteveroorzakende genetische variant gevonden na uitgebreide analyse van genoombrede sequentiegegevens. Hoewel het sequencen van volledige genomen tegenwoordig relatief goedkoop en van hoge kwaliteit is, blijft er een aanzienlijk aantal genetische varianten over die met de huidige technologieën en methoden niet of slechts gedeeltelijk geïdentificeerd kunnen worden. Voorbeelden hiervan zijn veranderingen in mitochondriaal DNA, varianten in de paraloge regio's van het genoom en structurele varianten (een overzicht hiervan wordt gegeven in **hoofdstuk 1**). Deze tekortkomingen in variantidentificatie zijn echter niet de enige reden waarom een genetische diagnose uitblijft. De klinische interpretatie van genetische varianten vormt een andere uitdaging die het stellen van meer genetische diagnoses in de weg staat. Het onderzoek uitgevoerd als onderdeel van dit PhD-traject focust op de ontwikkeling en toepassing van bio-informatica methoden om zowel de identificatie als de interpretatie van varianten te verbeteren, met als uiteindelijk doel het aantal diagnoses te vergroten.

In **hoofdstuk 2** beschrijf ik het ontwerp en de toepassing van een methode genaamd Chameleolyser, die de nauwkeurige sequentieanalyse van paraloge regio's in het genoom mogelijk maakt. Chameleolyser extraheert reads uit een conventioneel read alignment en hermapt deze op een referentiegenoom waarin alle paralogen behalve één gemaskeerd zijn, waarna een sensitieve variant calling wordt uitgevoerd. Naast het identificeren van varianten in identieke paralogen, stelt Chameleolyser ons ook in staat om homozygote deleties en ectopische genconversies te identificeren. De toepassing van onze methode op het diagnostische exoomcohort van het Radboudumc resulteerde in de identificatie van gemiddeld 60 genetische varianten per exoom die niet met conventionele analysepijplijnen konden worden geïdentificeerd. Nadat was aangetoond dat deze verborgen genetische varianten konden worden geïdentificeerd met onze nieuwe methode, evalueerden we het aandeel van genetische diagnosen die voorkomen bij ZZ-patiënten als gevolg van tot nu toe verborgen ziekteveroorzakende varianten. De klinische interpretatie van deze varianten bij nog niet gediagnosticeerde ZZ-patiënten resulteerde in 25 genetische diagnosen (0,14%) die niet konden worden vastgesteld met reguliere exoomanalysetechnieken.

Hoofdstuk 3 beschrijft een grootschalige heranalyse van exoom- en genoomsequentiegegevens van 6.004 ongediagnosticeerde ZZ-families. Deze studie, de grootste multicenter heranalysestudie ter wereld, maakt deel uit van het Solve-RD-onderzoeksproject. Sequentie-, fenotypische en stamboomgegevens werden verzameld vanuit 37 instituten in heel Europa en uniform geheranalyseerd in vijf verschillende data-analysegroepen. De varianten die door deze data-analyse-experts werden geïdentificeerd, werden vervolgens geïnterpreteerd door Europese ziekte-experts. We hebben aangetoond dat dit systeem van expertbeoordeling op twee niveaus zeer efficiënt is, aangezien het slechts 4.253 expertbeoordelingsuren kostte om varianten in 6.004 ZZ-families te interpreteren. In totaal werd in 506 (8,4%) van de families een genetische diagnose gesteld in gekende ziektegenen. De parallelle ad-hoc expertbeoordelingsanalyse (niet gecentraliseerd) leverde een aanvullende 249 (4,1%) diagnoses op. Verder werden 378 varianten in 333 ZZ-families geïdentificeerd als kandidaat-ziekteveroorzakend. Vervolgens onderzochten we waarom deze genetische diagnoses werden gesteld als onderdeel van onze systematische heranalyse-inspanning maar nog niet eerder aan het licht waren gekomen. Ten eerste was 12,5% van de diagnoses het resultaat van pathogene varianten in genen die pas in of na 2018 als ziektegenen werden erkend (het jaar waarin de studie is gestart). Ten tweede waren pathogene varianten in 32,6% van de gediagnosticeerde ZZ-families van onbekende klinische betekenis voor 2018, maar werden ze tussen 2018 en 2021 geherclassificeerd in ClinVar. Ten derde werd 37,7% van de ZZ-families gediagnosticeerd als gevolg van pathogene varianten die in 2021 nog steeds niet aanwezig waren in ClinVar (als pathogene variant), maar werden geïnterpreteerd als ziekteveroorzakend door samenwerkende klinische experts uit heel Europa. Deze drie redenen zijn allen gerelateerd aan variantinterpretatie en vormen samen 82,8% van de diagnosen. De resterende 17,2% van de nieuwe genetische diagnosen worden veroorzaakt door over het algemeen gemakkelijk te interpreteren, maar moeilijk te identificeren varianten. Deze meer exotische varianttypes omvatten CNVs, MEIs, STR expansies en andere SVs, alsook niet-canonieke splice-sitevarianten en mitochondriale DNA-varianten. De complete set van sequencing-, fenotypische en stamboomgegevens samen met de klinisch geïnterpreteerde varianten werden beschikbaar gesteld voor de gehele ZZ gemeenschap via het 'European Genome-Phenome Archive' (EGA) en het 'Genome-Phenome Analysis Platform' (GPAP).

In **hoofdstuk 4**, hetgeen ook deel uitmaakt van het Solve-RD onderzoeksproject, werd HiFi longread genoomsequencing toegepast op een unieke cohort van 293 individuen uit 114 niet gediagnosticeerde ZZ-families, samengesteld door ziekte-experts van vier Europese referentienetwerken (ERN-ITHACA, ERN-RND,

EURO-NMD en ERN EpiCARE). In deze cohort bevinden zich patiënten met een goed gekarakteriseerd klinisch syndroom waarvoor nog geen ziektegen is gekend ('the unsolvables'; 61 individuen in 21 families). De resterende ZZ-families zijn exoom- of genoom negatieve patiënten (en gezonde familieleden) die een zeldzame neurologische of neuromusculaire ziekte vertonen ('the unsolved'; 232 individuen in 93 families). Omdat de analyse van short-read sequentiegegevens de ziekteveroorzakende variant in geen van deze families aan het licht bracht, werd gehypothetiseerd dat tot nu toe verborgen SVs (inclusief STR expansies) of SNVs verantwoordelijk zouden kunnen zijn voor het fenotype. Om deze te identificeren, selecteerden we alle ultrazeldzame SVs en SNVs die co-segregeren met de ziekte en die zich bevinden in een gekend ziektegen. In ouder-nakomelingtrio's (gezonde ouders met een aangetast kind) werd een *de novo* variantontdekking uitgevoerd. Klinische interpretatie van deze varianten leverde een genetische diagnose op: in een patiënt die oorspronkelijk klinisch gediagnosticeerd werd met het Aicardi-syndroom werd een ziekteverklarende *de novo* missense variant in het *TUBA1A*-gen geïdentificeerd. Een zorgvuldige inspectie van de klinische kenmerken van de patiënt bevestigde inderdaad dat zij aanvankelijk klinisch verkeerd was gediagnosticeerd, waarschijnlijk als gevolg van de klinische overlap tussen Lissencephaly 3 (veroorzaakt door pathogene varianten in *TUBA1A*) en Aicardi-syndroom. In de subcohort van 'unsolved' ZZ-families werden 12 (13%) genetische diagnosen vastgesteld, allemaal in gekende ziektegenen: 2 SVs, 3 STR-expansies en 2 SNVs. Naast deze genetische diagnosen worden kandidaat-ziekteveroorzakende varianten gevonden in nog eens 4.3% van deze families. Onze analyse toont de toegevoegde waarde aan van HiFi longread genoomsequencing in exoom en/of genoom negatieve ZZ-families.

In **hoofdstuk 5** bespreek ik de resultaten en relevantie van het werk dat ik de afgelopen vijf jaar heb gedaan. Ik ga dieper in op onze bevindingen en de uitdagingen die nog resten. Verder reflecteer ik op hoe het veld zich in de nabije toekomst zou kunnen ontwikkelen naarmate de sequencingkosten verder dalen en de kwaliteit van output en analysealgoritmen verder verbeteren. Alle nieuwe technologieën die door de hele onderzoeksgemeenschap worden ontwikkeld en toegepast, bieden geleidelijk antwoorden voor de talrijke patiënten met diagnostische vragen.

Appendices

Description of the research data management

Ethics and privacy

All research conducted in this thesis involves data from human individuals, including rare disease patients and their healthy relatives. All data were pseudonymized to protect patient identities. This pseudonymization ensures that data cannot be directly traced back to individuals, thereby maintaining their privacy. The research strictly adhered to the principles set out in the Declaration of Helsinki, which outlines ethical guidelines for medical research involving human subjects.

Ethical considerations were crucial throughout the research process. Each study was reviewed and approved by the relevant institutional review boards and ethics committees. For Chapter 2, the institutional review board of the Radboud University Medical Center (approval number 2020-7142) provided ethical approval. For Chapters 3 and 4, the Ethics Committee of the University of Tübingen, Germany (ClinicalTrials.gov Nr.: NCT03491280), granted ethical approval.

Data collection and storage

The raw data that we have used for Chapter 2 were collected through routine genetic investigations at the Radboudumc. All of these data were analyzed on a secure and dedicated high-performance computing infrastructure (TURBO).

For Chapter 3, individuals were recruited via four European Reference Networks (ERNs). Inclusion criteria required a clinical rare disease diagnosis in at least one family member, with inconclusive exome or genome analysis at the time of submission. Data collection did not exclude participants based on sex, gender, ethnicity, race, age, or any other socially relevant groupings. We collected genomics, phenotypic and pedigree information. These data were uploaded by the various submitters to a dedicated server at CNAG (Centro Nacional de Análisis Genómico) in Barcelona. Subsequently, the bioinformatics laboratories at Radboudumc and Tübingen downloaded the data for analysis. Portions of this dataset were also downloaded to a central analysis server (the Sandbox), which is accessible to authorized Solve-RD analysts.

Chapter 4 followed similar inclusion criteria and ethical standards as Chapter 3. Participants were recruited via European Reference Networks, and the data collection did not exclude anyone based on socio-demographic factors. Each patient entry was linked to the respective ERN, and data suitability was the responsibility of the submitter. Informed consent was obtained from all participants, ensuring compliance with ethical standards.

Data sharing

Data sharing is a critical aspect of this research, facilitating further scientific exploration and validation. The raw sequencing data that we have used in chapter 2 can only be disclosed with specific consent from individual patients, respecting privacy and consent. Scientific results and summary data are available as supplementary material attached to the scientific publication. The data that we generated for validation purposes have been deposited in the European Genome-phenome Archive (EGA) under the accession codes EGAS00001006479 (long-read genome sequencing for individuals with biobank consent) and EGAS00001007513 (*STRC* amplicon sequencing). These datasets are available under restricted access, and their re-use will be evaluated by a data access committee to ensure it aligns with the provided consent.

For Chapter 3, pseudonymized phenotypic information for all individuals and their genetic variants are accessible through the RD-Connect Genome-Phenome Analysis Platform (GPAP) upon validated registration. The RD-Connect GPAP allows authorized users to query, analyze, interpret, and tag data within a secure environment, although direct download of full datasets is not permitted. All raw and processed data files are available at the EGA under the Solve-RD study EGAS00001003851, with datasets EGAD00001009767, EGAD00001009768, EGAD00001009769, and EGAD00001009770. Additionally, all novel and expert-curated variants identified in this study (n=207) have been submitted to ClinVar, ensuring broader accessibility and usability of the genetic data.

Raw sequencing data and variant calls for the different types of genetic variants that we generated for the research described in chapter 4 have been deposited in the EGA, ensuring secure long-term storage and controlled access for authorized researchers. Data access is granted to authorized researchers and clinicians, with decisions made by a Data Access Committee (DAC) based on whether the request aligns with the provided consent. The EGA serves as a secure archive adhering to the FAIR principles (Findable, Accessible, Interoperable, and Reusable).

List of abbreviations

ACMG: American College of Medical Genetics and Genomics
AMP: Association for Molecular Pathology
API: Application Programming Interface
ASD: Autism Spectrum Disorder
BAM: Binary Alignment Map
CNV: Copy Number Variant
DAC: Data Access Committee
DITF: Data Interpretation Task Force
DNM: De Novo Mutation
ES: Exome Sequencing
EGA: European Genome-phenome Archive
ERN: European Reference Network
FASTQ: A text-based format for storing nucleotide sequences
GS: Genome Sequencing
GUI: Graphical User Interface
HPO: Human Phenotype Ontology
LIS3: Lissencephaly type 3
LRS: Long-Read Sequencing
MAF: Minor Allele Frequency
MEI: Mobile Element Insertion
MLPA: Multiplex Ligation-dependent Probe Amplification
MME: Matchmaker Exchange
MQ: Mapping Quality
MT-TL1: Mitochondrially Encoded tRNA Leucine 1
OGM: Optical Genome Mapping
OMIM: Online Mendelian Inheritance in Man
ORDO: Orphanet Rare Disease Ontology
PCR: Polymerase Chain Reaction
PIK3CA: Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha
QC: Quality Control
RNA: Ribonucleic Acid
SCA: Spinocerebellar Ataxia
SMA: Spinal Muscular Atrophy
SMRT: Single Molecule Real-Time
SNV: Single Nucleotide Variant
STR: Short Tandem Repeat
SV: Structural Variant

UTR: Untranslated Region
VAF: Variant Allele Frequency
VCF: Variant Call Format
VEP: Variant Effect Predictor
VUS: Variant of Uncertain Significance
WES: Whole Exome Sequencing
WGS: Whole Genome Sequencing

Curriculum vitae

Wouter Steyaert was born on November 21, 1982, in Ghent, Belgium. He completed his secondary education at Don Bosco College in Zwijnaarde, focusing on science and mathematics. Wouter earned his Bachelor's degree in Biomedical Laboratory Technology, graduating magna cum laude. For his thesis, he contributed to a project on stress resistance in genetically modified maize at the University of Girona in Spain, under the guidance of Professor Maria Pla.

With the advent of exome sequencing and related technologies, there was an increasing demand for professionals skilled in programming and large-scale data analysis. Subsequently, after his bachelor's degree, Wouter began his career as a bioinformatician at the Center for Medical Genetics in Ghent. During the years he worked there, he developed various data analysis pipelines for examining exome, genome, and transcriptome data that were utilized in both research and diagnostics. The research projects he contributed to during this period primarily focused on unraveling the molecular basis of several connective tissue disorders. Recognizing the importance of continuous education, Wouter decided to pursue a Master's program in bioinformatics. As this program was initially unavailable in Ghent, he completed the necessary bridging courses part-time at KU Leuven while working full-time. Two years later, the master's program was introduced in Ghent, allowing him to continue his studies there. In 2019, Wouter completed his Master's in Bioinformatics with magna cum laude. His thesis, supervised by Professors Lieven Clement, Bert Callewaert, and Wim Van Criekinge, focused on the statistical methodology for comparing genetic variantion between two groups of individuals.

After obtaining his master's degree, Wouter applied for a PhD position at Radboud University Medical Center in Nijmegen. He was accepted to work on the Solve-RD project, marking a significant milestone in his academic and professional journey. In 2021, he was a predoctoral semifinalist for the Charles J. Epstein Trainee Awards for Excellence in Human Genetics and in 2023, he won an early career award from the European Society of Human Genetics.

Wouter has co-authored several publications in scientific journals, contributing valuable insights to the field of human genetics. His work continues to impact the scientific community, reflecting his dedication and expertise in his ongoing research endeavors. His commitment to advancing human genetics is evident through his collaborative and innovative approach to research.

PhD portfolio

PhD portfolio of Wouter Steyaert

Department: **Department** of **Human Genetics**
PhD period: **01/05/2019 – 30/09/2023**
PhD Supervisor(s): **Prof. Han Brunner, Prof. Christian Gilissen and Prof. Alexander Hoischen**

Training activities	**Hours**
Courses	
RIMLS - Introduction course "In the lead of my PhD" (2019)	15.00
RU - Mindfulness Based Stress Reduction (2019)	45.00
RU - Projectmanagement voor Promovendi (2020)	45.00
Clinical Genomics and NGS (2021)	40.00
Radboudumc - Scientific integrity (2023)	20.00
Seminars	
Human Genetics GDG group meeting (2021)	14.00
Theme discussion Human Genetics (weekly, 2 oral presentations) (2023)	112.00
Conferences	
Solve-RD annual meeting 2020 (2020)	14.00
European Human Genetics Conference 2020 (2020)	35.00
Solve-RD annual meeting 2021* (2021)	14.00
European Human Genetics Conference 2021 (2021)	35.00
American Human Genetics Conference 2021* (2021)	42.00
Solve-RD annual meeting 2022* (2022)	14.00
European Human Genetics Conference 2022+ (2022)	35.00
Solve-RD final meeting 2023* (2023)	14.00
Dutch Society of Human Genetics meeting 2023* (2023)	14.00
European Human Genetics Conference 2023* (2023)	35.00
Other	
Radboudumc - General Radboudumc introduction for research personnel (2019)	9.00
RIMLS PhD retreat (2021)	21.00
Total	**573.00**

* oral presentation + poster presentation

A

List of publications

Ariane J A G Van Tongerloo, Hannah Verdin, **Wouter Steyaert**, Paul J Coucke, Sandra Janssens (2024). Accepting or declining preconception expanded carrier screening: An exploratory study with 407 couples. Journal of Genetic Counseling, n/a(n/a), .

Robin Wijngaard, German Demidov, Luke O'Gorman, Jordi Corominas-Galbany, Burcu Yaldiz, **Wouter Steyaert**, Elke de Boer, Lisenka E L M Vissers, Erik-Jan Kamsteeg, Rolph Pfundt, Hilde Swinkels, Amber den Ouden, Iris B A W te Paske, Richarda M de Voer, Laurence Faivre, Anne-Sophie Denommé-Pichon, Yannis Duffourd, Antonio Vitobello, Martin Chevarin, Volker Straub, Ana Töpf, Anneke J van der Kooi, Francesca Magrinelli, Clarissa Rocca, Michael G Hanna, Jana Vandrovcova, Stephan Ossowski, Steven Laurie, Christian Gilissen, Solve-RD consortium (2024). Mobile element insertions in rare diseases: a comparative benchmark and reanalysis of 60,000 exome samples. European Journal of Human Genetics, 32(2), 200-208.

Adam Jackson, Sheng-Jia Lin, Elizabeth A Jones, Kate E Chandler, David Orr, Celia Moss, Zahra Haider, Gavin Ryan, Simon Holden, Mike Harrison, Nigel Burrows, Wendy D Jones, Mary Loveless, Cassidy Petree, Helen Stewart, Karen Low, Deirdre Donnelly, Simon Lovell, Konstantina Drosou, J C Ambrose, P Arumugam, R Bevers, M Bleda, F Boardman-Pretty, C R Boustred, H Brittain, M A Brown, M J Caulfield, G C Chan, A Giess, J N Griffin, A Hamblin, S Henderson, T J P Hubbard, R Jackson, L J Jones, D Kasperaviciute, M Kayikci, A Kousathanas, L Lahnstein, A Lakey, S E A Leigh, I U S Leong, F J Lopez, F Maleady-Crowe, M McEntagart, F Minneci, J Mitchell, L Moutsianas, M Mueller, N Murugaesu, A C Need, P O'Donovan, C A Odhams, C Patch, D Perez-Gil, M B Pereira, J Pullinger, T Rahim, A Rendon, T Rogers, K Savage, K Sawant, R H Scott, A Siddiq, A Sieghart, S C Smith, A Sosinsky, A Stuckey, M Tanguy, A L Taylor Tavares, E R A Thomas, S R Thompson, A Tucci, M J Welland, E Williams, K Witkowska, S M Wood, M Zarowiecki, Olaf Riess, Tobias B Haack, Holm Graessner, Birte Zurek, Kornelia Ellwanger, Stephan Ossowski, German Demidov, Marc Sturm, Julia M Schulze-Hentrich, Rebecca Schüle, Christoph Kessler, Melanie Wayand, Matthis Synofzik, Carlo Wilke, Andreas Traschütz, Ludger Schöls, Holger Hengel, Peter Heutink, Han Brunner, Hans Scheffer, Nicoline Hoogerbrugge, Alexander Hoischen, Peter A C 't Hoen, Lisenka E L M Vissers, Christian Gilissen, **Wouter Steyaert**, Karolis Sablauskas, Richarda M de Voer, Erik-Jan Kamsteeg, Bart van de Warrenburg, Nienke van Os, Iris te Paske, Erik Janssen, Elke de Boer, Marloes Steehouwer, Burcu Yaldiz, Tjitske Kleefstra, Anthony J Brookes, Colin Veal, Spencer Gibson, Marc Wadsley, Mehdi Mehtarizadeh, Umar Riaz, Greg Warren, Farid Yavari Dizjikan, Thomas Shorter, Ana Töpf, Volker Straub, Chiara Marini Bettolo, Sabine Specht, Jill Clayton-Smith,

Siddharth Banka, Elizabeth Alexander, Laurence Faivre, Christel Thauvin, Antonio Vitobello, Anne-Sophie Denommé-Pichon, Yannis Duffourd, Emilie Tisserant, Ange-Line Bruel, Christine Peyron, Aurore Pélissier, Sergi Beltran, Ivo Glynne Gut, Steven Laurie, Davide Piscia, Leslie Matalonga, Anastasios Papakonstantinou, Gemma Bullich, Alberto Corvo, Carles Garcia, Marcos Fernandez-Callejo, Carles Hernández, Daniel Picó, Ida Paramonov, Hanns Lochmüller, Gulcin Gumus, Virginie Bros-Facer, Ana Rath, Marc Hanauer, Annie Olry, David Lagorce, Svitlana Havrylenko, Katia Izem, Fanny Rigour, Giovanni Stevanin, Alexandra Durr, Claire-Sophie Davoine, Léna Guillot-Noel, Anna Heinzmann, Giulia Coarelli, Gisèle Bonne, Teresinha Evangelista, Valérie Allamand, Isabelle Nelson, Rabah Ben Yaou, Corinne Metay, Bruno Eymard, Enzo Cohen, Antonio Atalaia, Tanya Stojkovic, Milan Macek, Marek Turnovec, Dana Thomasová, Radka Pourová Kremliková, Vera Franková, Markéta Havlovicová, Vlastimil Kremlik, Helen Parkinson, Thomas Keane, Dylan Spalding, Alexander Senf, Peter Robinson, Daniel Danis, Glenn Robert, Alessia Costa, Christine Patch, Mike Hanna, Henry Houlden, Mary Reilly, Jana Vandrovcova, Francesco Muntoni, Irina Zaharieva, Anna Sarkozy, Vincent Timmerman, Jonathan Baets, Liedewei Van de Vondel, Danique Beijer, Peter de Jonghe, Vincenzo Nigro, Sandro Banfi, Annalaura Torella, Francesco Musacchia, Giulio Piluso, Alessandra Ferlini, Rita Selvatici, Rachele Rossi, Marcella Neri, Stefan Aretz, Isabel Spier, Anna Katharina Sommer, Sophia Peters, Carla Oliveira, Jose Garcia Pelaez, Ana Rita Matos, Celina São José, Marta Ferreira, Irene Gullo, Susana Fernandes, Luzia Garrido, Pedro Ferreira, Fátima Carneiro, Morris A Swertz, Lennart Johansson, Joeri K van der Velde, Gerben van der Vries, Pieter B Neerincx, Dieuwke Roelofs-Prins, Sebastian Köhler, Alison Metcalfe, Alain Verloes, Séverine Drunat, Caroline Rooryck, Aurelien Trimouille, Raffaele Castello, Manuela Morleo, Michele Pinelli, Alessandra Varavallo, Manuel Posada De la Paz, Eva Bermejo Sánchez, Estrella López Martín, Beatriz Martínez Delgado, F Javier Alonso García de la Rosa, Andrea Ciolfi, Bruno Dallapiccola, Simone Pizzi, Francesca Clementina Radio, Marco Tartaglia, Alessandra Renieri, Elisa Benetti, Peter Balicza, Maria Judit Molnar, Ales Maver, Borut Peterlin, Alexander Münchau, Katja Lohmann, Rebecca Herzog, Martje Pauly, Alfons Macaya, Anna Marcé-Grau, Andres Nascimiento Osorio, Daniel Natera de Benito, Rachel Thompson, Kiran Polavarapu, David Beeson, Judith Cossins, Pedro M Rodriguez Cruz, Peter Hackman, Mridul Johari, Marco Savarese, Bjarne Udd, Rita Horvath, Gabriel Capella, Laura Valle, Elke Holinski-Feder, Andreas Laner, Verena Steinke-Lange, Evelin Schröck, Andreas Rump, Gaurav K Varshney (2023). Clinical, genetic, epidemiologic, evolutionary, and functional delineation of TSPEAR-related autosomal recessive ectodermal dysplasia 14. Human Genetics and Genomics Advances, 4(2), 100186-100186.

Anne-Sophie Denommé-Pichon, Leslie Matalonga, Elke de Boer, Adam Jackson, Elisa Benetti, Siddharth Banka, Ange-Line Bruel, Andrea Ciolfi, Jill Clayton-Smith, Bruno Dallapiccola, Yannis Duffourd, Kornelia Ellwanger, Chiara Fallerini, Christian Gilissen, Holm Graessner, Tobias B Haack, Marketa Havlovicova, Alexander Hoischen, Nolwenn Jean-Marçais, Tjitske Kleefstra, Estrella López-Martín, Milan Macek, Maria Antonietta Mencarelli, Sébastien Moutton, Rolph Pfundt, Simone Pizzi, Manuel Posada, Francesca Clementina Radio, Alessandra Renieri, Caroline Rooryck, Lukas Ryba, Hana Safraou, Martin Schwarz, Marco Tartaglia, Christel Thauvin-Robinet, Julien Thevenon, Frédéric Tran Mau-Them, Aurélien Trimouille, Pavel Votypka, Bert B A de Vries, Marjolein H Willemsen, Birte Zurek, Alain Verloes, Christophe Philippe, Kristin M Abbott, Laurence Faivre, Mieke Kerstjens, Estrella López Martín, Isabelle Maystadt, Manuela Morleo, Vicenzo Nigro, Michele Pinelli, Francesca C Radio, Olaf Riess, Jean-Madeleine de Sainte Agathe, Gijs W E Santen, Christel Thauvin, Annalaura Torella, Lisenka Vissers, Antonio Vitobello, Kristina Zguro, Enzo Cohen, Daniel Danis, Fei Gao, Rita Horvath, Mridul Johari, Lennart Johanson, Shuang Li, Heba Morsy, Isabelle Nelson, Ida Paramonov, Iris B A W te Paske, Peter Robinson, Marco Savarese, **Wouter Steyaert**, Ana Töpf, Joeri K van der Velde, Jana Vandrovcova, Stephan Ossowski, German Demidov, Marc Sturm, Julia M Schulze-Hentrich, Rebecca Schüle, Jishu Xu, Christoph Kessler, Melanie Wayand, Matthis Synofzik, Carlo Wilke, Andreas Traschütz, Ludger Schöls, Holger Hengel, Holger Lerche, Josua Kegele, Peter Heutink, Han Brunner, Hans Scheffer, Nicoline Hoogerbrugge, Peter A C 't Hoen, Lisenka E L M Vissers, Karolis Sablauskas, Richarda M de Voer, Erik-Jan Kamsteeg, Bart van de Warrenburg, Nienke van Os, Iris te Paske, Erik Janssen, Marloes Steehouwer, Burcu Yaldiz, Anthony J Brookes, Colin Veal, Spencer Gibson, Vatsalya Maddi, Mehdi Mehtarizadeh, Umar Riaz, Greg Warren, Farid Yavari Dizjikan, Thomas Shorter, Volker Straub, Chiara Marini Bettolo, Jordi Diaz Manera, Sophie Hambleton, Karin Engelhardt, Elizabeth Alexander, Christine Peyron, Aurore Pélissier, Sergi Beltran, Ivo Glynne Gut, Steven Laurie, Davide Piscia, Anastasios Papakonstantinou, Gemma Bullich, Alberto Corvo, Marcos Fernandez-Callejo, Carles Hernández, Daniel Picó, Hanns Lochmüller, Gulcin Gumus, Virginie Bros-Facer, Ana Rath, Marc Hanauer, David Lagorce, Oscar Hongnat, Maroua Chahdil, Emeline Lebreton, Giovanni Stevanin, Alexandra Durr, Claire-Sophie Davoine, Léna Guillot-Noel, Anna Heinzmann, Giulia Coarelli, Gisèle Bonne, Teresinha Evangelista, Valérie Allamand, Rabah Ben Yaou, Corinne Metay, Bruno Eymard, Antonio Atalaia, Tanya Stojkovic, Marek Turnovec, Dana Thomasová, Radka Pourová Kremliková, Vera Franková, Markéta Havlovicová, Petra Lišková, Pavla Doležalová, Helen Parkinson, Thomas Keane, Mallory Freeberg, Coline Thomas, Dylan Spalding, Glenn Robert, Alessia Costa, Christine Patch, Mike Hanna, Henry Houlden, Mary Reilly, Stephanie Efthymiou, Elisa Cali, Francesca Magrinelli, Sanjay M Sisodiya, Jonathan Rohrer,

Francesco Muntoni, Irina Zaharieva, Anna Sarkozy, Vincent Timmerman, Jonathan Baets, Geert de Vries, Jonathan De Winter, Danique Beijer, Peter de Jonghe, Liedewei Van de Vondel, Willem De Ridder, Sarah Weckhuysen, Vincenzo Nigro, Margherita Mutarelli, Alessandra Varavallo, Sandro Banfi, Francesco Musacchia, Giulio Piluso, Alessandra Ferlini, Rita Selvatici, Francesca Gualandi, Stefania Bigoni, Rachele Rossi, Marcella Neri, Stefan Aretz, Isabel Spier, Anna Katharina Sommer, Sophia Peters, Carla Oliveira, Jose Garcia Pelaez, Ana Rita Matos, Celina São José, Marta Ferreira, Irene Gullo, Susana Fernandes, Luzia Garrido, Pedro Ferreira, Fátima Carneiro, Morris A Swertz, Lennart Johansson, Gerben van der Vries, Pieter B Neerincx, David Ruvolo, Wilhemina S Kerstjens Frederikse, Eveline Zonneveld-Huijssoon, Dieuwke Roelofs-Prins, Marielle van Gijn, Sebastian Köhler, Alison Metcalfe, Séverine Drunat, Delphine Heron, Cyril Mignot, Boris Keren, Didier Lacombe, Aurelien Trimouille, Gabriel Capella, Laura Valle, Elke Holinski-Feder, Andreas Laner, Verena Steinke-Lange, Maria-Roberta Cilio, Evelina Carpancea, Chantal Depondt, Damien Lederer, Yves Sznajer, Sarah Duerinckx, Sandrine Mary, Alfons Macaya, Ana Cazurro-Gutiérrez, Belén Pérez-Dueñas, Francina Munell, Clara Franco Jarava, Laura Batlle Masó, Anna Marcé-Grau, Roger Colobran, Peter Hackman, Bjarne Udd, Dimitri Hemelsoet, Bart Dermaut, Nika Schuermans, Bruce Poppe, Hannah Verdin, Andrés Nascimento Osorio, Christel Depienne, Andreas Roos, Isabell Cordts, Marcus Deschauer, Pasquale Striano, Federico Zara, Antonella Riva, Michele Iacomino, Paolo Uva, Marcello Scala, Paolo Scudieri, Ayşe Nazlı Başak, Kristl Claeys, Kaan Boztug, Matthias Haimel, Gijs W.E, Claudia A L Ruivenkamp, Daniel Natera de Benito, Rachel Thompson, Kiran Polavarapu, Bodo Grimbacher, Ioannis Zaganas, Evgenia Kokosali, Mathioudakis Lambros, Athanasios Evangeliou, Martha Spilioti, Elisabeth Kapaki, Mara Bourbouli, Peter Balicza, Maria Judit Molnar, Manuel Posada De la Paz, Eva Bermejo Sánchez, Beatriz Martínez Delgado, F Javier Alonso García de la Rosa, Evelin Schröck, Andreas Rump, Davide Mei, Annalisa Vetro, Simona Balestrini, Renzo Guerrini, Patrick F Chinnery, Thiloka Ratnaike, Katherine Schon, Ales Maver, Borut Peterlin, Alexander Münchau, Katja Lohmann, Rebecca Herzog, Martje Pauly, Patrick May, David Beeson, Judith Cossins, Simone Furini, Alexandra Afenjar, Alice Goldenberg, Alice Masurel, Alice Phan, Anne Dieux-Coeslier, Anne Fargeot, Anne-Marie Guerrot, Annick Toutain, Arnaud Molin, Arthur Sorlin, Audrey Putoux, Béatrice Jouret, Béatrice Laudier, Bénédicte Demeer, Bérénice Doray, Bertille Bonniaud, Bertrand Isidor, Brigitte Gilbert-Dussardier, Bruno Leheup, Bruno Reversade, Carle Paul, Catherine Vincent-Delorme, Cecilia Neiva, Céline Poirsier, Chloé Quélin, Christine Chiaverini, Christine Coubes, Christine Francannet, Cindy Colson, Claire Desplantes, Constance Wells, Cyril Goizet, Damien Sanlaville, Daniel Amram, Daphné Lehalle, David Geneviève, Dominique Gaillard, Einat Zivi, Elisabeth Sarrazin, Elisabeth Steichen, Élise Schaefer, Elodie Lacaze, Emmanuel Jacquemin,

Ernie Bongers, Esra Kilic, Estelle Colin, Fabienne Giuliano, Fabienne Prieur, Fanny Laffargue, Fanny Morice-Picard, Florence Petit, François Cartault, François Feillet, Geneviève Baujat, Gilles Morin, Gwenaëlle Diene, Hubert Journel, Isabelle Perthus, James Lespinasse, Jean-Luc Alessandri, Jeanne Amiel, Jelena Martinovic, Julian Delanne, Juliette Albuisson, Laëtitia Lambert, Laurence Perrin, Lilian Bomme Ousager, Lionel Van Maldergem, Lucile Pinson, Lyse Ruaud, Mahtab Samimi, Marie Bournez, Marie Noëlle Bonnet-Dupeyron, Marie Vincent, Marie-Line Jacquemont, Marie-Pierre Cordier-Alex, Marion Gérard-Blanluet, Marjolaine Willems, Marta Spodenkiewicz, Martine Doco-Fenzy, Massimiliano Rossi, Mathilde Renaud, Mélanie Fradin, Michèle Mathieu, Muriel H Holder-Espinasse, Nada Houcinat, Nadine Hanna, Nathalie Leperrier, Nicolas Chassaing, Nicole Philip, Odile Boute, Philippe Khau Van Kien, Philippe Parent, Pierre Bitoun, Pierre Sarda, Pierre Vabres, Pierre-Simon Jouk, Renaud Touraine, Salima El Chehadeh, Sandra Whalen, Sandrine Marlin, Sandrine Passemard, Sarah Grotto, Séverine Audebert Bellanger, Sophie Blesson, Sophie Nambot, Sophie Naudion, Stanislas Lyonnet, Sylvie Odent, Tania Attie-Bitach, Tiffany Busa, Valérie Drouin-Garraud, Valérie Layet, Varoona Bizaoui, Véronica Cusin, Yline Capri, Yves Alembik (2023). A Solve-RD ClinVar-based reanalysis of 1522 index cases from ERN-ITHACA reveals common pitfalls and misinterpretations in exome sequencing. Genetics in Medicine, 25(4), 100018-100018.

Burcu Yaldiz, Erdi Kucuk, Juliet Hampstead, Tom Hofste, Rolph Pfundt, Jordi Corominas Galbany, Tuula Rinne, Helger G Yntema, Alexander Hoischen, Marcel Nelen, Christian Gilissen, Olaf Riess, Tobias B Haack, Holm Graessner, Birte Zurek, Kornelia Ellwanger, Stephan Ossowski, German Demidov, Marc Sturm, Julia M Schulze-Hentrich, Rebecca Schüle, Jishu Xu, Christoph Kessler, Melanie Wayand, Matthis Synofzik, Carlo Wilke, Andreas Traschütz, Ludger Schöls, Holger Hengel, Holger Lerche, Josua Kegele, Peter Heutink, Han Brunner, Hans Scheffer, Nicoline Hoogerbrugge, Peter A C.'t Hoen, Lisenka E L M Vissers, **Wouter Steyaert**, Karolis Sablauskas, Richarda M de Voer, Erik-Jan Kamsteeg, Bart van de Warrenburg, Nienke van Os, Iris te Paske, Erik Janssen, Elke de Boer, Marloes Steehouwer, Tjitske Kleefstra, Anthony J Brookes, Colin Veal, Spencer Gibson, Vatsalya Maddi, Mehdi Mehtarizadeh, Umar Riaz, Greg Warren, Farid Yavari Dizjikan, Thomas Shorter, Ana Töpf, Volker Straub, Chiara Marini Bettolo, Jordi Diaz Manera, Sophie Hambleton, Karin Engelhardt, Jill Clayton-Smith, Siddharth Banka, Elizabeth Alexander, Adam Jackson, Laurence Faivre, Christel Thauvin, Antonio Vitobello, Anne-Sophie Denommé-Pichon, Yannis Duffourd, Ange-Line Bruel, Christine Peyron, Aurore Pélissier, Sergi Beltran, Ivo Glynne Gut, Steven Laurie, Davide Piscia, Leslie Matalonga, Anastasios Papakonstantinou, Gemma Bullich, Alberto Corvo, Marcos Fernandez-Callejo, Carles Hernández, Daniel Picó, Ida Paramonov, Hanns

Lochmüller, Gulcin Gumus, Virginie Bros-Facer, Ana Rath, Marc Hanauer, David Lagorce, Oscar Hongnat, Maroua Chahdil, Emeline Lebreton, Giovanni Stevanin, Alexandra Durr, Claire-Sophie Davoine, Léna Guillot-Noel, Anna Heinzmann, Giulia Coarelli, Gisèle Bonne, Teresinha Evangelista, Valérie Allamand, Isabelle Nelson, Rabah Ben Yaou, Corinne Metay, Bruno Eymard, Enzo Cohen, Antonio Atalaia, Tanya Stojkovic, Milan Macek, Marek Turnovec, Dana Thomasová, Radka Pourová Kremliková, Vera Franková, Markéta Havlovicová, Petra Lišková, Pavla Doležalová, Helen Parkinson, Thomas Keane, Mallory Freeberg, Coline Thomas, Dylan Spalding, Peter Robinson, Daniel Danis, Glenn Robert, Alessia Costa, Christine Patch, Mike Hanna, Henry Houlden, Mary Reilly, Jana Vandrovcova, Stephanie Efthymiou, Heba Morsy, Elisa Cali, Francesca Magrinelli, Sanjay M Sisodiya, Jonathan Rohrer, Francesco Muntoni, Irina Zaharieva, Anna Sarkozy, Vincent Timmerman, Jonathan Baets, Geert de Vries, Jonathan De Winter, Danique Beijer, Peter de Jonghe, Liedewei Van de Vondel, Willem De Ridder, Sarah Weckhuysen, Vincenzo Nigro, Margherita Mutarelli, Manuela Morleo, Michele Pinelli, Alessandra Varavallo, Sandro Banfi, Annalaura Torella, Francesco Musacchia, Giulio Piluso, Alessandra Ferlini, Rita Selvatici, Francesca Gualandi, Stefania Bigoni, Rachele Rossi, Marcella Neri, Stefan Aretz, Isabel Spier, Anna Katharina Sommer, Sophia Peters, Carla Oliveira, Jose Garcia Pelaez, Ana Rita Matos, Celina São José, Marta Ferreira, Irene Gullo, Susana Fernandes, Luzia Garrido, Pedro Ferreira, Fátima Carneiro, Morris A Swertz, Lennart Johansson, Joeri K van der Velde, Gerben van der Vries, Pieter B Neerincx, David Ruvolo, Kristin M Abbott, Wilhemina SKerstjens Frederikse, Eveline Zonneveld-Huijssoon, Dieuwke Roelofs-Prins, Marielle van Gijn, Sebastian Köhler, Alison Metcalfe, Alain Verloes, Séverine Drunat, Delphine Heron, Cyril Mignot, Boris Keren, Jean-Madeleine de Sainte Agathe, Caroline Rooryck, Didier Lacombe, Aurelien Trimouille, Manuel Posada De la Paz, Eva Bermejo Sánchez, Estrella López Martín, Beatriz Martínez Delgado, F Javier Alonso García de la Rosa, Andrea Ciolfi, Bruno Dallapiccola, Simone Pizzi, Francesca Clementina Radio, Marco Tartaglia, Alessandra Renieri, Simone Furini, Chiara Fallerini, Elisa Benetti, Peter Balicza, Maria Judit Molnar, Ales Maver, Borut Peterlin, Alexander Münchau, Katja Lohmann, Rebecca Herzog, Martje Pauly, Alfons Macaya, Ana Cazurro-Gutiérrez, Belén Pérez-Dueñas, Francina Munell, Clara Franco Jarava, Laura Batlle Masó, Anna Marcé-Grau, Roger Colobran, Andrés Nascimento Osorio, Daniel Natera de Benito, Rachel Thompson, Kiran Polavarapu, Bodo Grimbacher, David Beeson, Judith Cossins, Peter Hackman, Mridul Johari, Marco Savarese, Bjarne Udd, Rita Horvath, Patrick F Chinnery, Thiloka Ratnaike, Fei Gao, Katherine Schon, Gabriel Capella, Laura Valle, Elke Holinski-Feder, Andreas Laner, Verena Steinke-Lange, Evelin Schröck, Andreas Rump, Ayşe Nazlı Başak, Dimitri Hemelsoet, Bart Dermaut, Nika Schuermans, Bruce Poppe, Hannah Verdin, Davide Mei, Annalisa Vetro, Simona Balestrini, Renzo Guerrini, Kristl Claeys,

Gijs W E Santen, Emilia K Bijlsma, Mariette J V Hoffer, Claudia A L Ruivenkamp, Kaan Boztug, Matthias Haimel, Isabelle Maystadt, Isabelle Cordts, Marcus Deschauer, Ioannis Zaganas, Evgenia Kokosali, Mathioudakis Lambros, Athanasios Evangeliou, Martha Spilioti, Elisabeth Kapaki, Mara Bourbouli, Pasquale Striano, Federico Zara, Antonella Riva, Michele Iacomino, Paolo Uva, Marcello Scala, Paolo Scudieri, Maria-Roberta Cilio, Evelina Carpancea, Chantal Depondt, Damien Lederer, Yves Sznajer, Solve-RD consortium (2023). Twist exome capture allows for lower average sequence coverage in clinical exome sequencing. Human Genomics, 17(1), 39-39.

Wouter Steyaert, Lonneke Haer-Wigman, Rolph Pfundt, Debby Hellebrekers, Marloes Steehouwer, Juliet Hampstead, Elke de Boer, Alexander Stegmann, Helger Yntema, Erik-Jan Kamsteeg, Han Brunner, Alexander Hoischen, Christian Gilissen (2023). Systematic analysis of paralogous regions in 41,755 exomes uncovers clinically relevant variation. Nature Communications, 14(1), 6845-6845.

Anna K Sommer, Iris B A W te Paske, José Garcia-Pelaez, Andreas Laner, Elke Holinski-Feder, Verena Steinke-Lange, Sophia Peters, Laura Valle, Isabel Spier, David Huntsman, Gabriel Capella, Gareth Evans, Andreas Rump, Evelin Schröck, Alexander Hoischen, Nicoline Geverink, Marc Tischkowitz, Leslie Matalonga, Steven Laurie, Christian Gilissen, **Wouter Steyaert**, German Demidov, Carla Oliveira, Richarda M de Voer, Nicoline Hoogerbrugge, Stefan Aretz (2022). Solving the genetic aetiology of hereditary gastrointestinal tumour syndromes– a collaborative multicentre endeavour within the project Solve-RD. European Journal of Medical Genetics, 65(5), 104475-104475.

Elke de Boer, Burcu Yaldiz, Anne-Sophie Denommé-Pichon, Leslie Matalonga, Steve Laurie, **Wouter Steyaert**, Rick de Reuver, Christian Gilissen, Michael Kwint, Rolph Pfundt, Alain Verloes, Michèl A A P Willemsen, Bert B A de Vries, A Vitobello, Tjitske Kleefstra, Lisenka E L M Vissers, Enzo Cohen, Isabel Cuesta, Daniel Danis, Fei Gao, Rita Horvath, Mridul Johari, Lennart Johanson, Shuang Li, Heba Morsy, Isabelle Nelson, Ida Paramonov, Iris B A W te Paske, Peter Robinson, Marco Savarese, Ana Töpf, Aurélien Trimouille, Joeri K van der Velde, Jana Vandrovcova, Antonio Vitobello, Birte Zurek, Kristin M Abbot, Siddharth Banka, Elisa Benetti, Giorgio Casari, Andrea Ciolfi, Jill Clayton-Smith, Bruno Dallapiccola, Kornelia Ellwanger, Laurence Faivre, Holm Graessner, Tobias B Haack, Anna Hammarsjö, Marketa Havlovicova, Alexander Hoischen, Anne Hugon, Adam Jackson, Mieke Kerstjens, Anna Lindstrand, Estrella López Martín, Milan Macek, Isabelle Maystadt, Manuela Morleo, Vicenzo Nigro, Ann Nordgren, Maria Pettersson, Michele Pinelli, Simone Pizzi, Manuel Posada, Francesca C Radio, Alessandra Renieri, Caroline Rooryck, Lukas Ryba, Gijs W E Santen, Martin

Schwarz, Marco Tartaglia, Christel Thauvin, Annalaura Torella, Lisenka Vissers, Pavel Votypka, Klea Vyshka, Kristina Zguro (2022). Genome-wide variant calling in reanalysis of exome sequencing data uncovered a pathogenic TUBB3 variant. European Journal of Medical Genetics, 65(1), 104402-104402.

Wouter Steyaert, Matthew J Varney, Jeffrey L Benovic, John Creemers, Marijn M Speeckaert, Paul J Coucke, Joris R Delanghe (2022). Hypergastrinemia, a clue leading to the identification of an atypical form of diabetes mellitus type 2. Clinica Chimica Acta, 532(), 79-83.

Matthew J Varney, **Wouter Steyaert**, Paul J Coucke, Joris R Delanghe, David E Uehling, Babu Joseph, Richard Marcellus, Rima Al-awar, Jeffrey L Benovic (2022). G protein–coupled receptor kinase 6 (GRK6) regulates insulin processing and secretion via effects on proinsulin conversion to insulin. Journal of Biological Chemistry, 298(10), .

Gelana Khazeeva, Karolis Sablauskas, Bart van der Sanden, **Wouter Steyaert**, Michael Kwint, Dmitrijs Rots, Max Hinne, Marcel van Gerven, Helger Yntema, Lisenka Vissers, Christian Gilissen (2022). DeNovoCNN: a deep learning approach to de novo variant calling in next generation sequencing data. Nucleic Acids Research, 50(17), e97-e97.

Nika Schuermans, Dimitri Hemelsoet, Wim Terryn, Sanne Steyaert, Rudy Van Coster, Paul J Coucke, **Wouter Steyaert**, Bert Callewaert, Elke Bogaert, Patrick Verloo, Arnaud V Vanlander, Elke Debackere, Jody Ghijsels, Pontus LeBlanc, Hannah Verdin, Leslie Naesens, Filomeen Haerynck, Steven Callens, Bart Dermaut, Bruce Poppe, Jan De Bleecker, Patrick Santens, Paul Boon, Guy Laureys, Tessa Kerre, for UD-PrOZA (2022). Shortcutting the diagnostic odyssey: the multidisciplinary Program for Undiagnosed Rare Diseases in adults (UD-PrOZA). Orphanet Journal of Rare Diseases, 17(1), 210-210.

Wouter Steyaert, Shana Verschuere, Paul J Coucke, Olivier M Vanakker (2021). Comprehensive validation of a diagnostic strategy for sequencing genes with one or multiple pseudogenes using pseudoxanthoma elasticum as a model. Journal of Genetics and Genomics, 48(4), 289-299.

Birte Zurek, Kornelia Ellwanger, Lisenka E L M Vissers, Rebecca Schüle, Matthis Synofzik, Ana Töpf, Richarda M de Voer, Steven Laurie, Leslie Matalonga, Christian Gilissen, Stephan Ossowski, Peter A C 't Hoen, Antonio Vitobello, Julia M Schulze-

Hentrich, Olaf Riess, Han G Brunner, Anthony J Brookes, Ana Rath, Gisèle Bonne, Gulcin Gumus, Alain Verloes, Nicoline Hoogerbrugge, Teresinha Evangelista, Tina Harmuth, Morris Swertz, Dylan Spalding, Alexander Hoischen, Sergi Beltran, Holm Graessner, Tobias B Haack, German Demidov, Marc Sturm, Christoph Kessler, Melanie Wayand, Carlo Wilke, Andreas Traschütz, Ludger Schöls, Holger Hengel, Peter Heutink, Han Brunner, Hans Scheffer, **Wouter Steyaert**, Karolis Sablauskas, Erik-Jan Kamsteeg, Bart van de Warrenburg, Nienke van Os, Iris te Paske, Erik Janssen, Elke de Boer, Marloes Steehouwer, Burcu Yaldiz, Tjitske Kleefstra, Colin Veal, Spencer Gibson, Marc Wadsley, Mehdi Mehtarizadeh, Umar Riaz, Greg Warren, Farid Yavari Dizjikan, Thomas Shorter, Volker Straub, Chiara Marini Bettolo, Sabine Specht, Jill Clayton-Smith, Siddharth Banka, Elizabeth Alexander, Adam Jackson, Laurence Faivre, Christel Thauvin, Anne-Sophie Denommé-Pichon, Yannis Duffourd, Emilie Tisserant, Ange-Line Bruel, Christine Peyron, Aurore Pélissier, Ivo Glynne Gut, Davide Piscia, Anastasios Papakonstantinou, Gemma Bullich, Alberto Corvo, Carles Garcia, Marcos Fernandez-Callejo, Carles Hernández, Daniel Picó, Ida Paramonov, Hanns Lochmüller, Virginie Bros-Facer, Marc Hanauer, Annie Olry, David Lagorce, Svitlana Havrylenko, Katia Izem, Fanny Rigour, Giovanni Stevanin, Alexandra Durr, Claire-Sophie Davoine, Léna Guillot-Noel, Anna Heinzmann, Giulia Coarelli, Valérie Allamand, Isabelle Nelson, Rabah Ben Yaou, Corinne Metay, Bruno Eymard, Enzo Cohen, Antonio Atalaia, Tanya Stojkovic, Milan Macek, Marek Turnovec, Dana Thomasová, Radka Pourová Kremliková, Vera Franková, Markéta Havlovicová, Vlastimil Kremlik, Helen Parkinson, Thomas Keane, Alexander Senf, Peter Robinson, Daniel Danis, Glenn Robert, Alessia Costa, Christine Patch, Mike Hanna, Henry Houlden, Mary Reilly, Jana Vandrovcova, Francesco Muntoni, Irina Zaharieva, Anna Sarkozy, Vincent Timmerman, Jonathan Baets, Liedewei Van de Vondel, Danique Beijer, Peter de Jonghe, Vincenzo Nigro, Sandro Banfi, Annalaura Torella, Francesco Musacchia, Giulio Piluso, Alessandra Ferlini, Rita Selvatici, Rachele Rossi, Marcella Neri, Stefan Aretz, Isabel Spier, Anna Katharina Sommer, Sophia Peters, Carla Oliveira, Jose Garcia Pelaez, Ana Rita Matos, Celina São José, Marta Ferreira, Irene Gullo, Susana Fernandes, Luzia Garrido, Pedro Ferreira, Fátima Carneiro, Morris A Swertz, Lennart Johansson, Joeri K van der Velde, Gerben van der Vries, Pieter B Neerincx, Dieuwke Roelofs-Prins, Sebastian Köhler, Alison Metcalfe, Séverine Drunat, Caroline Rooryck, Aurelien Trimouille, Raffaele Castello, Manuela Morleo, Michele Pinelli, Alessandra Varavallo, Manuel Posada De la Paz, Eva Bermejo Sánchez, Estrella López Martín, Beatriz Martínez Delgado, F Javier Alonso García de la Rosa, Andrea Ciolfi, Bruno Dallapiccola, Simone Pizzi, Francesca Clementina Radio, Marco Tartaglia, Alessandra Renieri, Elisa Benetti, Peter Balicza, Maria Judit Molnar, Ales Maver, Borut Peterlin, Alexander Münchau, Katja Lohmann, Rebecca Herzog, Martje Pauly, Alfons Macaya, Anna Marcé-Grau, Andres Nascimiento Osorio,

Daniel Natera de Benito, Rachel Thompson, Kiran Polavarapu, David Beeson, Judith Cossins, Pedro M Rodriguez Cruz, Peter Hackman, Mridul Johari, Marco Savarese, Bjarne Udd, Rita Horvath, Gabriel Capella, Laura Valle, Elke Holinski-Feder, Andreas Laner, Verena Steinke-Lange, Evelin Schröck, Andreas Rump, Solve-RD consortium (2021). Correction to: Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. European Journal of Human Genetics, 29(9), 1459-1461.

Rebecca Schüle, Dagmar Timmann, Corrie E Erasmus, Jennifer Reichbauer, Melanie Wayand, Jonathan Baets, Peter Balicza, Patrick Chinnery, Alexandra Dürr, Tobias Haack, Holger Hengel, Rita Horvath, Henry Houlden, Erik-Jan Kamsteeg, Christoph Kamsteeg, Katja Lohmann, Alfons Macaya, Anna Marcé-Grau, Ales Maver, Judit Molnar, Alexander Münchau, Borut Peterlin, Olaf Riess, Ludger Schöls, Giovanni Stevanin, Matthis Synofzik, Vincent Timmerman, Bart van de Warrenburg, Nienke van Os, Jana Vandrovcova, Carlo Wilke, Andrea Bevot, Stephan Zuchner, Sergi Beltran, Steven Laurie, Leslie Matalonga, Holm Graessner, Birte Zurek, Kornelia Ellwanger, Stephan Ossowski, German Demidov, Marc Sturm, Julia M Schulze-Hentrich, Peter Heutink, Han Brunner, Hans Scheffer, Nicoline Hoogerbrugge, Alexander Hoischen, Peter A C 't Hoen, Lisenka E L M Vissers, Christian Gilissen, **Wouter Steyaert**, Karolis Sablauskas, Richarda M de Voer, Erik Janssen, Elke de Boer, Marloes Steehouwer, Burcu Yaldiz, Tjitske Kleefstra, Anthony J Brookes, Colin Veal, Spencer Gibson, Marc Wadsley, Mehdi Mehtarizadeh, Umar Riaz, Greg Warren, Farid Yavari Dizjikan, Thomas Shorter, Ana Töpf, Volker Straub, Chiara Marini Bettolo, Sabine Specht, Jill Clayton-Smith, Siddharth Banka, Elizabeth Alexander, Adam Jackson, Laurence Faivre, Christel Thauvin, Antonio Vitobello, Anne-Sophie Denommé-Pichon, Yannis Duffourd, Emilie Tisserant, Ange-Line Bruel, Christine Peyron, Aurore Pélissier, Ivo Glynne Gut, Davide Piscia, Anastasios Papakonstantinou, Gemma Bullich, Alberto Corvo, Carles Garcia, Marcos Fernandez-Callejo, Carles Hernández, Daniel Picó, Ida Paramonov, Hanns Lochmüller, Gulcin Gumus, Virginie Bros-Facer, Ana Rath, Marc Hanauer, Annie Olry, David Lagorce, Svitlana Havrylenko, Katia Izem, Fanny Rigour, Alexandra Durr, Claire-Sophie Davoine, Léna Guillot-Noel, Anna Heinzmann, Giulia Coarelli, Gisèle Bonne, Teresinha Evangelista, Valérie Allamand, Isabelle Nelson, Rabah Ben Yaou, Corinne Metay, Bruno Eymard, Enzo Cohen, Antonio Atalaia, Tanya Stojkovic, Milan Macek, Marek Turnovec, Dana Thomasová, Radka Pourová Kremliková, Vera Franková, Markéta Havlovicová, Vlastimil Kremlik, Helen Parkinson, Thomas Keane, Dylan Spalding, Alexander Senf, Peter Robinson, Daniel Danis, Glenn Robert, Alessia Costa, Christine Patch, Mike Hanna, Mary Reilly, Francesco Muntoni, Irina Zaharieva, Anna Sarkozy, Peter de Jonghe, Vincenzo Nigro, Sandro Banfi, Annalaura Torella, Francesco Musacchia, Giulio Piluso, Alessandra Ferlini, Rita

Selvatici, Rachele Rossi, Marcella Neri, Stefan Aretz, Isabel Spier, Anna Katharina Sommer, Sophia Peters, Carla Oliveira, Jose Garcia Pelaez, Ana Rita Matos, Celina São José, Marta Ferreira, Irene Gullo, Susana Fernandes, Luzia Garrido, Pedro Ferreira, Fátima Carneiro, Morris A Swertz, Lennart Johansson, Joeri K van der Velde, Gerben van der Vries, Pieter B Neerincx, Dieuwke Roelofs-Prins, Sebastian Köhler, Alison Metcalfe, Alain Verloes, Séverine Drunat, Caroline Rooryck, Aurelien Trimouille, Raffaele Castello, Manuela Morleo, Michele Pinelli, Alessandra Varavallo, Manuel Posada De la Paz, Eva Bermejo Sánchez, Estrella López Martín, Beatriz Martínez Delgado, F Javier Alonso García de la Rosa, Andrea Ciolfi, Bruno Dallapiccola, Simone Pizzi, Francesca Clementina Radio, Marco Tartaglia, Alessandra Renieri, Elisa Benetti, Maria Judit Molnar, Rebecca Herzog, Martje Pauly, Andres Nascimiento Osorio, Daniel Natera de Benito, Rachel Thompson, Kiran Polavarapu, David Beeson, Judith Cossins, Pedro M Rodriguez Cruz, Peter Hackman, Mridul Johari, Marco Savarese, Bjarne Udd, Gabriel Capella, Laura Valle, Elke Holinski-Feder, Andreas Laner, Verena Steinke-Lange, Evelin Schröck, Andreas Rump, Solve-RD-DITF-RND, The Solve-RD Consortium (2021). Solving unsolved rare neurological diseases—a Solve-RD viewpoint. European Journal of Human Genetics, 29(9), 1332-1336.

Elke de Boer, Charlotte W Ockeloen, Leslie Matalonga, Rita Horvath, Enzo Cohen, Isabel Cuesta, Daniel Danis, Anne-Sophie Denommé-Pichon, Yannis Duffourd, Christian Gilissen, Mridul Johari, Steven Laurie, Shuang Li, Isabelle Nelson, Sophia Peters, Ida Paramonov, Sivakumar Prasanth, Peter Robinson, Karolis Sablauskas, Marco Savarese, **Wouter Steyaert**, Ana Töpf, Joeri K van der Velde, Antonio Vitobello, Richard J Rodenburg, Marieke J H Coenen, Mirian Janssen, Dylan Henssen, Siddharth Banka, Elisa Benetti, Giorgio Casari, Andrea Ciolfi, Jill Clayton-Smith, Bruno Dallapiccola, Kornelia Ellwanger, Laurence Faivre, Holm Graessner, Tobias B Haack, Anna Hammarsjö, Marketa Havlovicova, Alexander Hoischen, Anne Hugon, Adam Jackson, Tjitske Kleefstra, Anna Lindstrand, Estrella López-Martín, Milan Macek, Manuela Morleo, Vicenzo Nigro, Ann Nordgren, Maria Pettersson, Michele Pinelli, Simone Pizzi, Manuel Posada, Francesca Clementina Radio, Alessandra Renieri, Caroline Rooryck, Lukas Ryba, Martin Schwarz, Marco Tartaglia, Christel Thauvin, Annalaura Torella, Alain Verloes, Lisenka Vissers, Pavel Votypka, Klea Vyshka, Birte Zurek, Aurélien Trimouille, Lisenka E L M Vissers, Solve-RD SNV-indel working group, Solve-RD-DITF-ITHACA (2021). A MT-TL1 variant identified by whole exome sequencing in an individual with intellectual disability, epilepsy, and spastic tetraparesis. European Journal of Human Genetics, 29(9), 1359-1368.

Birte Zurek, Kornelia Ellwanger, Lisenka E L M Vissers, Rebecca Schüle, Matthis Synofzik, Ana Töpf, Richarda M de Voer, Steven Laurie, Leslie Matalonga, Christian

Gilissen, Stephan Ossowski, Peter A C 't Hoen, Antonio Vitobello, Julia M Schulze-Hentrich, Olaf Riess, Han G Brunner, Anthony J Brookes, Ana Rath, Gisèle Bonne, Gulcin Gumus, Alain Verloes, Nicoline Hoogerbrugge, Teresinha Evangelista, Tina Harmuth, Morris Swertz, Dylan Spalding, Alexander Hoischen, Sergi Beltran, Holm Graessner, Tobias B Haack, German Demidov, Marc Sturm, Christoph Kessler, Melanie Wayand, Carlo Wilke, Andreas Traschütz, Ludger Schöls, Holger Hengel, Peter Heutink, Han Brunner, Hans Scheffer, **Wouter Steyaert**, Karolis Sablauskas, Erik-Jan Kamsteeg, Bart van de Warrenburg, Nienke van Os, Iris te Paske, Erik Janssen, Elke de Boer, Marloes Steehouwer, Burcu Yaldiz, Tjitske Kleefstra, Colin Veal, Spencer Gibson, Marc Wadsley, Mehdi Mehtarizadeh, Umar Riaz, Greg Warren, Farid Yavari Dizjikan, Thomas Shorter, Volker Straub, Chiara Marini Bettolo, Sabine Specht, Jill Clayton-Smith, Siddharth Banka, Elizabeth Alexander, Adam Jackson, Laurence Faivre, Christel Thauvin, Anne-Sophie Denommé-Pichon, Yannis Duffourd, Emilie Tisserant, Ange-Line Bruel, Christine Peyron, Aurore Pélissier, Ivo Glynne Gut, Davide Piscia, Anastasios Papakonstantinou, Gemma Bullich, Alberto Corvo, Carles Garcia, Marcos Fernandez-Callejo, Carles Hernández, Daniel Picó, Ida Paramonov, Hanns Lochmüller, Virginie Bros-Facer, Marc Hanauer, Annie Olry, David Lagorce, Svitlana Havrylenko, Katia Izem, Fanny Rigour, Giovanni Stevanin, Alexandra Durr, Claire-Sophie Davoine, Léna Guillot-Noel, Anna Heinzmann, Giulia Coarelli, Valérie Allamand, Isabelle Nelson, Rabah Ben Yaou, Corinne Metay, Bruno Eymard, Enzo Cohen, Antonio Atalaia, Tanya Stojkovic, Milan Macek, Marek Turnovec, Dana Thomasová, Radka Pourová Kremliková, Vera Franková, Markéta Havlovicová, Vlastimil Kremlik, Helen Parkinson, Thomas Keane, Alexander Senf, Peter Robinson, Daniel Danis, Glenn Robert, Alessia Costa, Christine Patch, Mike Hanna, Henry Houlden, Mary Reilly, Jana Vandrovcova, Francesco Muntoni, Irina Zaharieva, Anna Sarkozy, Vincent Timmerman, Jonathan Baets, Liedewei Van de Vondel, Danique Beijer, Peter de Jonghe, Vincenzo Nigro, Sandro Banfi, Annalaura Torella, Francesco Musacchia, Giulio Piluso, Alessandra Ferlini, Rita Selvatici, Rachele Rossi, Marcella Neri, Stefan Aretz, Isabel Spier, Anna Katharina Sommer, Sophia Peters, Carla Oliveira, Jose Garcia Pelaez, Ana Rita Matos, Celina São José, Marta Ferreira, Irene Gullo, Susana Fernandes, Luzia Garrido, Pedro Ferreira, Fátima Carneiro, Morris A Swertz, Lennart Johansson, Joeri K van der Velde, Gerben van der Vries, Pieter B Neerincx, Dieuwke Roelofs-Prins, Sebastian Köhler, Alison Metcalfe, Séverine Drunat, Caroline Rooryck, Aurelien Trimouille, Raffaele Castello, Manuela Morleo, Michele Pinelli, Alessandra Varavallo, Manuel Posada De la Paz, Eva Bermejo Sánchez, Estrella López Martín, Beatriz Martínez Delgado, F Javier Alonso García de la Rosa, Andrea Ciolfi, Bruno Dallapiccola, Simone Pizzi, Francesca Clementina Radio, Marco Tartaglia, Alessandra Renieri, Elisa Benetti, Peter Balicza, Maria Judit Molnar, Ales Maver, Borut Peterlin, Alexander Münchau, Katja Lohmann, Rebecca

Herzog, Martje Pauly, Alfons Macaya, Anna Marcé-Grau, Andres Nascimiento Osorio, Daniel Natera de Benito, Rachel Thompson, Kiran Polavarapu, David Beeson, Judith Cossins, Pedro M Rodriguez Cruz, Peter Hackman, Mridul Johari, Marco Savarese, Bjarne Udd, Rita Horvath, Gabriel Capella, Laura Valle, Elke Holinski-Feder, Andreas Laner, Verena Steinke-Lange, Evelin Schröck, Andreas Rump, Solve-RD consortium (2021). Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. European Journal of Human Genetics, 29(9), 1325-1331.

Ana Töpf, Angela Pyle, Helen Griffin, Leslie Matalonga, Katherine Schon, Enzo Cohen, Isabel Cuesta, Daniel Danis, Anne-Sophie Denommé-Pichon, Yannis Duffourd, Christian Gilissen, Mridul Johari, Steven Laurie, Shuang Li, Isabelle Nelson, Ida Paramonov, Sophia Peters, Sivakumar Prasanth, Peter Robinson, Karolis Sablauskas, Marco Savarese, **Wouter Steyaert**, Joeri K van der Velde, Antonio Vitobello, Jonathan Baets, Danique Beijer, Gisèle Bonne, Judith Cossins, Teresinha Evangelista, Alessandra Ferlini, Peter Hackman, Michael G Hanna, Henry Houlden, Jarred Lau, Hanns Lochmüller, William L Macken, Francesco Musacchia, Andres Nascimento, Daniel Natera-de Benito, Vincenzo Nigro, Giulio Piluso, Veronica Pini, Robert D S Pitceathly, Kiran Polavarapu, Pedro M Rodriguez Cruz, Anna Sarkozy, Rita Selvatici, Rachel Thompson, Annalaura Torella, Bjarne Udd, Liedewei Van de Vondel, Jana Vandrovcova, Irina Zaharieva, Albert Sickmann, Ulrike Schara–Schmidt, Andreas Hentschel, Patrick F Chinnery, Heike Kölbel, Andreas Roos, Rita Horvath, Solve-RD SNV-indel working group, Solve-RD DITF-euroNMD (2021). Exome reanalysis and proteomic profiling identified TRIP4 as a novel cause of cerebellar hypoplasia and spinal muscular atrophy (PCH1). European Journal of Human Genetics, 29(9), 1348-1353.

Leslie Matalonga, Carles Hernández-Ferrer, Davide Piscia, Enzo Cohen, Isabel Cuesta, Daniel Danis, Anne-Sophie Denommé-Pichon, Yannis Duffourd, Christian Gilissen, Mridul Johari, Steven Laurie, Shuang Li, Isabelle Nelson, Sophia Peters, Ida Paramonov, Sivakumar Prasanth, Peter Robinson, Karolis Sablauskas, Marco Savarese, **Wouter Steyaert**, Joeri K van der Velde, Antonio Vitobello, Rebecca Schüle, Matthis Synofzik, Ana Töpf, Lisenka E L M Vissers, Richarda de Voer, Stefan Aretz, Gabriel Capella, Richarda M de Voer, Gareth Evans, Jose Garcia Pelaez, Elke Holinski-Feder, Nicoline Hoogerbrugge, Andreas Laner, Carla Oliveira, Andreas Rump, Evelin Schröck, Anna Katharina Sommer, Verena Steinke-Lange, Iris te Paske, Marc Tischkowitz, Laura Valle, Siddharth Banka, Elisa Benetti, Giorgio Casari, Andrea Ciolfi, Jill Clayton-Smith, Bruno Dallapiccola, Elke de Boer, Kornelia Ellwanger, Laurence Faivre, Holm Graessner, Tobias B Haack, Anna Hammarsjö, Marketa Havlovicova, Alexander Hoischen, Anne Hugon, Adam Jackson, Tjitske Kleefstra,

Anna Lindstrand, Estrella López-Martín, Milan Macek, Manuela Morleo, Vicenzo Nigro, Ann Nordgren, Maria Pettersson, Michele Pinelli, Simone Pizzi, Manuel Posada, Francesca Clementina Radio, Alessandra Renieri, Caroline Rooryck, Lukas Ryba, Martin Schwarz, Marco Tartaglia, Christel Thauvin, Annalaura Torella, Aurélien Trimouille, Alain Verloes, Lisenka Vissers, Pavel Votypka, Klea Vyshka, Birte Zurek, Jonathan Baets, Danique Beijer, Gisèle Bonne, Judith Cossins, Teresinha Evangelista, Alessandra Ferlini, Peter Hackman, Michael G Hanna, Rita Horvath, Henry Houlden, Jarred Lau, Hanns Lochmüller, William L Macken, Francesco Musacchia, Andres Nascimento, Daniel Natera-de Benito, Vincenzo Nigro, Giulio Piluso, Veronica Pini, Robert D S Pitceathly, Kiran Polavarapu, Pedro M Rodriguez Cruz, Anna Sarkozy, Rita Selvatici, Rachel Thompson, Bjarne Udd, Liedewei Van de Vondel, Jana Vandrovcova, Irina Zaharieva, Peter Balicza, Patrick Chinnery, Alexandra Dürr, Tobias Haack, Holger Hengel, Erik-Jan Kamsteeg, Christoph Kamsteeg, Katja Lohmann, Alfons Macaya, Anna Marcé-Grau, Ales Maver, Judit Molnar, Alexander Münchau, Borut Peterlin, Olaf Riess, Ludger Schöls, Rebecca Schüle-Freyer, Giovanni Stevanin, Vincent Timmerman, Bart van de Warrenburg, Nienke van Os, Melanie Wayand, Carlo Wilke, Raul Tonda, Marcos Fernandez-Callejo, Daniel Picó, Carles Garcia-Linares, Anastasios Papakonstantinou, Alberto Corvó, Ricky Joshi, Hector Diez, Ivo Gut, Sergi Beltran, Stephan Ossowski, German Demidov, Marc Sturm, Julia M Schulze-Hentrich, Christoph Kessler, Peter Heutink, Han Brunner, Hans Scheffer, Peter A C 't Hoen, Erik Janssen, Marloes Steehouwer, Burcu Yaldiz, Anthony J Brookes, Colin Veal, Spencer Gibson, Marc Wadsley, Mehdi Mehtarizadeh, Umar Riaz, Greg Warren, Farid Yavari Dizjikan, Thomas Shorter, Volker Straub, Chiara Marini Bettolo, Sabine Specht, Elizabeth Alexander, Emilie Tisserant, Ange-Line Bruel, Christine Peyron, Aurore Pélissier, Ivo Glynne Gut, Gemma Bullich, Alberto Corvo, Carles Garcia, Carles Hernández, Gulcin Gumus, Virginie Bros-Facer, Ana Rath, Marc Hanauer, Annie Olry, David Lagorce, Svitlana Havrylenko, Katia Izem, Fanny Rigour, Alexandra Durr, Claire-Sophie Davoine, Léna Guillot-Noel, Anna Heinzmann, Giulia Coarelli, Valérie Allamand, Rabah Ben Yaou, Corinne Metay, Bruno Eymard, Antonio Atalaia, Tanya Stojkovic, Marek Turnovec, Dana Thomasová, Radka Pourová Kremliková, Vera Franková, Markéta Havlovicová, Vlastimil Kremlik, Helen Parkinson, Thomas Keane, Dylan Spalding, Alexander Senf, Glenn Robert, Alessia Costa, Christine Patch, Mike Hanna, Mary Reilly, Francesco Muntoni, Peter de Jonghe, Sandro Banfi, Rachele Rossi, Marcella Neri, Isabel Spier, Ana Rita Matos, Celina São José, Marta Ferreira, Irene Gullo, Susana Fernandes, Luzia Garrido, Pedro Ferreira, Fátima Carneiro, Morris A Swertz, Lennart Johansson, Gerben van der Vries, Pieter B Neerincx, Solve-RD SNV-indel working group, Solve-RD DITF-GENTURIS, Solve-RD DITF-ITHACA, Solve-RD DITF-euroNMD, Solve-RD DITF-RND, the Solve-RD Consortia (2021). Solving

patients with rare diseases through programmatic reanalysis of genome-phenome data. European Journal of Human Genetics, 29(9), 1337-1347.

Ilse Meerschaut, Shana De Coninck, **Wouter Steyaert**, Angela Barnicoat, Allan Bayat, Francesco Benedicenti, Siren Berland, Edward M Blair, Jeroen Breckpot, Anna de Burca, Anne Destrée, Sixto García-Miñaúr, Andrew J Green, Bernadette C Hanna, Kathelijn Keymolen, Marije Koopmans, Damien Lederer, Melissa Lees, Cheryl Longman, Sally Ann Lynch, Alison M Male, Fiona McKenzie, Isabelle Migeotte, Ercan Mihci, Banu Nur, Florence Petit, Juliette Piard, Frank S Plasschaert, Anita Rauch, Pascale Ribaï, Iratxe Salcedo Pacheco, Franco Stanzial, Irene Stolte-Dijkstra, Irene Valenzuela, Vinod Varghese, Pradeep C Vasudevan, Emma Wakeling, Carina Wallgren-Pettersson, Paul Coucke, Anne De Paepe, Daniël De Wolf, Sofie Symoens, Bert Callewaert (2020). A clinical scoring system for congenital contractural arachnodactyly. Genetics in Medicine, 22(1), 124-131.

Scott Barish, Tahsin Stefan Barakat, Brittany C Michel, Nazar Mashtalir, Jennifer B Phillips, Alfredo M Valencia, Berrak Ugur, Jeremy Wegner, Tiana M Scott, Brett Bostwick, David R Murdock, Hongzheng Dai, Elena Perenthaler, Anita Nikoncuk, Marjon van Slegtenhorst, Alice S Brooks, Boris Keren, Caroline Nava, Cyril Mignot, Jessica Douglas, Lance Rodan, Catherine Nowak, Sian Ellard, Karen Stals, Sally Ann Lynch, Marie Faoucher, Gaetan Lesca, Patrick Edery, Kendra L Engleman, Dihong Zhou, Isabelle Thiffault, John Herriges, Jennifer Gass, Raymond J Louie, Elliot Stolerman, Camerun Washington, Francesco Vetrini, Aiko Otsubo, Victoria M Pratt, Erin Conboy, Kayla Treat, Nora Shannon, Jose Camacho, Emma Wakeling, Bo Yuan, Chun-An Chen, Jill A Rosenfeld, Monte Westerfield, Michael Wangler, Shinya Yamamoto, Cigall Kadoch, Daryl A Scott, Hugo J Bellen (2020). BICRA, a SWI/SNF Complex Member, Is Associated with BAF-Disorder Related Phenotypes in Humans and Model Organisms. The American Journal of Human Genetics, 107(6), 1096-1112.

Ilse Meerschaut, Aude Beyens, **Wouter Steyaert**, Riet De Rycke, Katrien Bonte, Tine De Backer, Sandra Janssens, Joseph Panzer, Frank Plasschaert, Daniël De Wolf, Bert Callewaert (2019). Myhre syndrome: A first familial recurrence and broadening of the phenotypic spectrum. American Journal of Medical Genetics Part A, 179(12), 2494-2499.

Wouter Steyaert, Steven Callens, Paul Coucke, Bart Dermaut, Dimitri Hemelsoet, Wim Terryn, Bruce Poppe (2018). Future perspectives of genome-scale sequencing. Acta Clinica Belgica, 73(1), 7-10.

Paul C Marcogliese, Vandana Shashi, Rebecca C Spillmann, Nicholas Stong, Jill A Rosenfeld, Mary Kay Koenig, Julián A Martínez-Agosto, Matthew Herzog, Agnes H Chen, Patricia I Dickson, Henry J Lin, Moin U Vera, Noriko Salamon, John M Graham, Damara Ortiz, Elena Infante, **Wouter Steyaert**, Bart Dermaut, Bruce Poppe, Hyung-Lok Chung, Zhongyuan Zuo, Pei-Tseng Lee, Oguz Kanca, Fan Xia, Yaping Yang, Edward C Smith, Joan Jasien, Sujay Kansagra, Gail Spiridigliozzi, Mays El-Dairi, Robert Lark, Kacie Riley, Dwight D Koeberl, Katie Golden-Grant, Steven Callens, Paul Coucke, Dimitri Hemelsoet, Wim Terryn, Rudy Van Coster, David R Adams, Mercedes E Alejandro, Patrick Allard, Mahshid S Azamian, Carlos A Bacino, Ashok Balasubramanyam, Hayk Barseghyan, Gabriel F Batzli, Alan H Beggs, Babak Behnam, Anna Bican, David P Bick, Camille L Birch, Devon Bonner, Braden E Boone, Bret L Bostwick, Lauren C Briere, Donna M Brown, Matthew Brush, Elizabeth A Burke, Lindsay C Burrage, Shan Chen, Gary D Clark, Terra R Coakley, Joy D Cogan, Cynthia M Cooper, Heidi Cope, William J Craigen, Precilla D'Souza, Mariska Davids, Jyoti G Dayal, Esteban C Dell'Angelica, Shweta U Dhar, Ani Dillon, Katrina M Dipple, Laurel A Donnell-Fink, Naghmeh Dorrani, Daniel C Dorset, Emilie D Douine, David D Draper, David J Eckstein, Lisa T Emrick, Christine M Eng, Ascia Eskin, Cecilia Esteves, Tyra Estwick, Carlos Ferreira, Brent L Fogel, Noah D Friedman, William A Gahl, Emily Glanton, Rena A Godfrey, David B Goldstein, Sarah E Gould, Jean-Philippe F Gourdine, Catherine A Groden, Andrea L Gropman, Melissa Haendel, Rizwan Hamid, Neil A Hanchard, Lori H Handley, Matthew R Herzog, Ingrid A Holm, Jason Hom, Ellen M Howerton, Yong Huang, Howard J Jacob, Mahim Jain, Yong-hui Jiang, Jean M Johnston, Angela L Jones, Isaac S Kohane, Donna M Krasnewich, Elizabeth L Krieg, Joel B Krier, Seema R Lalani, C Christopher Lau, Jozef Lazar, Brendan H Lee, Hane Lee, Shawn E Levy, Richard A Lewis, Sharyn A Lincoln, Allen Lipson, Sandra K Loo, Joseph Loscalzo, Richard L Maas, Ellen F Macnamara, Calum A MacRae, Valerie V Maduro, Marta M Majcherska, May Christine V Malicdan, Laura A Mamounas, Teri A Manolio, Thomas C Markello, Ronit Marom, Julian A Martínez-Agosto, Shruti Marwaha, Thomas May, Allyn McConkie-Rosell, Colleen E McCormack, Alexa T McCray, Matthew Might, Paolo M Moretti, Marie Morimoto, John J Mulvihill, Jennifer L Murphy, Donna M Muzny, Michele E Nehrebecky, Stan F Nelson, J Scott Newberry, John H Newman, Sarah K Nicholas, Donna Novacic, Jordan S Orange, J Carl Pallais, Christina G S Palmer, Jeanette C Papp, Neil H Parker, Loren D M Pena, John A Phillips, Jennifer E Posey, John H Postlethwait, Lorraine Potocki, Barbara N Pusey, Chloe M Reuter, Amy K Robertson, Lance H Rodan, Jacinda B Sampson, Susan L Samson, Kelly Schoch, Molly C Schroeder, Daryl A Scott, Prashant Sharma, Rebecca Signer, Edwin K Silverman, Janet S Sinsheimer, Kevin S Smith, Kimberly Splinter, Joan M Stoler, Jennifer A Sullivan, David A Sweetser, Cynthia J Tifft, Camilo Toro, Alyssa A Tran, Tiina K Urv, Zaheer M Valivullah, Eric Vilain, Tiphanie P Vogel, Colleen E Wahl,

Nicole M Walley, Chris A Walsh, Patricia A Ward, Katrina M Waters, Monte Westerfield, Anastasia L Wise, Lynne A Wolfe, Elizabeth A Worthey, Shinya Yamamoto, Guoyun Yu, Diane B Zastrow, Allison Zheng, Michael F Wangler, Ghayda Mirzaa, Brendan Lee, Stanley F Nelson, Hugo J Bellen (2018). IRF2BPL Is Associated with Neurological Phenotypes. The American Journal of Human Genetics, 103(2), 245-260.

Annekatrien Boel, Hanna De Saffel, **Wouter Steyaert**, Bert Callewaert, Anne De Paepe, Paul J Coucke, Andy Willaert (2018). CRISPR/Cas9-mediated homology-directed repair by ssODNs in zebrafish induces complex mutational patterns resulting from genomic integration of repair-template fragments. Disease Models & Mechanisms, 11(10), dmm035352-dmm035352.

Laura Muiño-Mosquera, Felke Steijns, Tjorven Audenaert, Ilse Meerschaut, Anne De Paepe, **Wouter Steyaert**, Sofie Symoens, Paul Coucke, Bert Callewaert, Marjolijn Renard, Julie De Backer (2018). Tailoring the American College of Medical Genetics and Genomics and the Association for Molecular Pathology Guidelines for the Interpretation of Sequenced Variants in the FBN1 Gene for Marfan Syndrome. Circulation: Genomic and Precision Medicine, 11(6), e002039-e002039.

Wouter Steyaert, Annekatrien Boel, Paul Coucke, Andy Willaert (2018). BATCH-GE: Analysis of NGS Data for Genome Editing Assessment. Xenopus: Methods and Protocols, (), 83-90.

Tim Van Damme, Thatjana Gardeitchik, Miski Mohamed, Sergio Guerrero-Castillo, Peter Freisinger, Brecht Guillemyn, Ariana Kariminejad, Daisy Dalloyaux, Sanne van Kraaij, Dirk J Lefeber, Delfien Syx, **Wouter Steyaert**, Riet De Rycke, Alexander Hoischen, Erik-Jan Kamsteeg, Sunnie Y Wong, Monique van Scherpenzeel, Payman Jamali, Ulrich Brandt, Leo Nijtmans, G Christoph Korenke, Brian H Y Chung, Christopher C Y Mak, Ingrid Hausser, Uwe Kornak, Björn Fischer-Zirnsak, Tim M Strom, Thomas Meitinger, Yasemin Alanay, Gulen E Utine, Peter K C Leung, Siavash Ghaderi-Sohi, Paul Coucke, Sofie Symoens, Anne De Paepe, Christian Thiel, Tobias B Haack, Fransiska Malfait, Eva Morava, Bert Callewaert, Ron A Wevers (2017). Mutations in ATP6V1E1 or ATP6V1A Cause Autosomal-Recessive Cutis Laxa. The American Journal of Human Genetics, 100(2), 216-227.

Sofie Symoens, **Wouter Steyaert**, Lynn Demuynck, Anne De Paepe, Karin E M Diderich, Fransiska Malfait, Paul J Coucke (2017). Tissue-specific mosaicism for a lethal osteogenesis imperfecta COL1A1 mutation causes mild OI/EDS overlap syndrome. American Journal of Medical Genetics Part A, 173(4), 1047-1050.

Caroline Van Cauwenbergh, Kristof Van Schil, Robrecht Cannoodt, Miriam Bauwens, Thalia Van Laethem, Sarah De Jaegere, **Wouter Steyaert**, Tom Sante, Björn Menten, Bart P Leroy, Frauke Coppieters, Elfride De Baere (2017). arrEYE: a customized platform for high-resolution copy number analysis of coding and noncoding regions of known and candidate retinal dystrophy genes and retinal noncoding RNAs. Genetics in Medicine, 19(4), 457-466.

R H Ali, K Shah, A Nasir, W Steyaert, P J Coucke, W Ahmad (2016). Exome sequencing revealed a novel biallelic deletion in the DCAF17 gene underlying Woodhouse Sakati syndrome. Clinical Genetics, 90(3), 263-269.

Khadim Shah, Raja Hussain Ali, Muhammad Ansar, Kwanghyuk Lee, Muhammad Salman Chishti, Izoduwa Abbe, Biao Li, Joshua D Smith, Deborah A Nickerson, Jay Shendure, Paul J Coucke, **Wouter Steyaert**, Michael J Bamshad, Regie Lyn P Santos-Cortez, Suzanne M Leal, Wasim Ahmad, University of Washington Center for Mendelian Genomics (2016). Mitral regurgitation as a phenotypic manifestation of nonphotosensitive trichothiodystrophy due to a splice variant in MPLKIP. BMC Medical Genetics, 17(1), 13-13.

Thomas Naert, Robin Colpaert, Tom Van Nieuwenhuysen, Dionysia Dimitrakopoulou, Jannick Leoen, Jurgen Haustraete, Annekatrien Boel, **Wouter Steyaert**, Trees Lepez, Dieter Deforce, Andy Willaert, David Creytens, Kris Vleminckx (2016). CRISPR/Cas9 mediated knockout of rb1 and rbl1 leads to rapid and penetrant retinoblastoma development in Xenopus tropicalis. Scientific Reports, 6(1), 35264-35264.

Annekatrien Boel, Woutert Steyaert, Nina De Rocker, Björn Menten, Bert Callewaert, Anne De Paepe, Paul Coucke, Andy Willaert (2016). BATCH-GE: Batch analysis of Next-Generation Sequencing data for genome editing assessment. Scientific Reports, 6(1), 30330-30330.

Delfien Syx, Sofie Symoens, **Wouter Steyaert**, Anne De Paepe, Paul J Coucke, Fransiska Malfait (2015). Ehlers-Danlos Syndrome, Hypermobility Type, Is Linked to Chromosome 8p22-8p21.1 in an Extended Belgian Family. Disease Markers, 2015(), 828970-828970.

O Essawi, M Farraj, K De Leeneer, W Steyaert, K De Pauw, A De Paepe, K Claes, T Essawi, P J Coucke (2015). Next Generation Sequencing to Determine the Cystic Fibrosis Mutation Spectrum in Palestinian Population. Disease Markers, 2015(), 458653-458653.

Sofie Symoens, Aileen M. Barnes, Charlotte Gistelinck, Fransiska Malfait, Brecht Guillemyn, **Wouter Steyaert**, Delfien Syx, Sanne D'hondt, Martine Biervliet, Julie De Backer, Eckhard P. Witten, Sergey Leikin, Elena Makareeva, Gabriele Gillessen-Kaesbach, Ann Huysseune, Kris Vleminckx, Andy Willaert, Anne De Paepe, Joan C. Marini, Paul J. Coucke (2015). Genetic Defects in TAPT1 Disrupt Ciliogenesis and Cause a Complex Lethal Osteochondrodysplasia. The American Journal of Human Genetics, 97(4), 521-534.

Mohammad J Hosen, Filip Van Nieuwerburgh, **Wouter Steyaert**, Dieter Deforce, Ludovic Martin, Georges Leftheriotis, Anne De Paepe, Paul J Coucke, Olivier M Vanakker (2015). Efficiency of Exome Sequencing for the Molecular Diagnosis of Pseudoxanthoma Elasticum. Journal of Investigative Dermatology, 135(4), 992-998.

Kim De Leeneer, Jan Hellemans, **Wouter Steyaert**, Steve Lefever, Inge Vereecke, Eveline Debals, Brecht Crombez, Machteld Baetens, Mattias Van Heetvelde, Frauke Coppieters, Jo Vandesompele, Annelies De Jaegher, Elfride De Baere, Paul Coucke, Kathleen Claes (2015). Flexible, Scalable, and Efficient Targeted Resequencing on a Benchtop Sequencer for Variant Detection in Clinical Practice. Human Mutation, 36(3), 379-387.

Patrick Santens, Tim Van Damme, **Wouter Steyaert**, Andy Willaert, Bernard Sablonnière, Anne De Paepe, Paul J Coucke, Bart Dermaut (2015). RNF216 mutations as a novel cause of autosomal recessive Huntington-like disorder. Neurology, 84(17), 1760-1766.

Sofie Symoens, Fransiska Malfait, Sanne D'hondt, Bert Callewaert, Annelies Dheedene, **Wouter Steyaert**, Hans Peter Bächinger, Anne De Paepe, Hulya Kayserili, Paul J Coucke (2013). Deficiency for the ER-stress transducer OASIS causes severe recessive osteogenesis imperfecta in humans. Orphanet Journal of Rare Diseases, 8(1), 154-154.

A

Dankwoord

Hoewel ik tijdens mijn PhD-traject hard heb gewerkt, zou het onmogelijk zijn geweest om de verkregen resultaten te presenteren zonder de toewijding en steun van een zeer groot aantal andere mensen. Hen wil ik hier bedanken.

Vooreerst wil ik Han bedanken, promotor van mijn proefschrift. Zonder zijn steun, begrip en oplossingsgerichtheid zou deze thesis er nooit zijn gekomen. Han heeft me bijzonder geïnspireerd, gemotiveerd en mij vertrouwen gegeven. Uit de gesprekken die we hebben gehad heb ik bijzonder veel geleerd, wat ik zowel meeneem op mijn verder wetenschappelijke pad alsook in mijn persoonlijk leven. De rijkdom aan constructieve gedachten blijven me bij in mijn herinnering.

Mijn beide supervisors-promotors, Christian en Alex, hebben ook een enorme bijdrage geleverd aan dit werk. Waarin ik zeker ben geëvolueerd de voorbije jaren is het inzicht in de noodzaak om wetenschappelijke projecten tot het allerlaatste punt af te werken. Dat heb ik aan Christian te danken. Alex heeft een bijzonder grote hoeveelheid energie. Het is inspirerend om te zien hoeveel er kan worden opgebouwd door die energie aan het werk te zetten. Door nooit op te geven, door blijvend door te zetten komen zaken in beweging en zorg je voor verandering. Dat is motiverend.

Tijdens mijn PhD-traject heb ik wetenschap gepresenteerd op posters, in papers en tijdens wetenschappelijke praatjes op nationale en internationale conferenties. Dat dit telkens netjes en gestructureerd was, is in eerste plaats de verdienste van mijn drie promotoren. De feedbackmomenten over deze wetenschappelijke presentaties hebben me bijzonder geholpen om het verhaal strak te krijgen en de overbodige details achterwege te laten.

Voor heel veel van de data die is geanalyseerd als onderdeel van dit werk is de primaire analyse gebeurd door de collega's van het fantastische bio-informatica team aan het Radboudumc onder leiding van Christian en Steven. In het bijzonder hebben Jordi, Galuh en Luke een hele grote bijdrage geleverd aan de verwerking van de Solve-RD data. Ook wanneer ik een technische of organisatorische vraag had, was er altijd iemand die me vrijwel onmiddellijk ter hulp schoot. In Christians onderzoeksgroep wil ik uitdrukkelijk Gelana, Erdi, Brechtje, Juliet, Karolis, Laurens en Jakob bedanken. We hadden tal van inspirerende, leerrijke en ontspannende gesprekken. We lachten met elkanders grappen en boden een luisterend oor wanneer nodig.

Ik had ook het voorrecht om deel te mogen uitmaken van Alex zijn onderzoeksgroep. Ik ben dankbaar voor de samenwerkingen, de gesprekken en de warmte die hiervan uitging. Grote dank hiervoor aan Bart, Konny, Cas en Emil, en in het bijzonder ook aan Marloes en Nick die een hele grote bijdrage hebben geleverd aan de staal- en dataverwerking en de validatie van onze Solve-RD bevindingen. Wanneer ik een vraag had over een bepaald staal en trio, kwam er vrijwel ogenblikkelijk een antwoord. Deze ogenschijnlijke vanzelfsprekendheid heeft een bijzonder positieve invloed gehad op Solve-RD en dus ook op dit proefschrift. Special thanks also to Lydia for taking me under her wing in the creation of the long-read sequencing paper. Without the variant interpretations, excellent figures and guidance, the long-read paper would not be anything close to what it has ultimately become.

Naast de collega's aan het Radboudumc in Nijmegen wil ik zeker ook de onderzoekers bedanken waarmee we bijna op dagelijkse basis hebben samengewerkt binnen het Solve-RD project. In het bijzonder heeft Solve-RD veel te danken aan het bio-informatica werk geleverd door het team aan het Centro Nacional de Análisis Genómico (CNAG) in Barcelona. Steve, Leslie en Sergi maar ook alle andere bioinformatici en wetenschappers daar hebben ervoor gezorgd dat de NGS dataverwerking van topkwaliteit was. Ook onze collega's in Tübingen, in het bijzonder German en Stephan, hebben heel wat werk gedaan voor hoofdstuk drie van dit proefschrift. De identificatie van structurele varianten in short-read sequencing data werd voornamelijk door hen naar een hoger niveau getild.

Naast de vele nieuwe contacten en samenwerkingen die mijn PhD-traject me heeft geboden, zowel nationaal als internationaal, heb ik ook steeds contact gehouden met Paul, de supervisor die ik had in het Centrum voor Medische Genetica Gent (CMGG). Ik ben bijzonder dankbaar voor de blijvende steun, aanmoedigingen, advies en het vertrouwen. Ik ben dan ook erg blij om terug deel te mogen uitmaken van het CMGG.

Tenslotte, mijn ouders, familie en naaste vrienden zijn blijvend in mij geloven. Door hun voortdurende en onvoorwaardelijke steun is het gelukt om te blijven volharden en door te zetten in dit uitdagende traject.

Radboudumc university medical center

IN·DEI·NOMINE·FELICITER

Radboud University