Muhammet Erdi Küçük

Unraveling Rare Diseases:

Bioinformatic Evaluation of Novel Sequencing Technologies

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Unraveling Rare Diseases

Bioinformatic Evaluation of Novel Sequencing Technologies

Muhammet Erdi Küçük

Unraveling Rare Diseases: Bioinformatic Evaluation of Novel Sequencing Technologies
Muhammet Erdi Küçük

Radboud Dissertation 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 | Annelies Lips
Cover: Deniz Özyurda Ergen
Printing: DPN Rikken/Pumbo

ISBN: 9789465152301
DOI: 10.54195/9789465152301
Free download at: https://doi.org/10.54195/9789465152301

RADBOUD
UNIVERSITY
PRESS

Unraveling Rare Diseases

Bioinformatic Evaluation of Novel Sequencing Technologies

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 19 maart 2026

om 12.30 uur precies

door

Muhammet Erdi Küçük

Promotor:
prof. dr. C.F.H.A. Gilissen

Copromotor:
dr. J.E. Hampstead

Manuscriptcommissie:
Prof. dr. M.A. Huijnen
Dr. J.H.F. de Baaij
Prof. dr. E.A. Sistermans (Vrije Universiteit Amsterdam)

Unraveling Rare Diseases

Bioinformatic Evaluation of Novel Sequencing Technologies

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, March 19, 2026

at 12.30 pm

by

Muhammet Erdi Küçük

Supervisor:
prof. dr. C.F.H.A. Gilissen

Co-supervisor:
dr. J.E. Hampstead

Manuscript Committee:
Prof. dr. M.A. Huijnen
Dr. J.H.F. de Baaij
Prof. dr. E.A. Sistermans (Vrije Universiteit Amsterdam)

Table of contents

1. Introduction

1.1 Background

One of the primary aims of human genetics research is to characterize the relationship between genetic variation and biomedical traits, especially diseases. To date it is estimated that there are around 7,000 single-gene inherited disorders, which are caused by defects in one particular gene (Nguengang-Wakap et al., 2019). Virtually all of these are classified as rare diseases, which are defined by affecting less than 1 in 2000 individuals according to EU criteria (Vissers et al., 2016). Rare diseases collectively affect more than 300 million people globally and include many conditions with severe medical consequences, such as Huntington's disease, retinitis pigmentosa, and cystic fibrosis. Since most of these diseases are chronic, degenerative and life-threatening, they provide a substantial challenge to public health systems (Claussnitzer et al., 2020).

Among rare diseases, one of the most common symptoms is intellectual disability (ID), with a worldwide prevalence estimated at 1% (Bamshad et al., 2011). ID is characterized by serious impairment of mental faculties and social behavior, and is often identified early in life due to developmental delays. ID can occur in combination with other neurodevelopmental disorders, such as epilepsy and autism spectrum disorders, with varying levels of severity (Vasudevan & Suri, 2017). Causes of ID include infections, maternal alcohol abuse during pregnancy, complications at birth, and malnutrition, but up to half of syndromic ID cases are estimated to be genetic in nature. Historically, only about 25-35% of ID patients obtained a molecular diagnosis (Vissers et al., 2016), although recent studies with extensive analysis were able to diagnose up to 41% of participants (Wright et al., 2023).

1.2 Research and Diagnostics of Rare Diseases

Obtaining a molecular diagnosis is of critical importance for the care and treatment of rare disease patients. For instance, clinicians might not be able to predict the prognosis of a patient without the knowledge of the underlying genetic causes (Shendure et al., 2019). Identification of damaging mutations is also crucial for carrier screening and family planning (Ng et al., 2009).

1.2.1 Genetic Architecture of Rare Diseases

Both research and diagnostics rely on an understanding of the genetic basis of the disease in question. The underlying genetic cause of a rare disease can

vary greatly depending on the specific rare disease and the genetic mutations involved. Most rare diseases are caused by one or two mutations in a single gene, which may result in the production of an abnormal/reduced protein or altered gene expression that disrupts normal cellular processes. These conditions are referred to as monogenic or Mendelian disorders, and include examples such as cystic fibrosis, sickle cell anemia, and Huntington's disease. In other rare diseases, concurrent mutations in multiple genes or complex interactions between genetic and environmental factors can produce the phenotype. These are called polygenic or multifactorial disorders, and include conditions such as autism, schizophrenia, and diabetes (W. Li, 2023). For many disease phenotypes, pathogenic variation can arise from either mechanism (monogenic or polygenic), depending on the genes and variants involved.

Rare diseases are often distinguished based on their inheritance pattern. Some rare diseases follow an autosomal dominant pattern, meaning that a person needs only one copy of the mutated gene to develop the disease. This often arises through haploinsufficiency, a reduction in gene dosage in which one functional allele produces insufficient protein amount or activity to sustain the normal phenotype, so a single wild-type copy cannot compensate (Deutschbauer et al., 2005). Dominance can also result from dominant-negative effects, in which an abnormal protein interferes with the wild-type protein's activity (e.g., by forming non-functional multimers or sequestering binding partners) (Billant et al., 2016). In contrast, autosomal recessive conditions typically require biallelic pathogenic variants; the presence of one functional allele usually maintains adequate protein level or activity, so heterozygous individuals are clinically unaffected carriers. Disease manifests when both alleles are affected—either homozygous for the same variant or compound heterozygous for two different pathogenic variants—driving loss of function below a critical threshold (Yuan et al., 2022).

1.2.2 Rare disease research

In almost all cases, the molecular diagnosis of rare diseases is predicated on prior biomedical research. For instance, genetic association studies investigate the frequency and distribution of genetic variants in individuals with the disease compared to individuals without the disease. This approach includes family-based studies (comparing variants between affected and unaffected relatives) and population-based studies (comparing variants between patient cohorts and healthy controls). The goal of such case–control designs is to find statistically significant variant–disease associations, suggesting that a variant found more

often in patients may contribute to disease risk (Cutting, 2014). Other statistical studies use so-called de novo mutations from large patient cohorts, to assess whether a gene harbours more mutations than would be expected based on a mutational model (Samocha et al., 2014; Kaplanis et al., 2020).

Once a variant's involvement in a rare disease is established through research, that knowledge can be used for further research and in routine diagnostics. Functional characterization of newly identified mutations often increases our understanding of disease biology at the molecular level (Bustos et al., 2021; Gilissen et al., 2014; Przybyla & Gilbert, 2021) which is necessary for the development of treatments and therapies. In recent years, researchers have even designed targeted gene therapies for conditions such as spinal muscular atrophy (Mendell et al., 2017), childhood-onset blindness (Bennett et al., 2016) and adenosine deaminase deficiency (Aiuti et al., 2009) based on the initial identification of pathogenic variants and genes.

1.2.3 Challenges in Molecular Diagnosis of Rare Diseases

Molecular diagnosis of rare diseases remains a complex, expensive and lengthy process. One major challenge is the genetic heterogeneity of many rare diseases, especially ID and other developmental disorders (Vissers et al., 2016). Embryonic development is an extremely complex process requiring many proteins to function in the right amounts at precise times. Indeed, over 1,700 distinct genetic disorders can lead to intellectual disability, underscoring that mutations in numerous different genes may produce a similar developmental phenotype (Maia et al., 2021). This heterogeneity makes it hard to determine the genetic cause of a rare disease, as a one-to-one relationship between a phenotype and a genetic variant may not be established statistically (De Ligt et al., 2012; Hoischen et al., 2014).

Another challenge is that rare disease genetics can be caused by many different classes of genetic variants. The spectrum of genetic variation ranges from single base pair changes to large chromosomal rearrangements (**Table 1**). Among these, SNVs or point mutations involve the substitution of a single nucleotide at a specific locus. These single-nucleotide changes can have various consequences: loss-of-function (LoF) mutations (e.g., nonsense or frameshift variants) that truncate the protein or abolish its function; missense mutations that substitute one amino acid and may alter protein structure or function; and even variants in non-coding or intronic regions (including some traditionally considered 'synonymous' mutations) that can disrupt splicing or gene regulation. SNVs have been relatively well-studied and found to be a

prominent cause for a wide range of rare diseases. Depending on the particular disorder, in about 25-30% of patients the cause of disease can be attributed to SNVs (Dong et al., 2020).

Larger genetic alterations, collectively termed structural variants (SVs), can also cause rare diseases. SVs are defined as insertions (either novel sequences or duplications), deletions, translocations or inversions greater than 50 bases in size (Alkan et al., 2011). Recent systematic discovery studies have highlighted the substantial contribution of SVs to rare disease. For example, a study of 960 familial rare disease cases by Cohen et al. (2022) found that incorporating SV analysis yielded up to 13% additional diagnoses. Structural variants can disrupt genes in a myriad of ways, often leading to non-functional or missing proteins. Additionally, SVs in non-coding regions (e.g. promoters or enhancers) and copy-number variants (CNVs) can cause disease by altering gene expression levels (Weiner et al., 2023).

Given the sheer number of variants present in every human genome, a critical practical challenge is distinguishing the pathogenic mutation(s) from the many benign or unrelated variants that a patient carries. Researchers and clinicians must apply filters based on inheritance, predicted functional impact, population frequency, and other criteria to narrow down candidate variants. Even so, interpretation remains difficult – especially when the causative variant lies in poorly characterized parts of the genome or does not produce an obvious loss of function.

1.2.4 Genome variation

The characterization and cataloging of full human genetic variation has become an ongoing effort in genomics research. According to the 1000 Genomes study the typical human genome has on average 4 million SNVs, 1 million indels. Among these, around 40,000 to 200,000 are considered rare variants, with an allele frequency of less than 0.5%. The actual number of rare variants varies depending on the individual's particular ancestry and population structure (Mathieson & Reich, 2017). Such rare variants with functional impact on gene expression or protein structure are especially important for biomedical research. These include missense mutations, which alter a single amino acid that might impede the protein function. On the other hand, a nonsense mutation causes a protein to terminate or prevent its translation by creating an early stop codon. Similarly, frameshift mutations are insertions or deletions of a length that is not divisible by three, therefore disrupting the reading frame of triplet codons.

Another case of disruptions happens when a variant occurs at the boundary of an exon and an intron, preventing the splicing of the latter. These splice variants, along with nonsense and frameshift mutations are collectively called loss-of-function (LoF) or protein-truncating variants (PTVs) since they usually do not result in functional protein products. It is estimated that each human genome contains approximately 11,000 SVs with potential functional impact, including 250-300 with moderate to high functional impact (Lek et al., 2016).

1.2.5 *De novo* mutations

Genetic variants are also classified according to their different modes of inheritance, which adds another layer of complexity for researchers. One form of rare genetic variation that is crucial for human disease research is called *de novo* mutations (DNMs). Unlike most genetic variants, DNMs are not inherited but instead arise as a result of mutagenesis in germline cells or early embryogenesis (Veltman & Brunner, 2012). Compared to the inherited genetic variation, which are subject to purifying selection, DNMs arise spontaneously so they are more deleterious on average. DNMs are not rare at the genome level: recent long-read trio analyses estimate ~95 DNMs per child on average (≈88 single nucleotide variants (SNVs) and 8 insertion/deletions (indels)), with 15% of SNV DNMs arising postzygotically (Noyes et al., 2025). DNMs have a strong paternal bias (4:1 paternal:maternal), alongside a paternal age effect of roughly +1.3 SNVs per year (Noyes et al., 2025).

In the last decade, extensive studies revealed that DNMs are a major cause of sporadic genetic disorders, such as Kabuki syndrome, Cri du chat syndrome and Schinzel-Giedion syndrome (Barbosa et al., 2018). Furthermore, damaging DNMs have shown to be responsible for a substantial proportion of more common neurodevelopment disorders such as intellectual disability, autism spectrum disorders (ASDs) and early-onset epileptic encephalopathies (e.g., SCN1A-related Dravet syndrome) (Wright et al., 2023; Stawicka et al., 2024). They are less immediately impactful in autosomal recessive diseases, although rare instances occur in which a patient inherits one pathogenic allele and acquires a second de novo variant in trans, resulting in a recessive presentation (Wright et al., 2023). Together, these observations highlight both the frequency of DNMs in the human germline and their disproportionate clinical impact on severe, early-onset disorders (Noyes et al., 2025; Wright et al., 2023).

However, routine and reliable detection of DNMs remains technically challenging. Only with the advent of next-generation sequencing (NGS), unbiased discovery

of these mutations became feasible by sequencing trios of both parents and the patient. In the trio approach genetic information from parents is used to check Mendelian inheritance of each genetic variant in the proband (Allen et al., 2013; Iossifov et al., 2014). Since DNMs by definition are not present in parental genomes except in germline cells, they will break the rules of Mendelian inheritance when detected in the proband. These Mendelian inheritance errors (MIEs) can also be caused by erroneous genotyping calls and these need to be distinguished from true DNMs. Therefore, the sequencing technology that is used for the trio approach needs to have a genotyping error rate as low as 1-2% to allow for detection of true DNMs (Vulto-van Silfhout et al., 2013). Additionally, certain regions of the genome have structural and functional features, such as repetitive sequences or biased GC composition, that increase the error rate. Accurate and sensitive detection is especially salient in these regions, as they contain the most de novo SVs and CNVs. These variants are estimated to be as rare as 0.02 per genome (Conrad et al., 2010), but play an important role in neurodevelopmental disorders (Girirajan et al., 2013).

1.3 The role of Sequencing Technologies in the Detection of Genetic Variants

In the last 50 years, a variety of technologies have been developed to detect different types of genetic variation. The definitions of many variant types that we distinguish are based upon the original technologies that were applied to detect them.

1.3.1 Historical developments

Historically, Sanger sequencing was the dominant method to detect novel genetic variation. Developed by Frederick Sanger and colleagues in 1977, it involves in vitro DNA replication, during which chain-terminating dideoxynucleotides are randomly incorporated to produce DNA fragments with different lengths (Heather & Chain, 2016). These fragments are then separated by gel or capillary electrophoresis and read out by fluorescent imaging to determine nucleotide sequence. Sanger sequencing produces highly-accurate (>99.99%) reads that are up to 700-900 bp in length (Coe et al., 2014). However, the process is time-consuming – less than 1 Mb of DNA can be sequenced per hour (Shendure et al., 2019). Therefore, researchers need to localize the genetic signal using linkage analysis and fine mapping before performing Sanger sequencing in a targeted fashion. For this reason,

researchers traditionally performed linkage analysis and fine mapping to localize a genetic signal before sequencing candidate regions with Sanger technology. Despite its laborious nature, this targeted approach led to the characterization of around 1,000 (of an estimated 7,000) monogenic disorders by the year 2000, including many with high biomedical importance such as cystic fibrosis and Huntington's disease (Claussnitzer et al., 2020).

At the turn of the century, the draft human genome sequence was completed, marking a significant milestone in human genetics (Lander et al., 2001). At the same time, microarray technology had been developed to genotype multiple regions of the genome simultaneously. DNA microarrays contain short, specific oligonucleotides (also called probes) attached to a solid surface. These probes hybridize to complementary sequences in a denatured DNA sample, and hybridization is detected via fluorescence. Since a single array can contain millions of probes, microarrays enabled high-throughput screening of known genomic positions (Srivastava et al., 2019). However, because the probes need to be designed in advance, microarrays rely on existing knowledge about the genome sequence. This limits their utility for discovering truly novel or rare variants that were not included on the array.

1.3.2 Next Generation Sequencing technology

The limitations of the aforementioned methods, notably, the need for targeted sequencing and prior variant knowledge, led to the development of next-generation sequencing (NGS) in the late 2000s. NGS made it possible to identify the full spectrum of genetic variation in a high-throughput, unbiased manner (Shendure et al., 2017). In general, NGS technologies employ a "shotgun" approach: the DNA of interest is randomly fragmented in short (~300 – 600 bp) pieces, amplified (often by cloning or PCR), and then sequenced in a massively parallel fashion. Various methods were commercialized for the last step, but in the end the sequencing-by-synthesis (SBS) approach developed by Solexa/Illumina became the most widely adopted. SBS utilizes a specially designed chip called a flow cell, on which immobilized DNA fragments are replicated by the incorporation of chain-terminating fluorescent nucleotides. Flow cells can contain up to 1,000 clones of the same fragment, in order to amplify the fluorescent signal, which is detected by a charge-coupled device (CCD) camera. Unlike Sanger sequencing, chain-termination is reversible, since the fluorescent part of the incorporated nucleotides can be enzymatically cut off after the signal is detected. Then the next cycle of extension can be started by simply adding new fluorescent nucleotides. Automating this cycle of chain-

termination, signal detection and extension greatly reduces the amount of time and labor required for the sequencing. The latest iteration of Illumina machines can sequence up to 60,000 megabases of genetic material in an hour. Due to the high-throughput nature of the platform, it became possible to sequence large cohorts of patients in genetics research.

With the adoption of NGS, whole-exome sequencing (WES) emerged as a popular strategy for identifying disease-associated variants (Allen et al., 2013; Gilissen et al., 2011; Iossifov et al., 2014). WES focuses on the ~1–2% of the genome that codes for proteins (the exons), where a high proportion of known disease-causing mutations are found. In WES, the exonic regions are first captured or enriched from the DNA sample, then sequenced using a high-throughput NGS platform (often Illumina). This targeted approach is cost-effective for clinical diagnostics, since it concentrates sequencing effort on the most relevant portion of the genome, as exons constitute less than 2% of the human genome but contain many pathogenic variants. Early exome capture methods used array-based platforms: a microarray of probes representing all human exons used capture coding DNA fragments from the sample (Sulonen et al., 2011). However, this method required relatively expensive equipment, and large amounts of input DNA. Newer protocols switched to in-solution capture, using pools of biotinylated oligonucleotide probes in a test tube to hybridize with exonic sequences. The probe-bound fragments can then be pulled down with streptavidin-coated beads and recovered enzymatically (Parla et al., 2011).

As sequencing costs have dropped further, whole-genome sequencing (WGS) has become a feasible option in diagnostics. WGS involves sequencing an individual's entire genome with no capture or targeting step, thereby providing a comprehensive view of both coding and non-coding variation. WGS offers advantages not only in detecting variants outside of exons, but also in identifying structural variants that might be missed by targeted approaches. Detection of SVs from sequencing data relies on computational approaches to gauge read depth variability and needs a reliable baseline against which copy number gains or losses can be compared. WGS data (free from exome-capture biases) yields more even coverage across the genome, improving sensitivity for SNVs and SVs (Lelieveld et al., 2015). These benefits have made WES and WGS invaluable tools for identifying rare genetic variants (De Ligt et al., 2012; Gilissen et al., 2014).

Table 1 Overview of different genetic variant detection approaches (van der Sanden et al., 2023; Vissers et al., 2017).

Technology	Year	Read Length	Can detect	Approx. Diagnostic Yield	Advantages	Drawbacks
Sanger Sequencing	1977	~1000 bp	SNVs, Indels	1.5%	High accuracy, relative long read length	Low throughput
Genomic microarrays	2004	NA	SNVs, CNVs	11.6%	High throughput, low cost	Relies upon already known SNV variants
SRS Exomes	2012	100 -300 bp	Coding SNVs, Indels	30%	Cost-effective, unbiased discovery	No coverage of the non-coding regions, biased by the exome capture
SRS Genomes	2016	100 -300 bp	SNVs, Indels, SVs	50%	Comprehensive screening of whole genome	Relatively high cost, requires handling of large-scale data
LRS Genomes	2010	10-20 kb mean	SVs, SNVs, Indels, CNVs	?	Detection of all variant types, including large SVs	Significantly costlier than SRS, requires novel software

Columns, from left to right, indicate the sequencing technology, year of introduction, variant types that can be detected, diagnostic yield reported in the literature, main advantages and drawbacks.

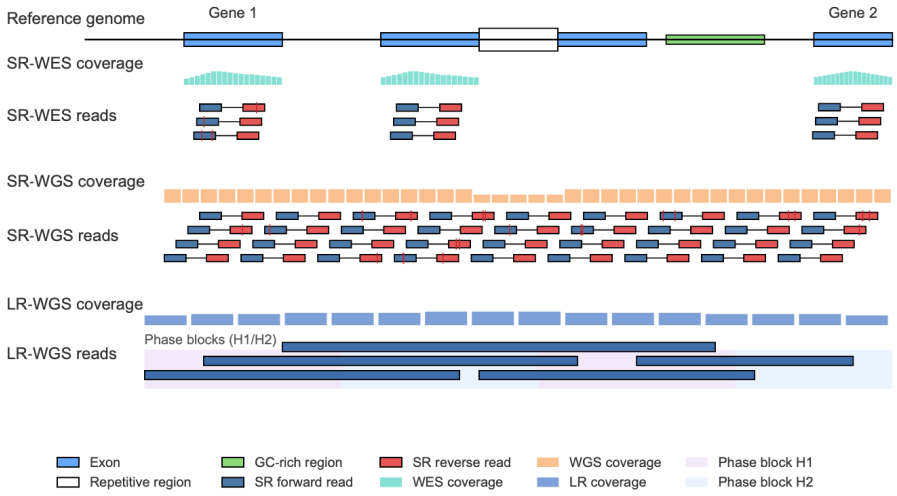

Figure 1. Schematic illustration of sequencing technologies and the abilities to detect genome variation, namely short read whole exome sequencing (SR-WES), short read whole genome sequencing (SR-WGS) and long read whole genome sequencing (LR-WGS).

Horizontal bars indicate sequencing reads. In the genome representation, red box indicates a repetitive genome region and yellow boxes indicate exons.

1.3.3 Long read sequencing technology

Sequencing technologies and computational methods have vastly improved our ability to detect genomic variants (L. Yang, 2020). Nevertheless, next generation sequencing approaches have various technical limitations. Short-read sequencing (SRS)—the mainstay of WES/WGS—reads the genome in ~100–150 bp fragments, which leads to uneven coverage in repetitive or low-complexity DNA. These regions often have extreme GC content (≥60%), which interferes with PCR/library preparation and produces coverage "gaps" or drop-outs (Benjamini & Speed, 2012). Consequently, short-read methods have higher false-negative rates in such regions. Another often-mentioned limitation of short-read sequencing is its difficulty in certain repetitive or biased regions, rather than systemic error per se. The error rate of modern short reads is quite low and largely random (≥99% raw read accuracy), but short reads struggle with mapping to highly repetitive sequences or GC-rich regions, which can lead to coverage gaps and missed variants (Heather & Chain, 2016). Moreover, a single short read cannot span most SV breakpoints, so large insertions/deletions and complex rearrangements must be inferred indirectly from discordant pairs or depth fluctuations, which constrains sensitivity and precision (Alkan et al., 2011).

These limitations have spurred the development of long-read sequencing (LRS) as a complementary/alternative approach (Figure 1). Two main platforms are in use: Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio). ONT measures ionic-current changes as single DNA molecules traverse a nanopore and can yield ultra-long reads (tens to hundreds of kilobases), with recent chemistries/basecallers delivering ~98–99% single-read accuracy for small-variant calling when evaluated in head-to-head benchmarks (Jain et al., 2018; Mahmoud et al., 2024). PacBio's single-molecule real-time (SMRT) technology observes fluorescent nucleotide incorporations in zero-mode waveguides (Eid et al., 2009). A key advance for PacBio has been circular consensus sequencing (CCS), which repeatedly reads the same molecule to produce HiFi reads of ~10–25 kb in length with >99.5–99.9% per-base accuracy (Wenger et al., 2019; Hon et al., 2020). Because standard HiFi library prep is amplification-free, GC bias is markedly reduced: difficult high-GC targets that fail in capture/PCR (≥60% GC) are generally covered near genomic averages by HiFi WGS, with only rare motif-specific deficits (Nurk et al., 2022). In recent head-to-head benchmarks, PacBio HiFi achieved small-variant F1 scores ≈99–99.9% (e.g., F1 = 99.87% on HG002), implying false-positive/false-negative rates on the order of ~0.1–1%; ONT (merged callers) reached F1 ≈ 98.7% in the same setting (Mahmoud et al., 2024). For structural variants (SVs ≥50 bp), Pacbio LRS provides a clear advantage: long reads routinely detect ~20,000–25,000 SVs per human genome, and benchmark SV F1 ≈ 0.90–0.93 has been reported for Pacbio HiFi LRS pipelines versus ~0.45 for short-read callers in medically relevant regions (Mahmoud et al., 2024). In contrast to inference from paired-end signatures, long reads frequently span entire SVs and resolve breakpoints at base-pair resolution (M. J. Chaisson et al., 2015; Huddleston et al., 2017). Beyond variant discovery, LRS confers additional benefits: haplotype phasing across tens of kilobases (enabling parent-of-origin assignment for de novo mutations and compound-heterozygote resolution) and epigenetic readouts—direct current-signal methylation in ONT and kinetics-based 5-mC detection in PacBio—without separate assays (Hon et al., 2020).

However, like all technologies, LRS has drawbacks that should be considered. Historically, LRS offered lower throughput and higher reagent cost than SRS, although this gap is narrowing. PacBio's latest Revio platform yields ~100–120 Gb HiFi per SMRT Cell and runs four cells in parallel, enabling ~400–480 Gb (~3–4 human genomes at 30×) per ~24 h run, with annual capacity on the order of ~1,000–1,300 genomes per instrument; reagent cost for a 30× HiFi genome is now ≈ US $1,000 (Hale, 2022; Pacific Biosciences, 2023).

Earlier concerns about long-read error rates (8–15% raw) are largely mitigated by HiFi mode (Figure 2), yielding QV ≥ 30–40 consensus reads and small-variant accuracy that approaches or matches short-read performance (Wenger et al., 2019; Hon et al., 2020; Mahmoud et al., 2024).

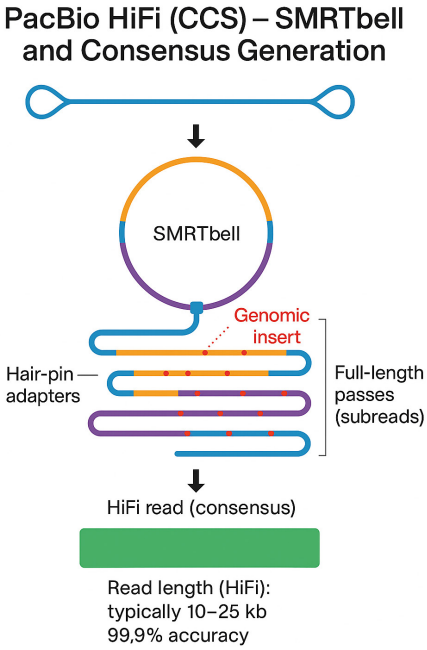

Figure 2. Circular consensus (HiFi) sequencing with PacBio SMRT. Genomic DNA is ligated to hairpin adapters to form a closed SMRTbell template (top). A single DNA polymerase repeatedly traverses the circular insert, generating multiple subreads (forward and reverse passes). Random errors present in individual subreads (red dots) are averaged during consensus calculation to produce one high-accuracy HiFi read (bottom). This CCS process yields long inserts (typically ~10–25 kb) with per-base accuracies >Q30–Q40 (>99.5–99.9%), enabling precise small-variant calling, phasing across kilobases, and stable coverage in GC- and repeat-rich regions.

Importantly, highly accurate long reads make it feasible to detect all major variant classes—from SNVs/indels to large SVs and tandem repeats—in a single assay without capture bias. Benchmarks consistently show that HiFi matches short reads for small variants while substantially improving detection and breakpoint resolution of insertions, complex rearrangements, and other variants concentrated in repetitive or GC-extreme regions (Mahmoud et al., 2024; Nurk et al., 2022). Meanwhile, ONT's ultra-long reads uniquely aid assembly across multi-megabase repeats and segmental duplications, complementing HiFi's per-base accuracy (Nurk et al., 2022). Together, these

capabilities highlight the potential of LRS to reveal "hidden" variants in genomic dead zones that SRS struggles to interrogate, while also enabling phasing and methylation analysis within the same dataset.

1.4 The Aim of This Thesis

The advent of highly-accurate long-read sequencing has important implications for clinical research and diagnostics. In the past, the use of new technologies improved our understanding of genetic disorders and increased the number of patients with a molecular diagnosis (Claussnitzer et al., 2020). However, employing a new technology has many challenges, since each technology has its own characteristics that require a different approach in data analysis and interpretation (Sedlazeck et al., 2018). The aim of this thesis is to investigate the advantages of new technologies, such as improved exome capture and long-read sequencing technology for improving the identification of genetic variants in patients.

In **Chapter 2** we investigated whether new exome capture kits can improve variant detection in coding regions. We evaluated the ability of three different exome capture kits and whole genome sequencing to detect variation in the coding regions. In this study I designed and performed the comparison analysis with long-read whole-genome sequencing (WGS) data. This included performing SNV/SV variant calling using long-read sequencing (LRS) data, and performing a comprehensive comparison to the short-read data from different enrichment kits. The analysis aimed to evaluate the consistency and accuracy of variant detection, especially for structural and complex variants, which are more accessible through LRS.

In **Chapter 3** we hypothesized that new long-read sequencing technologies can uncover variation that is missed by short-read sequencing technology. We performed long-read sequencing for 5 proband-mother-father trios and compared its ability to identify SNVs and structural genome variation to current short-read sequencing technology. In this study I designed and applied the SNV calling pipeline and conducted a comparative analysis with short-read sequencing (SRS) data. I also performed quality assessment of both structural variants (SVs) and SNVs using Mendelian inheritance error analysis, the identification of de novo SNVs, and an investigation of recessive inheritance patterns. Furthermore, I compared sequencing coverage between SRS and

LRS platforms to assess their impact on variant detection. I was responsible for drafting, finalizing, and submitting the manuscript for publication.

Chapter 4 investigates whether the latest (HiFi) long-read sequencing has become sufficiently accurate to replace short-read sequencing technology for the detection of SNVs. We sequenced 8 proband-mother-father trios with both short and long-read sequencing and compared the ability of both technologies to detect *de novo* point mutations. I was solely responsible for designing and conducting all the bioinformatic analyses, which encompassed variant calling and annotation from both LRS and SRS data, comparative studies of inherited and small de novo variants, titration analyses to assess sensitivity, and phasing of de novo mutations to determine their parental origin. I was responsible for drafting, finalizing, and submitting the manuscript for publication.

Finally, **Chapter 5** discusses key findings of the thesis, highlighting how recent advancements in sequencing technologies enhance rare disease clinical research and diagnostics. The chapter also acknowledges current limitations, such as the need for larger cohort studies to validate clinical utility, lack of standardized tools for analyzing LRS data, and the difficulty of interpreting novel variants due to limited annotations. Future directions emphasize the need for gold-standard datasets, better validation technologies, and improved variant interpretation tools to translate these genomic insights into better clinical outcomes.

1.5 Research Data Management

Ethics and Privacy: This thesis is based on the results of research involving human participants, which were conducted in accordance with relevant national and international legislation and regulations, guidelines, codes of conduct and Radboudumc policy. Chapters 2, 3, and 4 use data collected from human participants in the context of healthcare. *The privacy of the participants in these studies was warranted by the use of pseudonymization.* Informed consent was obtained from participants to collect and process their data for this research project. Consent was also obtained for sharing and reuse of the (pseudonymized) data for future research.

Data collection and storage: Data for chapters 2, 3 and 4 were obtained by sequencing, stored and analyzed using Radboudumc's internal high-

performance computing systems and archived in the Human Genetics department. The metadata associated with this data, such as age, sex, disease and family relation information, is also collected by the Human Genetics department and treated as protected health information.

Data Sharing via FAIR principles: Chapters 2, 3 and 4 are published, all with open access. Target files for enrichment kits used in Chapter 2 are available online from the respective manufacturers (Agilent and TWIST). Data for all samples sequenced in Chapter 3 are available with restricted access at EGA under accession number EGAS00001004319. The datasets from Chapter 4 have also been uploaded to EGA, under accession number EGAS00001006479 and are available with restricted access. For both datasets, requests for access will be checked by the Data Access Committee (Christian Gilissen, Alexander Hoischen and Lisenka Vissers), against the conditions for sharing the data as described in the signed Informed Consent.

1

2. Twist exome capture allows for lower average sequence coverage in clinical exome sequencing

Burcu Yaldiz[1,*], Erdi Küçük[1,*], Juliet Hampstead[1], Tom Hofste[1], Rolph Pfundt[2], Jordi Corominas Galbany[1], Tuula Rinne[1], Helger G. Yntema[2], Solve-RD consortium, Alexander Hoischen[1], Marcel Nelen[1], Christian Gilissen[1]

[1] Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Centre, Geert Grooteplein 10, 6525 GA, Nijmegen, The Netherlands

[2] Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Centre, Geert Grooteplein 10, 6525 GA, Nijmegen, The Netherlands

[*] These authors contributed equally to this work.

2.1 Abstract

Exome and genome sequencing are the predominant techniques in the diagnosis and research of genetic disorders. Sufficient, uniform and reproducible/consistent sequence coverage is a main determinant for the sensitivity to detect single nucleotide (SNVs) and copy number variants (CNVs). Here we compared the ability to obtain comprehensive exome coverage for recent exome capture kits and genome sequencing techniques.

We compared three different widely used enrichment kits (Agilent SureSelect Human All Exon V5, Agilent SureSelect Human All Exon V7 and Twist Bioscience) as well as Short Read- and Long Read-WGS. We show that the Twist exome capture significantly improves complete coverage and coverage uniformity across coding regions compared to other exome capture kits. Twist performance is comparable to that of both short and long-read whole genome sequencing. Additionally, we show that even at a reduced average coverage of 70x there is only minimal loss in sensitivity for SNV and CNV detection.

We conclude that exome sequencing with Twist represents a significant improvement and could be performed at lower sequence coverage compared to other exome capture techniques.

2.2 Background

Next generation sequencing (NGS) techniques are widely used across clinical and research applications in genetics. With the improvements in targeted sequencing approaches, whole exome sequencing (WES) has become a standard tool in clinical diagnostics (Arts et al., 2019; Corominas et al., 2022; Hartman et al., 2019; McNeill, 2022; Neveling et al., 2013; Rabbani et al., 2014).

There are various exome capture kits with different target enrichment strategies. Selection of target genomic regions, sequence features, length of probes and exome capture mechanisms are the major differences among these kits. These characteristics may give rise to differences in the overall coverage uniformity and capture efficiency of specific targets, resulting in decreased variant calling sensitivity. Several studies that compared exome capture technologies have shown that there are major differences in their performance (Belova et al., 2022; Díaz-de Usera et al., 2020; Parla et al., 2011; Zhou et al., 2021) and that high average read depth does not guarantee coverage for individual targets. In these comparative studies, extreme GC content (Clark et al., 2011; Meienberg et al., 2015; Q. Wang et al., 2017) and mappability issues (Barbitoff et al., 2020; Q. Wang et al., 2017) are shown to be the major sources of coverage bias.

Sufficient, uniform and reproducible/consistent sequence coverage is required for robust and sensitive single nucleotide variant (SNV) and copy number variant (CNV) detection in exome data. While CNVs are not routinely detected from WES in each laboratory or pipeline, their additional clinical utility (Dong et al., 2020b; Pfundt et al., 2017; Royer-Bertrand et al., 2021) urges for reliable CNV detection from exomes, especially when patient cohorts are not routinely pre-screened by CNV-microarrays. CNV detection from WES data particularly depends on the analysis of read depth variations at sequencing targets. Large sets of reference samples are typically required in order to robustly compare CNV coverage profiles in exome data. Therefore, over- and underrepresentation of target regions due to extreme GC content and mappability issues can dramatically affect the robustness of CNV calling from exome data (Royer-Bertrand et al., 2021). Short- and Long read Whole Genome Sequencing (SR-WGS and LR-WGS respectively) approaches generally yield more uniform and complete coverage profiles than exome sequencing and the gapless nature of WGS data enables more accurate detection of CNVs and structural variants (SVs). However, lower sequencing and storage costs as well as the demonstration of diagnostic yield of CNV detection have led WES to

be proposed as a first tier diagnostic test in recent studies (Martinez-Granero et al., 2021; Srivastava et al., 2019).

In the last few years, new exome capture and sequencing technologies, particularly the Twist exome capture kit and long read sequencing (LRS) technologies, have been applied in clinical sequencing studies (Diaz-Horta et al., 2019; Pauper et al., 2021; Shakked et al., 2021). Here, we compared the Twist exome capture kit's coding sequence coverage and SNV detection sensitivity to other widely used exome kits as well as to SR- and LR-WGS. As further benchmarks, we utilized the SR- and LR-WGS methods which are purported to provide optimal uniformity and coverage profiles (Pauper et al., 2021). We assessed the sensitivity of SNV and CNV calling of Twist exome capture kit at reduced average coverage levels.

2.3 Methods

2.3.1 Sample Collection

2.3.1.1 Whole Exome Sequencing

Various studies have evaluated the effectiveness of established enrichment technologies such as Agilent SureSelect, Nimblegen SeqCap and Illumina TruSeq. These comparisons have shown relatively modest differences between the most recent versions of these technologies, mostly due to differences in target design. In this study we investigated a completely novel capture method by Twist Bioscience (Twist). Twist uses a silicon-based DNA synthesis technology that allows for the production of larger quantities of oligonucleotides, resulting in more probes and improved rebalancing, which was expected to yield significant improvements in target coverage and coverage uniformity. We compared Twist exome capture to one of the latest Agilent SureSelect captures (V7) which has been shown to perform on par with other commonly used exome capture technologies. In addition, we included an older version of the Agilent SureSelect (V5) which has been widely used in the past to provide a point of reference (Meienberg et al., 2015; Neveling et al., 2013; Sekhar et al., 2014; Shigemizu et al., 2015). All samples were sequenced using Illumina HiSeq4000 2x150bp sequencing (Lelieveld et al., 2015) in our center. We collected 20 whole blood patient samples sequenced using each of the three kits randomly (**Table 1**). These samples were downsampled to 100x as described below:

- Samples sequenced using the Agilent V5 enrichment kit with a mean coverage of 274.8x.
- Samples sequenced using the Agilent V7 with a mean coverage of 239.6x.
- Samples sequenced using the Twist enrichment kit with a mean coverage of 139.2x.

Table 1 Overview of samples used in this study

Enrichment/Library	Average Coverage	Coverage Range (min – max)	Number of Samples
Agilent V5	274.8	163.8 - 345.4	20
Agilent V7	239.6	131.1 - 370.6	20
Twist	139.2	119.7 - 158.5	20
WGS	59.3	50.86 - 69.33	20
LRS	29.4	24.24 - 38.86	18

Columns depict (from left to the right) the exome kits and the platforms; the average coverage across the target regions of the enrichment kits for the exomes; the range of coverage; the number of samples used in the analysis.

In addition to these samples, 7 exome samples captured with Twist enrichment kit with lower average coverage of 69.95x, five exome samples collected from three different tissues (amniotic fluid, basal mucosa (buccal swab) and fibroblasts) captured by Twist enrichment kit were also used for further comparisons **(Additional file 1: Table S1)**. Besides, 14 Twist samples with previously validated CNVs and an additional 100 Twist samples as a reference pool were used for performing CNV analysis **(Additional file 1: Table S2)**. These additional samples were used as control samples for normalization of the read counts and they were not involved in other comparisons.

All samples were sequenced on an Illumina NovaSeq 6000 sequencer using 2x150 paired-end sequencing. All exome samples were aligned by the Burrows Wheeler Aligner (BWA) (H. Li & Durbin, 2009) to the hg19/GRCh37 assembly of the human reference genome. Duplicates were marked, and GATK best practices were followed during the mapping process.

2.3.1.2 Short Read Whole Genome Sequencing

20 SR-WGS samples were sequenced using 2x150 bp paired-end on an Illumina NovaSeq 6000 sequencer to 59.3x mean coverage (**Additional file 1: Table S1**). Alignment was performed by using Burrows-Wheeler Aligner (BWA) (H. Li & Durbin, 2009) to the hg19/GRCh37 assembly of the human reference genome.

2.3.1.3 Long Read Whole Genome Sequencing

We also sequenced 6 trios (18 samples) with a Pacific Biosciences Sequel II instrument. We used three SMRT chips per sample, targeting 30x mean coverage with HiFi reads **(Additional file 1: Table S1)**. Reads were aligned to the hg19/GRCh37 assembly of the human reference genome with pbmm2 (version 1.4.0) using default parameters.

2.3.2 Gene Definitions

Genes and coding regions were defined using NCBI RefSeq (Release 61) (Pruitt et al., 2014) and EMBL-EBI Ensembl GENCODE (Release 91) (Cunningham et al., 2019) transcripts of the hg19/GRCh37 assembly of the human reference genome. Transcripts of both databases were downloaded from the UCSC Table Browser (Karolchik et al., 2004). We generated transcript files for only protein coding regions on chromosomes 1-22 and X in bed format using a custom Python script. Overlapping regions were merged using BEDTools v2.28.0 (Quinlan & Hall, 2010). RefSeq contained 197,736 exons and 19,259 genes and Ensembl 209,103 exons and 20,691 genes.

Disease genes were derived from the Online Mendelian Inheritance in Man (OMIM)'s Synopsis. The coding regions for the longest transcripts of 4,531 OMIM genes with the highest level of evidence were extracted from the RefSeq transcripts.

2.3.3 Downsampling, Coverage Calculation, GC Content and Evenness Scores

Sequence data was downsampled using SAMTools v1.10. (H. Li, Handsaker, Wysoker, Fennell, Ruan, Homer, Marth, Abecasis, Durbin, & Subgroup, 2009) for all samples. Single base-pair coverage of human protein coding regions was calculated for samples in all coverage level groups using BEDTools v.2.28.0. GC content was also calculated using BEDTools v2.28.0. The distribution of coverage over target regions was assessed by calculating an evenness score as defined by Mokry *et al.* (Mokry et al., 2010). The evenness score represents the fraction of sequenced bases that do not have to be redistributed from above-average coverage to below-average coverage positions to obtain completely even coverage for all targeted positions. This is a measurement that is relatively independent of sequencing depth.

2.3.4 Variant Comparison

Variants for all Illumina samples (WES and SR WGS) were called using the GATK HaplotypeCaller (version 3.4) (der Auwera & O'Connor, 2020). Target exonic regions for respective kits were extended 200bp upstream and downstream for variant calling. DeepVariant (version 1.1.0) was used for variant calling with default parameters for LR WGS samples. All variants were subsequently annotated by our in-house pipeline based on the Ensembl Variant Effect Predictor (VEP). Coding variants were compared by selecting true positive variants with allele frequencies > 0.001 (ExAC v0.2).

2.3.5 CNV Comparison for Twist

To examine the effect of coverage level on the sensitivity of copy number variation (CNV) detection, we used two independent data sets as described in Sample Collection. We used 20 randomly selected Twist samples (**Additional file 1: Table S1**) and additional 14 Twist samples with previously validated CNVs (**Additional file 1: Table S2**). We used an additional 100 Twist samples as a reference pool for CNV calling (**Additional file 1: Table S2**). All samples were downsampled to both 100x and 70x coverage for comparison. Since Conifer is used in inhouse diagnostic pipeline CNV calling was performed using Conifer v.0.2.2. We considered true CNVs to be calls with SVD-ZRPKM values smaller than -1.7 (deletions) or 1.7 (duplications). We additionally removed 3 singular values based on the inflection point of scree plots (**Additional file 1: Figure S1**).

2.4 Results

We compared three different widely used enrichment kits (Agilent V5, Agilent V7 and Twist) as well as SR- and LR-WGS. Randomly selected whole blood and tissue samples for all kits and SR-WGS were sequenced on an Illumina NovaSeq 6000 sequencer using 2x150 paired-end sequencing and LR-WGS samples were sequenced on a Pacific Biosciences Sequel II instrument.

2.4.1 Percentage of Coding Regions Covered (RefSeq and Ensembl) in WES and WGS

Differences in sequence coverage foremost stem from differences in the target design. Therefore, we compared the overlap between the extended targets (+/-200bp) of three capture kits analyzed (Agilent v5, Agilent v7, and Twist) with coding regions as defined using RefSeq and Ensembl data (see Methods). While the older Agilent v5 capture kit did not target about 980 kb of RefSeq

coding sequence, the newer Agilent v7 and Twist kits perform substantially better (148kb missing, Agilent v7; 83kb missing, Twist; **Additional file 1: Table S3**). The coding regions as defined by Ensembl data are broader than those defined using RefSeq data. We found that Twist does not target about 753 kB of these regions whereas Agilent v7 does not target about 348 kB (**Additional file 2**).

We then compared the percentage of the coding regions covered by at least 20x across WES data sequenced using each of the three exome capture kits, SR-WGS data, and LR-WGS data. All exome samples were downsampled to 100x average coverage (**Additional file 1: Table S4**). The highest coverage ratio at >20x for both RefSeq and Ensembl coding regions was obtained with Twist enrichment kits (**Figure 1A**). Twist covered 99.4% of the RefSeq and 97.5% of the Ensembl coding regions by 20x, while Agilent v7 and Agilent v5 covered 96.7% and 87.6% of RefSeq coding regions and 96% and 87.4% of Ensembl coding regions respectively. However, SR-WGS is superior to all three WES capture kits by this metric, covering 99.7% and 99.6% of RefSeq and Ensembl coding regions at 20x. LR-WGS reached only 89.5%, likely due to the lower average coverage of only 30x (**Additional file 1: Table S5a**). This is also the reason for the high standard deviation for LR-WGS. When we considered 10x minimal coverage sufficient in all LR-WGS samples, we found that LR-WGS performed similarly to SR-WGS (SR-WGS: 99.90%, LR-WGS: 99.2% for 10x RefSeq coverage; **Additional file 1: Table S5b**).

2.4.2 Evenness of Coverage

We also calculated an evenness of coverage score for all samples (**Methods**). Twist exomes have better uniformity of sequence coverage using this metric compared to Agilent v5 and v7 exomes (**Figure 1B**, **Additional file 1: Table S6**). An advantage of uniform coverage is that samples can potentially be sequenced at lower average coverage, thereby providing considerable cost savings. To investigate this in our data, we downsampled Agilent v7 and Twist exome samples to 50x mean coverage. Downsampled Twist exomes achieved a 97.2% and 95.2% coverage ratio for RefSeq and Ensembl coding regions respectively, constituting a 2.2% and 2.3% decrease in sufficiently covered regions (**Figure 1C**). In downsampled Agilent v7 exomes the decrease in sufficiently covered regions was 7.2% and 7.3% resulting in 89.5% and 88.7% coverage ratios for RefSeq and Ensembl coding regions respectively.

2.4.3 GC Content

A well-known reason for poor performing enrichment targets is extreme GC content. Therefore, we assessed the GC content of regions with insufficient

coverage (<20x) (**Methods**). The median GC ratio of insufficiently covered regions in our data was 38.8%, 37.5%, 66.6%, 53.1% and 55% for Agilent v5, Agilent v7, Twist, WGS and LRS samples respectively (**Figure 1D**). In regions that were well-covered the median GC content for all platforms was between 50%-53.2%. Interestingly, while Agilent v5 and v7 typically perform poorly in low GC regions, in Twist samples most low coverage regions have an high GC content (>65%). As expected, the GC content distribution of well and poorly covered regions in SR- and LR-WGS data are similar.

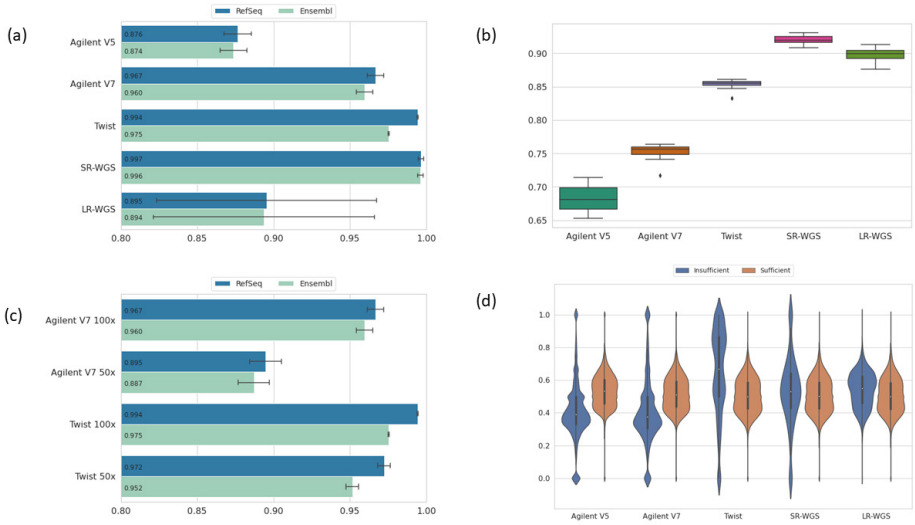

Figure 1. Comparison of exome kits and sequencing platforms

A Ratio of coding regions covered at ≥20x for different enrichment and sequencing platforms for RefSeq and Ensembl. **B** Boxplots of evenness scores for different enrichment kits and sequencing platforms. **C** Ratio of coding regions covered at ≥ 20x for different enrichment platforms when down-sampled to 50x. **D** GC content of insufficiently and sufficiently covered targets are significantly different for all kits and platforms (Mann Whitney U-Test p-value <0.001).

2.4.4 Twist Enrichment Kits Have Lower Minimum Average Coverage Requirements Than Agilent V7 Kits

Next, we wanted to establish a minimum level of average coverage sufficient to obtain results comparable to 100x average coverage in exome data. To do this, we assessed the effect of gradually downsampling average coverage to 20x in exome data (Twist and Agilent v7 kits) and 10x in genome data (**Figure 2A**, **Additional file 1: Table S7**). We show that the percentage of covered coding regions declines more rapidly in downsampled Agilent v7 exomes compared

to Twist exomes. For example, when downsampling from 70x to 60x average coverage the percentage of covered coding regions declines by 1.7% in Agilent v7 exomes (94.2% to 92.5) versus just 0.1% in Twist exomes (99% to 98.9%). When average coverage is reduced to 30x, only 74% and 82% of coding sequence is covered more than 20x for Agilent v7 and Twist respectively. We verified that these results are also valid for samples with DNA from other tissues (amniotic fluid, basal mucosa and fibroblasts) than blood enriched with Twist (**Additional file 1: Table S8; Figure S2**).

To investigate how lower average coverage might impact variant detection, we selected all common coding variants with an ExAC allele frequency > 0.001 (0.1%) in all WES and WGS samples. In gradually downsampled Twist exomes the median number of coding variants decreased only slightly up to 40x. While the difference between median number of coding variants was 360 between 100x and 40x this difference increased to 690 variants between 40x and 20x for Twist samples (**Figure 2B**). Similarly, the median number of coding variants remains relatively consistent down to 20x for SR-WGS samples, after which we observed a strong decline. However, for Agilent V7 samples median number of coding variants decreased by 255 when average coverage of samples reduced to 60x from 100x and this difference was 2019 when average coverage reduced to 40x from 60x. On average the number of detected coding variants with ExAC allele frequency > 0.1% was consistently smaller for Agilent V7 samples compared to Twist samples at each level of average coverage.

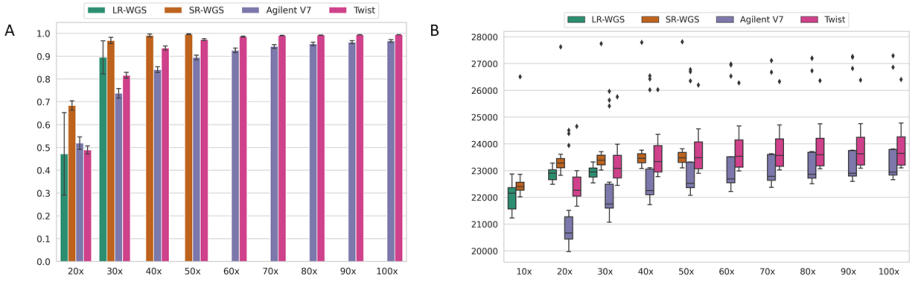

Figure 2. Comparison of enrichment kits and sequencing platforms at different coverage levels

A Overview of basepair coverage ratio at least 20x per platform for RefSeq coding regions. X-axis represents the mean coverage levels of the samples in each platform, y-axis represents the average ratio of base pairs that exceeds 20x coverage level for all samples in the corresponding kit/platform. **B** Boxplots represents the distribution of the number of coding variants for samples of each platform at different coverage levels. X-axis depicts the coverage levels and y-axis shows the number of coding variants.

2.4.5 Coverage of Clinically Relevant Genes

Our results show that Twist outperforms other kits and performs similar to WGS in terms of coverage and SNV detection. Additionally, we show that reducing average coverage to 70x in Twist exome data would likely have a negligible impact on the percentage of sufficiently covered regions and sensitivity of SNV detection. To determine whether 70x Twist exomes could be used in clinical diagnostics, we performed further detailed comparisons between Twist samples with average 70x and 100x coverage. RefSeq coding regions were used for further comparisons since Twist targets cover RefSeq regions better than Ensembl regions.

First, we verified that our downsampling procedure did not affect our results by repeating the coding region coverage analyses for 7 samples that were originally sequenced at 70x average coverage. On average 98.8% of the RefSeq coding regions were covered by at least 20x in these samples (**Additional file 1: Table S1**; **Table S7a)**.

To better understand the clinical importance of differences in coding region coverage, we assessed the coverage of transcripts of 4,531 OMIM transcripts which consist of ±10 mb distributed over 62,233 exons extracted from RefSeq coding regions (**Methods**). We examined the percentage of these transcripts with at least 20x coverage at all bases. In 100x Twist exome samples an average of 91% of OMIM transcripts were fully covered. In 70x Twist exome samples we observe a substantial decrease in the complete coverage of these transcripts (74.8%, **Figure 3A**). This drop is driven by a relatively small proportion of coding bases: 95% of bases exceed 20x coverage in 95% of OMIM transcripts in 70x Twist data **(Additional file 1: Table S9)**.

2.4.6 Genuine SNVs Can Still Be Detected in 70x Twist Exomes

The number of ExAC AF > 0.001 variants detected in 70x Twist exomes was comparable to that in 100x Twist exomes (0.5% of variants not detected at 70x). Although the total number of detected variants decreased only slightly for Twist when down-sampling from 100 to 70x coverage, we were interested in which variants specifically were lost. 20% were located in genes such as *MUC6, TAS2R45, HLA-DRB5* and *MUC4* that have previously been associated with mapping artifacts due homologous regions (**Additional file 1: Figure S3;** (Mandelker et al., 2016)). 80% were mapped to various genes in different samples. In addition, we wondered whether down-sampling had an effect on GATK quality scores, since these are commonly used to select less reliable

calls for orthogonal validation. While we observed that GATK quality scores were highly correlated in 70x and 100x Twist exomes (**Additional file 1: Figure S4**), we also show that the tails of the quality score distribution may be affected by the drop in coverage. Only 9% of variants had GATK quality scores less than 500 in 100x Twist exomes, while this increased to 19% in 70x samples.

2.4.7 CNVs Can Still Be Detected in 70x Twist Exomes

Another potential concern with having lower average coverage is the ability to call copy number variants (CNVs) based on depth-of-coverage using a relatively heterogeneous reference pool of only 100 samples. To address this, we examined the effect of lower coverage on CNV detection using Conifer. We compared CNV calls in 20 Twist samples with downsampled 70x coverage to those with 100x coverage (see Methods). To do this, samples in the reference pool were also downsampled to 100x and 70x average coverage. SVD normalization enables Conifer to remove coverage biases introduced by the capture and sequencing of exomes and detect only rare CNVs. Accordingly, in this study 67 CNVs were called in samples at both 100x (75 CNVs in total) and 70x (71 CNVs in total) coverage (**Figure 3B**). In downsampled 70x Twist exomes 6 duplications and 1 deletion did not exceed the SVD-ZRPKM value threshold (**Methods**) and 1 duplication was not called. In comparison, 1 duplication and 1 deletion did not exceed filtering thresholds in 100x Twist samples (**Additional file 1: Table S10**).

We also compared the CNV calls for 100x and 70x average coverage levels in another group of unsampled Twist exomes with a set of previously validated CNVs **(Additional file 1: Table S2)**. In 100x Twist samples, 10 out of 15 CNVs were called, 3 CNVs did not exceed the filter thresholds and 2 CNVs were not called (**Table 2**). In 70x Twist samples, 8 CNVs were called and 5 CNVs did not exceed the filter threshold. The same 2 CNVs that were missed in 100x Twist samples were also undetected. Although 3 CNVs did not exceed the SVD-ZRPKM threshold for both coverage levels they could be easily identified based on visual inspection of the coverage bedgraphs (**Additional file 1: Figure S5**). Almost all CNVs detected by 100x samples were also detected by 70x samples however a few of them were filtered out since they did not exceed the SVD-ZRPKM threshold value in both sample sets.

Table 2 CNV Status of 100x and 70x samples for the validated CNVs

Validated CNVs	CNV Status of 100x and 70x Samples					
Sample	**Chromosome**	**Start position**	**End position**	**CNV Type**	**100x Samples**	**70x Samples**
CNV_Sample_1	17	2516458	2808662	Deletion	Cannot exceed threshold	Cannot exceed threshold
CNV_Sample_2	15	23572075	28567878	Deletion	Called	Cannot exceed threshold
CNV_Sample_3	8	116085	43218462	Duplication	Not called	Not called
CNV_Sample_4	22	21562426	22937526	Deletion	2/3 segments	2/2 segments
CNV_Sample_5	16	14927708	16367932	Duplication	2/3 segments	1/2 segments
CNV_Sample_6	22	18893887	21414817	Deletion	3/4 segments	3/5 segments
CNV_Sample_7	23	24190859	26236246	Duplication	Cannot exceed threshold	Cannot exceed threshold
CNV_Sample_8	19	11105503	11141569	Deletion	Not called	Not called
CNV_Sample_9	11	pter	926088	Duplication	Called	Cannot exceed threshold
CNV_Sample_10	16	15457515	17564653	Deletion	2/2 segments	2/3 segments
CNV_Sample_11	17	1082960	1490254	Duplication	Called	Called
CNV_Sample_12	8	12051483	43218462	Duplication	2/5 segments	6/19 segments
CNV_Sample_12	8	pter	7079475	Deletion	1/2 segments	1/2 segments
CNV_Sample_13	6	160638463	qter	Deletion	4/7 segments	3/9 segments
CNV_Sample_14	22	50297485	50757432	Deletion	Cannot exceed threshold	Cannot exceed threshold

Columns depict (from left to right): sample ID of the samples with previously validated CNVs by visual inspection and concordance with phenotype; chromosome number; start position; end position of the validated CNV; CNV type; status of validated CNV for samples at 100x coverage; status of known CNV for sample at 70x coverage.

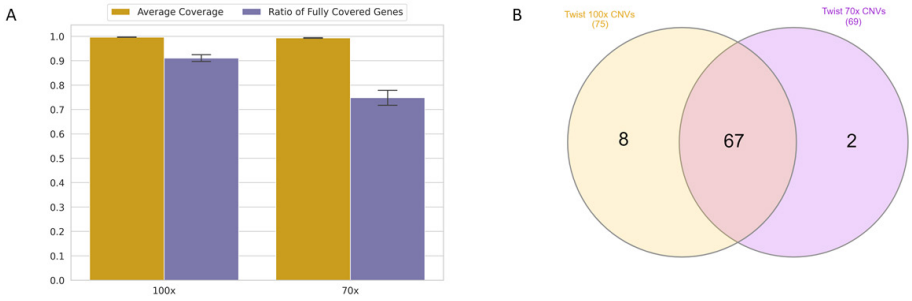

Figure 3. Comparison of Twist enrichment kit for 100x and 70x coverage levels

A Percentage of base pairs that exceeds 20x coverage level for OMIM genes (Yellow) percentage of genes which were fully covered with at least 20x coverage (Purple) **B** Venn Diagram that represents the number of CNVs for samples enriched with TWIST at 100x (Yellow) and 70x (Purple) coverage levels.

2.5 Discussion

Whereas for whole genome sequencing it is customary to only obtain 30-40x average coverage, this is not the same for exome sequencing due to the more uneven coverage that is the result of differences in capture efficiency for individual probes. Various studies have tried to help investigators make an informed decision on which sequencing platform to choose by comparing the performance of different WES kits with each other and with WGS using coverage and variant identification statistics (Iadarola et al., 2020; S. W. Kong et al., 2018; Lelieveld et al., 2015). Here we showed that Twist exome coverage is more uniform and consistent than coverage from other exome kits, and that there is a substantially smaller fraction of insufficiently covered coding bases. Although not as good as WGS, the results are very similar. These improvements are likely a result of the more or better balanced pool of oligonucleotides, i.e. baits, in the exome kit, however usually the individual sequence details and molarities are not shared by the providers.

Our results suggest that with lower average coverage than the commonly used 100-120x (Jamal et al., 2013), Twist exomes will achieve a similar performance as other exome kits at higher coverage. We find that at 70x average coverage the sensitivity for SNV detection is hardly affected and that there is only a small effect on the sensitivity of detecting CNVs. In our experience the sensitivity of CNV detection is likely to be more dependent on the size and quality of reference cohort that is used for CNV detection. We verified that these results

are consistent for samples that are originally sequenced at 70x and for different tissues than blood. However, QC thresholds may be adjusted by considering the strong increase in the variants with score below 500 and missed CNVs due to the SVD-ZRPKM thresholds in Twist 70x samples.

One class of variants that was not considered here are mosaic variants. It is unavoidable that the detection of mosaic variants will suffer from reduced overall coverage and this could be a reason to sequence at higher coverage. However, mosaic variants are relatively rare, and the sensitivity to detect high level mosaic variants (>10% VAF) will not substantially decrease (Acuna-Hidalgo et al., 2015).

We estimate that by performing WES at only 70x average coverage compared to 120x a 40% reduction ((120-70)/120) in sequencing costs can be achieved. Depending on the price for library preparation and exome capture kit and an overall price reduction for WES of 20-30% could be possible. In addition, our results may be used to re-evaluate minimal average coverage thresholds, for clinical exome sequencing and lead to fewer resequencing of samples with insufficient coverage.

We also compared our results to LR-WGS data. Whereas we previously found that LR-WGS provides coverage in regions that are missed by short read sequencing (Pauper et al., 2021), we find that for coding regions on average LRS has slightly lower coverage than SR-WGS, although still better than WES. This may have to do with the novelty of the technology and may improve over time to surpass SR-WGS.

In conclusion we found that Twist exome capture represents a significant improvement compared to other exome capture techniques. Exome coverage of Twist is more uniform and consistent than other enrichment kits. Because of more uniform coverage distribution, a minimum average coverage of 70x will provide sensitivity to detect both SNVs and CNVs similar to 150x WES samples with other enrichment kits.

Acknowledgements
Not applicable.

Author Contributions
Study conception and design:BY, EK and CG; Sample collection and sequencing:TH, MN, RP, TR and HGY; Sample QC check and alignment: JCG; Data analysis: BY and EK; Manuscript writing: BY and EK; Manuscript revision:

CG, JH, AH and HGY; Project supervision:CG and AH. All authors approved the final version of the manuscript.

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

Ethical Approval
Data was analyzed anonymously.

Consent for publication
Not applicable.

Competing Interests
The authors declare no competing interests.

Supplementary Information
The following files can be accessed via this DOI:
https://doi.org/10.1186/s40246-023-00485-5

Additional file 1

SuppInf1_TablesandFigures.pdf

Supplementary Tables

Table S1 Initial mean coverage of all samples

Table S2 Initial mean coverage of all samples used in CNV analysis

Table S3 Overview coding bases that are not targeted by different enrichment kits, based on the extended target regions (+/-200bp) of manufacturers. Length of Ensembl is 35,123,365 bp and RefSeq is 33,879,640bp

Table S4 Coverage statistics of the samples after downsampling

Table S5 Overview of base pair coverage for RefGene and Ensembl coding regions (a) Mean and standard deviation of base pair coverage by at least 20x per platform (b) Mean and standard deviation of base pair coverage by at least 10x per platform

Table S6 Average Evenness of coding regions (RefSeq) for different platforms

Table S7 Overview of basepair coverage ratio for samples with different coverage levels (RefSeq coding regions) (a) Ratio of covered regions by at least 20x (b) Ratio of covered regions by at least 10x

Table S8 Overview of basepair coverage ratio by at least 20x for samples with different coverage levels for blood and tissue samples enriched with Twist (based on RefSeq coding regions)

Table S9 Percentage of the OMIM transcripts that are covered at certain level of base pair coverage ratio by at least 20x

TableS10 CNVs called for 20 Twist samples

Supplementary Figures

Figure S1 Scree plots of singular values generated with Conifer a. Scree plot generated for 20 Twist samples at 100xcoverage b. Scree plot generated for 20 Twist samples at 70x coverage c. Scree plot generated for 14 Twist samples with validated CNVs at 100x coverage d. Scree plot generated for 14 Twist samples with validated CNVs at 70x coverage

Figure S2 Overview of basepair coverage ratio by at least 20x for blood samples and tissue samples enriched with Twist

Figure S3 Missing variants in samples with average coverage 70x compared to 100x.

Figure S4 A) GATK quality scores of variants identified in 100x average coverage samples compared to 70xaverage samples. B) Zoom in of the plot in Afor scores smaller than 10000

Figure S5 CNVs can't exceed the threshold for samples in both 100x and 70x coverage levels

Figure S6 CNVs called by samples with 100x average coverage and not exceed threshold for 70x coverage level

Figure S7 Visual graphs for segmentally called CNVs

Additional file 2

SupInf2_Ensembl Coding Regions Missed by Kits.pdf
Ensembl Coding Regions Missed by Twist and Agilent V7 Kits

Additional file 3

Solve-RD consortium.docx
List of SolveRD consortium members with affiliations

3. Long-read trio sequencing of individuals with unsolved intellectual disability

Marc Pauper[1,7] Erdi Küçük[1,2,7] Aaron Wenger[3], Shreyasee Chakraborty[3], Primo Baybayan[3], Michael Kwint[1], Bart van der Sanden[1,4], Marcel R Nelen[1], Ronny Derks[1], Han G Brunner[1,2,5], Alexander Hoischen[1,2,6,8] Lisenka ELM Vissers[1,4,8], Christian Gilissen[1,2,8]

[1] Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands.

[2] Radboud Institute for Molecular Life Sciences, Radboud University, Nijmegen, The Netherlands.

[3] Pacific Biosciences, Menlo Park, CA, USA.

[4] Donders Institute for Brain, Cognition and Behaviour, Radboud University 6525 HR Nijmegen, The Netherlands.

[5] Department of Clinical Genetics, Maastricht University Medical Center, Maastricht, The Netherlands.

[6] Department of Internal Medicine, Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, Netherlands

[7] These authors contributed equally

[8] These authors contributed equally

3.1 Abstract

Long-read sequencing (LRS) has the potential to comprehensively identify all medically relevant genome variation, including variation commonly missed by short-read sequencing (SRS) approaches. To determine this potential, we performed LRS around 15x-40x genome coverage using the Pacific Biosciences Sequel I System for five trios. The respective probands were diagnosed with intellectual disability (ID) whose etiology remained unresolved after SRS exomes and genomes. Systematic assessment of LRS coverage showed that ~35 Mb of the human reference genome was only accessible by LRS and not SRS. Genome-wide structural variant (SV) calling yielded on average 28,292 SV calls per individual, totalling 12.9 Mb of sequence. Trio-based analyses which allowed to study segregation, showed concordance for up to 95% of these SV calls across the genome, and 80% of the LRS SV calls were not identified by SRS. *De novo* mutation analysis did not identify any *de novo* SVs, confirming that these are rare events. Because of high sequence coverage, we were also able to call single nucleotide substitutions. On average, we identified 3 million substitutions per genome, with a Mendelian inheritance concordance of up to 97%. Of these, approximately 100,000 were located in the ~35 Mb of the genome that was only captured by LRS. Moreover, these variants affected the coding sequence of 64 genes, including 32 known Mendelian disease genes. Our data show the potential added value of LRS compared to SRS for identifying medically relevant genome variation.

3.2 Background

In the last decade short-read sequencing (SRS) approaches, such as whole exome sequencing (WES) and more recently whole genome sequencing (WGS), have revolutionized the field of medical genetics. Especially for clinically and genetically heterogenic disorders, such as intellectual disability (ID), WES has become the method of choice, allowing the identification of the underlying genetic defect in 40-60% of patients (Vissers et al., 2016). A substantial fraction (25%–30%) of the diagnostic success is due to recent progress in the discovery of new genes underlying disease (Farwell et al., 2015; Vissers et al., 2017; Y. Yang et al., 2014).

It has however been shown that, due to technical limitations, SRS approaches often lack sensitivity and specificity for a large proportion of structural variants (SVs) (Huddleston et al., 2017; Sedlazeck et al., 2018; Tattini et al., 2015). These limitations can be overcome by long-read sequencing (LRS). For instance, recent LRS studies have revealed that each human genome harbors thousands of SVs, in total spanning more than 10 Mb, that have largely remained undetected with conventional SRS (Eid et al., 2009; Huddleston et al., 2017; Pendleton et al., 2015; Seo et al., 2016; Shi et al., 2016). In addition, LRS of a haploid human genome resulted in SV call-sets three to sevenfold larger than those produced by standard SRS studies such as the 1,000 genomes project (M. J. Chaisson et al., 2015). This makes SVs an even more important source of human genome variation than anticipated so far, accounting for the greatest number of divergent bases across the human genome (Alkan et al., 2011; Weischenfeldt et al., 2013).

Structural variants, in particular copy number variations (CNVs), have long been recognized as an important cause for severe human diseases (Carvalho & Lupski, 2016; Cooper et al., 2007, 2008). As inversions and translocations were more difficult to study in a genome-wide fashion with good resolution, their biological and clinical relevance is likely underestimated so far (Escaramís et al., 2015).

There have now been some examples of studies in which LRS was applied to individual patients to resolve the genetic origin of disease (Merker et al., 2018; Reiner et al., 2018) (see also Mantere et al. (Mantere et al., 2019) for an overview). In addition to higher yield of structural variants, these studies indicate that LRS allows researchers to study genomic regions that are often challenging to sequence with SRS (Ebbert et al., 2019). Therefore, we hypothesized that LRS

may also enhance clinical diagnosis in an unbiased and genome-wide fashion for patients whose genetic etiology remained elusive after SRS WES and WGS approaches. To test this hypothesis, we here use a trio-based LRS approach for 5 individuals with unresolved ID and their parents (Gilissen et al., 2014), and compare LRS and SRS results to determine the added value of LRS for identifying all medically relevant genome variation in a single experiment.

3.3 Materials and Methods

3.3.1 Patient inclusion

Five patients with severe ID and their parents were selected for this study. All 5 patients were born to non-consanguineous parents with a negative family history. All were diagnosed with severe developmental delay, and co-morbidities, including epilepsy and/or behaviour problems. In addition, three of them showed facial dysmorphisms frequently observed in patients with genetic disorders (Supplementary Clinical Notes). Prior testing to detect the genetic cause of disease included genomic microarray (Vulto-van Silfhout et al., 2013), exome sequencing(De Ligt et al., 2012), genome sequencing (Gilissen et al., 2014) and methylome analysis (Barbosa et al., 2018), which had not resulted in a molecular diagnosis (**Table S1**). This study was conducted and approved by the Institutional Review Board of the Radboud university medical center (2017-3831).

3.3.2 Genomic DNA extraction, shearing and library preparation for long-read sequencing

Genomic DNA was extracted from whole blood using the Qiagen (Hilden, Germany) Puregene Blood Core Kit C. gDNA integrity was assessed with pulsed field gel electrophoresis (PFGE) 115 ng/well, 17 h run time at 70 V (**Figure S1**). gDNA was sheared with the Diagenode (Liege, Belgium) Megaruptor using long hydropores. A total of 12 µg gDNA was sheared to 60 Kb fragments in a total volume of 300 µl using the pre-installed settings. DNA was concentrated using 0.45x bead/sample ratio of Ampure PB beads and was eluted in 73 µl elution buffer. Qubit dsDNA BR assay was used to quantify DNA concentration.

All libraries were prepared using SMRTbell™ Template Prep Kit 1.0, according to the Procedure & Checklist – Preparing >30 Kb SMRTbells™ (Pacific Biosciences, Menlo Parc, CA, USA). As 10 µg DNA input was used instead of 5 µg, all reaction volumes were doubled until the size-selection step. DNA was sheared using the Megaruptor®, after which size selection was performed

using the BluePippin high-pass DNA size selection with 0.75% DF marker U1 high-pass 30-40 Kb v3 cassette. The range selection mode was set from 25-80 Kb. After size selection, Ampure PB bead cleanup steps were performed using 1x bead/sample ratio. DNA damage repair after size selection was performed with the reaction volumes described in the protocol. Qubit dsDNA HS assay was used to quantify DNA concentration.

3.3.3 Long-read sequencing

Sequencing primer v3 was annealed to the SMRTbell™ library. Polymerase was bound using the Sequel Binding Kit 2.0. SMRTbell™ complexes were purified using the Procedure & Checklist – Sample Clean-up using MicroSpin™ columns S-400 for diffusion loading. Sequencing reaction was performed using the Sequel sequencing plate 2.0 on a SMRT-cell 1M chip. On plate sample concentration was 10 pM, movie time was set to 600 minutes with an immobilization time of 120 minutes.

For 4 trios (Trio 1, 2, 3, 4) sequencing was performed at Radboud university medical center using a Pacific Biosciences (PacBio) Sequel I System. In total 145 SMRT Cells 1M were used, resulting in an average genome coverage of approximately 15x (**Table S2**). A single trio (Trio 5) was sequenced using a Sequel I System at Pacific Biosciences to an average genome coverage of 40x using 89 SMRT Cells 1M (**Table S2**).

For analyses on the percentage of the genome with specific fold-coverage we split the genome into "easily accessible" and "difficult" regions, and made the percentage calculations disregarding the latter. As "difficult" regions we defined those annotated as "scaffold", "contig", "clone", "telomere", "centromere" or "heterochromatin" in the GRCh38 assembly. Additionally, chrY was included in the "difficult" regions to enable better comparison between samples of different gender.

3.3.4 Variant calling from long-read sequencing data

For each trio, long reads were aligned to the GRCh38 reference genome (version GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set, with all non-primary contigs concatenated), using minimap2 (2.11-r797) with parameters "-a --eqx -L -O 5,56 -E 4,1 -B 5 --secondary=no -z 400,50 -r 2k -Y" (H. Li, 2016) . Structural variants were called using PBSV (https://github.com/PacificBiosciences/pbsv) (version 2.1.0) with default parameters for both the "discover" and "call" steps of PBSV's workflow. The "call" step was run jointly on

all 15 samples. PBSV is most effective for insertions with sizes from 50 bp to 5 Kb, deletions with sizes from 50 bp to 100 Kb and inversions with sizes from 200 bp to 5 Kb. Therefore, here we only considered variants of 50 bp and above as SVs. We annotated our SV calls using AnnotSV (version 2.0) (Geoffroy et al., 2018).

Single nucleotide substitution variants (SNVs) were identified with Longshot (0.2.0) (Edge & Bansal, 2019) using default parameters except for parameter max_cov, which was set to 50 for Trios 1, 2, 3 and 4 and 100 for Trio 5, in order to improve analysis runtime. The output was annotated with Annovar (2018-04-16) (K. Wang et al., 2010), using the core annotate_variation.pl script with RefGene hg38 build. We then compared VCF files for identifying SNVs that are uniquely detected by LRS or SRS using a custom script.

Read and mapping metrics were obtained from the aligned BAM files using SAMtools, Qualimap (Okonechnikov et al., 2015) and manual processing (**Table S2**). The number of sequenced bases was obtained from the "total length" field of the SAMtools' stats subcommand output. From the same output, the field "bases mapped" was divided by "total length" and by 3×10^9 in order to estimate the percentage of bases mapped and the mean coverage, respectively. The mean, median and N50 of read lengths were calculated by filtering out non-primary alignment, duplicate reads and supplementary alignments from the BAM files, and then manually processing the remaining mapped reads using awk and R. The mean mapping quality and error rate were calculated using Qualimap.

3.3.5 Short-read sequencing and variant calling

Whereas all five trios were previously exome and genome sequenced using SRS technology (**Table S1**) (De Ligt et al., 2012; Gilissen et al., 2014), Trio 5 was re-sequenced using 2x150bp paired-end reads on the Illumina NovaSeq 6000 instrument to an average coverage of 29x to allow for comparison of LRS to today's standard whole genome SRS. Reads were aligned using BWA mem (version 0.7.12-r1039) and SNVs were called using the xAtlas caller with default parameters (version 0.1) (Farek et al., 2018), and only variants passing all filters were retained. Structural variants were identified using three different variant callers: Manta (version 1.1.0) (Chen et al., 2016), LUMPY (version 0.2.13) (Layer et al., 2014) and DELLY (version 0.7.8) (Rausch et al., 2012). The output of LUMPY and DELLY were genotyped using SVTyper (version 0.6.0) (Chiang et al., 2015). Structural variants smaller than 50 bp in size were filtered out, and minimum genotype quality was set to 20.

3.3.6 Quality assessment of SVs and SNVs

For each trio we quantified the Mendelian inheritance errors in SV calls in the LRS data. For this analysis we used only deletions, insertions and inversions. Mendelian inheritance errors (MIEs) were quantified using the plugin "mendelian" from BCFtools (version 1.9). For comparison, we performed the same analysis on SVs detected with SRS in Trio 5.

Furthermore, for each proband, we identified SVs from LRS that were uniquely found in that proband and not in any of the other probands. For these SVs, we looked for MIE using BCFtools (version 1.9) plugin "mendelian". Similarly, we performed a Mendelian concordance analysis of SNVs with vcftools –mendel option. Only filtered SNVs on autosomal chromosomes were considered for this analysis. Specifically, for Trio 5, we analyzed both Longshot calls from LRS as well as xAtlas calls from SRS.

3.3.7 Comparison of sequence coverage between short and long-read sequencing

In Trio 5, we identified genomic regions that had no coverage with SRS but were well covered by LRS using BEDtools genomecov (version 2.25.0) (Quinlan & Hall, 2010). For this, we considered only reads with a minimum mapping quality of 10 and used Gencode Basic gene set, release 21 to define genic and exonic regions. We determined what percentage of these regions without coverage are in telomeric and centromeric sequences. We obtained a set of telomeric and centromeric regions from UCSC genome browser and used BEDtools intersect (2.25.0) to find overlaps.

We also determined genes that were not well-covered by SRS using these Gencode annotations. A gene was considered not well-covered when at least 10% of its length was not covered by SRS. We then compared the GC-content of these genes to a randomized set of well-covered genes with comparable lengths. GC-percentages of all genes were obtained from the UCSC Genome Browser.

3.3.8 Comparison of structural variants

The SV callsets from Manta, LUMPY, and DELLY from SRS data were each compared to the LRS SV calls of Trio 5. The comparisons were run on large (≥ 50 bp) deletions, insertions and inversion, using Truvari (version 1.3.2). Truvari matches SVs between different datasets. For a match, the maximum distance between start and end coordinates of two SVs was set to 1 Kb with parameter –refdist. Additionally, by default, two calls should be of the same

type and the ratio of the size of the smaller call over the larger call should be at least 0.7 for a match. The parameter --pctsim was set to 0 as it can only be applied on sequence-resolved variants. Parameter --sizemax was set to 100 Mb to circumvent the default 50 Kb.

The SV call sets of Manta, LUMPY and DELLY were also compared to each other to examine the concordance between the three methods. The comparison was done with Truvari (version 1.3.2) for all SV types and sizes, using similar parameters as described above.

We also compared our SV call set with two published datasets that were also produced by PacBio instruments:

We used SV calls available from Pacific Biosciences for the HG002 reference sample from the Genome in a Bottle (GIAB) consortium, sequenced on a PacBio Sequel instrument at 10x coverage (https://downloads.pacbcloud.com/public/dataset/HG002/Sequel-201810/). The comparison was done using Truvari for deletions, insertions and inversions larger than 50 bp and passing all filters (Wenger et al., 2019).

In addition, we compared our LRS SV dataset to the results obtained by Audano *et al*. (Audano et al., 2019) based on long-read WGS sequencing of 15 individuals across different populations. As Audano *et al*. only supplied SV calls in bed format, the comparison for overlap was performed using BEDtools (2.25.0) using a 50% reciprocal overlap setting (https://ars.els-cdn.com/content/image/1-s2.0-S0092867418316337-mmc1.xlsx).

3.3.9 Identification of *de novo* SVs

For each trio, potential *de novo* SVs were identified based on the genotyped SV calls of PBSV. Initially, we selected all variants with a heterozygous genotype (e.g. one wildtype allele and one allele with SV) in a proband, and a homozygous reference genotype (e.g. two wildtype alleles) in both parents. These variants were then subjected to additional filtering in order to identify high quality *de novo* candidates:

- Heterozygous alternative allele only in proband
- Read ratio supporting alternative allele between 0.3 - 0.7 in proband
- Minimum depth of coverage of 6 reads at SV coordinates for all samples
- Homozygous for the reference allele in all other samples
- Zero reads supporting alternative allele in other samples

Each candidate *de novo* SV was visually inspected using the Integrated Genomics Viewer (IGV, version 2.4.14) (Robinson et al., 2011), which adds better support for visualizing LRS, such as grouping and coloring of alignments based on ZMW or sequencing movie name, and allows for better identification of false positive SV calls due to sequencing artifacts. The subset of variants passing visual inspection was sent for PCR validation.

3.3.10 Identification of *de novo* SNVs

For each trio, potential *de novo* SNVs were identified from LRS data using VCFtools (0.1.13) –mendel option, which produces a list of Mendelian inheritance errors. We selected a quality score cutoff of 30 for missense variants such that all previously identified *de novo* mutations were included in the result. Filtered missense variants and all loss-of-function variants were visually inspected in IGV before being sent for validation with Sanger sequencing.

3.3.11 Recessive inheritance analysis

We used a custom script to parse AnnotSV annotations and identify genes that are affected by homozygous or compound heterozygous SVs and SNVs affecting the coding regions in all trios. Only loss-of-function SNVs with quality score higher than 30 were considered for this analysis.

3.3.12 Validation of SV and SNV events in LRS

Candidate *de novo* SVs and SNVs were visually inspected in the BAM files of the patient as well as both parents by using the Integrative Genomics Viewer (IGV). Based on the examination of the mode of inheritance, read quality, and mapping quality, each variant was classified as follows:

1. Inherited variant: even though the variant was classified as possible *de novo* event, the BAM files showed that the variant was also present in one of the parents (e.g. missing call in the parental data).
2. False positive variant: the quality and mapping of the reads at the region of the variant was substandard.
3. The variant appears as a true *de novo* event.

The remaining candidate *de novo* SVs were subsequently validated by breakpoint spanning PCRs and evaluation by Agarose Gel Electrophoresis. Primers were designed using Primer3. PCRs were performed by using Amplitaq Gold 360 Master Mix (Thermo Fisher Scientific) or Phusion Hot Start (Finnzymes) both according to the manufacturer protocol. The agarose gels were visually

inspected to assess whether the SVs appeared to be genuine *de novo* events. Hereto, it firstly needed to show a second PCR product representing the variant allele, next to the product of the expected size for the wildtype allele, and secondly, the second PCR product was only to be present in the proband and absent in the respective parents, indicating a *de novo* event.

The SNVs retained as potential *de novo* events were validated using Sanger sequencing. Primers were designed using Primer3Input. PCRs were using Amplitaq Gold 360 Master Mix (Thermo Fisher Scientific) according to the manufacturer protocol. PCR products were enzymatically cleaned by using Exonuclease I and FastAP, after which samples were Sanger sequenced. Finally, Sanger sequencing traces were analyzed using the VectorNTI software package (Thermo Fisher Scientific).

3.3.13 Titration analysis of LRS

Titration analysis was performed by subsampling *in silico* the LRS data of samples from Trio 5. Subsampling was done on the BAM files using SAMtools view (version 1.6) and the –s parameter to pass the desired fraction of data to retain. The original LRS data was subsampled to coverages 30x, 20x, 15x, 10x and 5x. Each subsampled BAM file was then used for SV discovery with PBSV (version 2.1), and SV calling was performed jointly for the trio at each coverage.

The SV call set of Trio 5 proband at each coverage was compared to the SV calls of the original full coverage dataset. The comparison was performed using Truvari (version 0.4) with parameters: --multimatch, --passonly, --pctsim 0 and --refdist 1000. Parameters --sizemin and --sizemax were set to 50 and 1,000,000.

We performed a similar analysis for SNV calling and used the same subsampled BAM files as described above, and ran Longshot to call SNVs at different coverages. We then used the original full coverage SNV calls as a truth set to calculate precision and recall values for each level of coverage.

3.4 Results

3.4.1 Long-read WGS characteristics

Five patient-parent trios were subjected to LR-WGS. From all sequenced reads per sample, the read mapping rate varied between 94.5% and 98.7% (**Table S2**). Whereas the longest read obtained was 60 Kb in size, the average read length was 9.5 Kb (N50 average: 17481.5 Kb; **Figure S2**). The error rate was consistent across samples (0.16 errors per aligned base), which is in line with what is reported in the literature (Weirather et al., 2017). Sequencing and mapping resulted in an average coverage of 16.01x for Trios 1, 2, 3 and 4, and 39.77x for Trio 5 (**Table 1**). On average 99.9% of the easily accessible genome regions were covered at least 5x and more than 85.4% of the complete genome including decoy sequences, centromeres, assembly gaps and chromosomes Y and M (**Tables S2 and S3**).

3.4.2 Structural variation across the cohort

We identified an average of 28,292 SVs (≥ 50 bp) per sample for Trios 1, 2, 3, 4. For Trio 5, sequenced at higher depth, we identified 33,157 SVs (**Table 1**, **Table S4a, Figure S3**), suggesting that greater sequencing depth enhances sensitivity for SV detection. Across all 15 samples, we identified 55,025 unique SVs, including 34,690 insertions (63%), 20,307 deletions (37%), and 28 inversions (0.05%) (**Table S4b**). There was a gradual decline of SVs abundance with increasing size, with smaller SVs being more abundant than larger SVs. Exceptions were noted for SVs at ~300 bp, representing *Alu* short interspersed nuclear elements (SINEs) and those at ~6.4 Kb representing LINE1 long interspersed nuclear elements (LINEs; **Figure S4a**, **Figure S4b**).

All detected SVs affected in total approximately 13 Mb of genome sequence per sample (deletions: 6.5 Mb, insertions: 6.4 Mb and inversions: 35.8 Kb, **Table S5**). On average, 173 SVs per individual affected the coding sequences, including a total of 81 genes with a known disease relationship according to OMIM (**Table S6, Figure 3**). However, none of these SV events could be linked to the patient phenotype. For Trio 5, we performed sub-sampling of the LRS data to determine the effect of coverage on SV discovery. As expected, we found that SV yield increases with coverage, but that the number of additional SVs diminishes beyond 10X.

Table 1 Overview of samples, sequencing statistics, identified variation and Mendelian inheritance errors.

Sample	Coverage (x)	# SVs	Total affected sequence (bp)	SVMIE (%)	Unique SVs in cohort	Study-specific SVs	SNV	SNV MIE (%)
T1P	15.7	29,030	12,775,459	4,409	75	18,601	3,142,916	448,808
T1F	12.6	26,475	11,340,949	(15.2%)	245	14,172	2,663,675	(14.3%)
T1M	14.9	27,421	12,998,704		236	18,334	3,007,973	
T2P	17.3	28,766	12,974,306	3,186	81	19,098	3,383,890	290,676
T2F	14.2	27,412	12,595,695	(11.1%)	244	19,549	3,127,690	(8.6%)
T2M	17.2	28,649	12,851,920		253	18,539	3,353,748	
T3P	15.0	28,111	12,689,520	2,681	57	19,321	3,147,660	219,392
T3F	18.3	28,971	13,327,881	(9.5%)	247	19,198	3,382,461	(7.0%)
T3M	18.0	28,962	12,884,749		264	19,419	3,352,767	
T4P	16.1	28,640	12,830,123	2,763	33	19,470	3,166,126	301,415
T4F	16.7	28,746	12,540,739	(9.6%)	261	21,653	3,261,812	(9.5%)
T4M	16.1	28,322	12,683,375		256	19,523	3,101,729	
T5P	41.6	33,056	14,085,392	2,130	16	21,539	3,956,435	125,023
T5F	37.6	33,277	14,235,204	(6.4%)	228	18,974	3,905,927	(3.2%)
T5M	40.1	33,138	13,949,579		273	22,023	3,932,800	

Columns from (left to right) indicate: sample identifier, average coverage across the genome (GRCh38), number of identified SVs (>=50 bp), total number of bases affected by SVs, number of SVs in proband with a Mendelian inheritance error (% is indicated below), number of SVs only occurring in this sample, number of SVs only found in this study (not in HG0002 and Audano *et al.*), number of identified SNVs, number of SNVs in proband with a Mendelian inheritance error (% is indicated below). F: father; M: mother; P: proband; SV: structural variant; MIE: Mendelian Inheritance Error; SNV: single nucleotide variant.

3.4.3 Overlap with published datasets

We compared our results to other published LRS datasets based on the GRCh38 reference genome, and sequenced using Pacific Biosciences Sequel instrument. First, we used SV calls from the HG002 reference sample that was sequenced at 10-fold depth (Wenger et al., 2019). We found that from the HG002 dataset 77.4% (n=5,920) of the deletions and 73.9% of the insertions (n=7,043) were also detected in our dataset, showing a high degree of overlap (**Table S7**). We also compared our SV calls with the published study of Audano *et al*. (Audano et al., 2019), whose work included the genomes (50x coverage) from 13 diploid individuals, the majority of whom were of non-European descent. Almost 80% of the variants they reported were not previously published to that date. We found that 33% of deletions (n=4,526) and 73% of insertions (n=15,963) in our cohort were novel compared to Audano *et al*. (**Table S8**), highlighting the fact that there is still a large degree of previously hidden structural variation to be identified in the human genome.

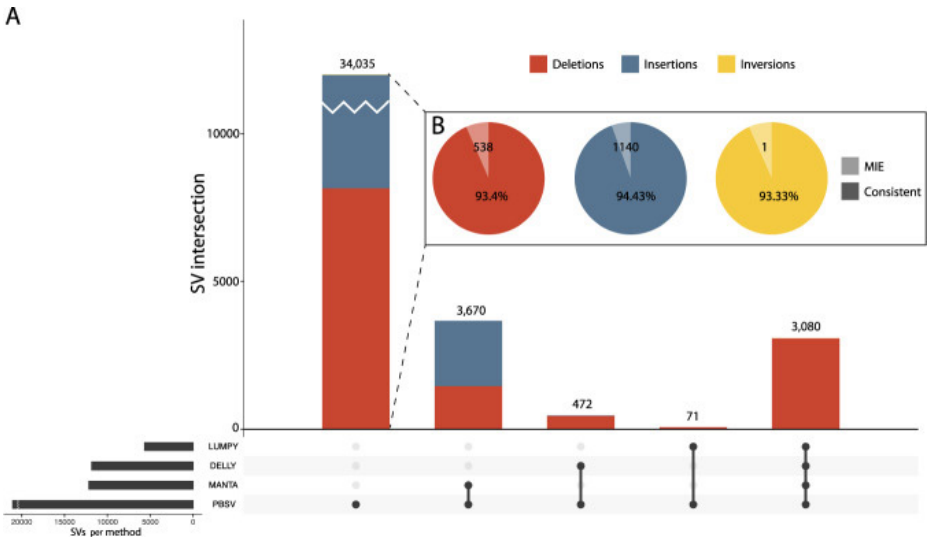

Figure 1. Comparison of structural variants called with long-read sequencing and short-read sequencing. A) comparison of structural variants identified in Trio 5 between long-read sequencing using PBSV and short-read sequencing using three algorithms for structural variant detection. The plot depicts the number of different structural variants that were identified by each combination of methods, indicated below the corresponding bar. Deletions in red, insertions in blue and inversions in yellow. The bottom left bar plot depicts the total number of SVs identified with each method. **B)** Pie charts show the number ofMendelian inheritance errors for the three types of SVs identified by LRS and the percentage of concordant calls.

3.4.4 Comparison of SRS and LRS structural variation

In order to compare the performance of LRS and SRS, Trio 5 was sequenced on an Illumina NovaSeq 6000 instrument to an average coverage of 29x (**Tables S9 and S10**). SV calls for SRS were obtained by three different calling algorithms: Manta, LUMPY and DELLY. We compared each of the SRS SV call sets separately to the SVs from LRS, considering only deletions, insertions and inversions (**Figure 1**). Between 51-78% of deletions identified in SRS were detected in LRS whereas only 25-38 % of deletions in LRS were detected in SRS by any of the three different calling algorithms (**Table S11**). For insertions, 83-91% of calls detected in SRS were detected in LRS but only 1-9% of insertions from LRS were also detected in SRS. The large differences in concordance between the results of the three different SRS SV calling algorithms and SVs from LRS emphasizes the challenges of SV detection based on SRS.

3.4.5 Quality of SV calls based on Mendelian inheritance

Our trio-based sequencing design allowed us to assess the Mendelian inheritance of SV calls. We define Mendelian inheritance errors (MIEs) as SNVs in a proband that could not have been inherited from either parent, resulting in a genotype that is inconsistent with Mendelian transmission. MIEs are commonly attributed to erroneous sequencing calls (Pilipenko et al., 2014). Conversely, proper Mendelian inheritance of SVs lends additional support to their reliability. We found that more than 90% of SV calls were concordant with Mendelian inheritance within Trios 2, 3 and 4 (**Figure 2, Table S12**). We obtained a lower concordance (87%) for Trio 1, likely due to a lower coverage in the father (12.6x) and, consistent with this, highest concordance in Trio 5 (96%) that was sequenced at higher coverage (39x). Moreover, in comparison to SRS of the same samples (Trio 5), LRS had a comparable overall percentage of Mendelian inheritance errors (MIE) to SRS (~5%). If we consider Mendelian concordant SVs detected by either technology as our truth set, then LRS has almost five times higher sensitivity than SRS (93% *vs.* 19%, respectively).

The high quality of the SV calls is also apparent from the unique SV events that were only identified within a single trio in the proband, but in none of the other trios. Mendelian inheritance concordance for unique deletions was 90.7%, similar to the overall concordance (**Table S13**). However, for unique insertion events concordance was only 76.0% suggesting that detection of these events is more challenging in LRS data.

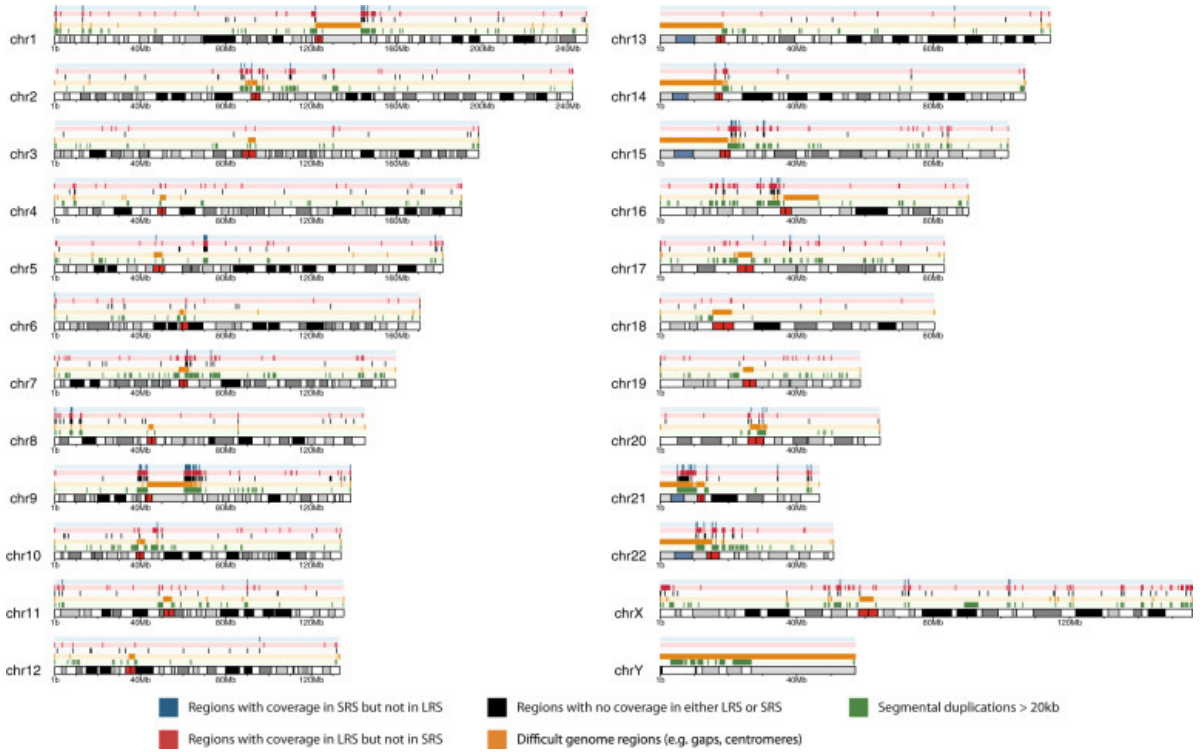

Figure 2. Schematic digital ideogram depicting genomic regions larger than 1 Kb without LRS or SRS coverage. From top to bottom tracks indicate: Regions with sequence coverage in SRS but no coverage in LRS (blue); Regions with sequence coverage in LRS but no coverage in SRS (red); Regions with no sequence coverage in neither LRS nor SRS (black); Genome regions that are difficult to assess like centromeres, telomeres and gaps (orange); Regions with segmental duplications larger than 20 Kb (green). Note that blue, red, black and orange regions are mutually exclusive but that suggestive overlap from the figure is due to the limited resolution.

3.4.6 *De novo* SV discovery

De novo mutations are a well-known cause of ID (Vissers et al., 2016). We therefore set out to filter our dataset for *de novo* SVs based on SV calling genotypes and minimal quality criteria. On average, this led to the identification of 10 candidate *de novo* SVs per trio (range 2-17; **Table S14**). After visual inspection of read mappings, the number of candidates was manually curated to a set of 8 possible *de novo* SVs ranging between 0 and 3 *de novo* candidates per trio (**Table S15, Figure S5**). Notably, the candidate *de novo* SVs that were removed after inspection were mostly false positive SV calls due to repetitive sequence, or inherited as a consequence of a missed SV call in one of the parents. In accordance with the hypothesis that increased sequencing coverage results in increased specificity, the lowest number of *de novo* candidates was found in the high coverage Trio 5.

3.4.7 *De novo* SV validation

Systematic validation of the 8 potential *de novo* SVs left in trios 1-4 after visual inspection and follow-up by breakpoint spanning PCRs, 4 SVs were confirmed as genuine SV events, for 2 others the PCRs remained inconclusive (non-unique or no PCR product), and 2 SVs were a likely false positive call from LRS SV data. None of the genuine SV events were however of *de novo* origin (**Table S15**), but rather confirmed as parentally inherited SVs which were likely missed due to low coverage (**Figure S5**).

3.4.8 Single nucleotide variants

In addition to SVs, our sequencing depth allowed us to identify SNVs in all five trios from LRS data. We identified, on average, 3.33 million substitutions per genome, of which 23,672 were located in the coding regions (**Table S16**). Detailed comparison for Trio 5 for all SNVs called from SRS and LRS showed substantial overlap. Over 95% of SNVs identified in SRS were also identified in LRS. In the coding regions, the overlap was more pronounced, reaching 97% (**Table S17**). Furthermore, we looked at the transition/transversion ratio (Ti/Tv) of called SNVs, which is often used as a metric to detect biases. SNVs called in both LRS and SRS data had a T/Tv ratio of 2.1, which is in line with expectations regarding WGS data (J. Wang et al., 2015), suggesting that most variants are genuine biological events (**Table S17**). In contrast, SNVs uniquely detected by SRS or LRS had Ti/Tv ratios of 1.17 and 0.99 respectively, which may indicate a higher degree of false positive calls in these sets. Evaluation of MIE rates for SNVs on LRS data showed that for Trio 5, with 40x coverage, the MIE rate was as low as 3%, while for Trios 1-4 it was around 8% (**Table S18**). The MIE rate in the

SRS data of Trio 5 was only 1.5%, almost half of that obtained for the LRS, which demonstrates the overall higher accuracy for SNV calling in SRS.

We overlaid the identified coding SVs with the coding SNVs called in each individual in order to find potential compound heterozygous SV-SNV pairs affecting the same gene and potentially explaining recessive disease inheritance. In total, we found 13 total SV-SNV pairs, but none were likely causal for the patient phenotype (**Table S19**).

3.4.9 *De novo* SNV discovery

Because we could have potentially missed *de novo* SNVs in poorly covered regions when we performed exome and genome SRS (De Ligt et al., 2012; Gilissen et al., 2014), we also identified potential *de novo* SNVs in all trios based on the LRS data. For this analysis, we used a minimum quality score of 30 as a threshold, as all previously identified *de novo* point mutations (DNMs) from SRS had scores above this threshold in the LRS data (**Tables S1 and S20)**. We considered missense and loss-of-function mutations as potentially damaging *de novo* candidates. This resulted in 67 candidate DNMs across all 5 trios (**Table S21**), in addition to the 6 *de novo* variants reported previously (De Ligt et al., 2012) (**Table S1, Table S20**).

3.4.10 *De novo* SNV validation

Routine Sanger sequencing validations confirmed 58 of 73 (79%) of those candidate variants. Confirmed variants had significantly higher average quality scores than variants that could not be confirmed in the proband (quality scores 67 versus 40 respectively, p=8.25e-4, Welch two-sample T-test). Similar to the LRS SV validations, all Sanger sequencing-confirmed variants, apart from the 6 already known *de novo* SNVs, appeared inherited from one of the parents.

3.4.11 Genome variation in previously uncovered regions

Long-read sequencing is expected to sequence across genomic regions that are difficult to assess using SRS. Therefore, we identified regions in the complete genome of Trio 5 proband that lacked sequence coverage in only one technology. Genome-wide we found that on average 191.7 Mb of the reference genome remains uncovered by both technologies, including 229.3 Kb of protein-coding sequence and 319.4 Kb telomeric and centromeric sequence (**Table S22**, **Figure 2**). We found an additional 35.2 Mb, including 634.9 Kb of protein-coding sequence corresponding to 105 genes, only covered in LRS data. *Vice versa*, only 12.5 Mb were uniquely covered by the SRS data,

including 20.1 Kb of coding sequence. We compared these results to Ebbert *et al.*'s (2019) (Ebbert et al., 2019) study on "dark" gene regions that cannot be adequately assembled or aligned using standard SRS. Of the 35.2 mb that is missed by the SRS data in our study, a substantial portion (67%) overlapped with the regions that are identified in Ebbert *et al.* (**Table S22d**).

Importantly, in the 35.2 Mb of LRS-only regions we identified on average 3,874 SVs and 32,540 high-quality substitutions in the proband, of which 50 and 672 respectively were overlapping with genic regions in the proband of Trio 5 (**Figure 3**). These 672 genic substitutions included 171 missense and 3 loss-of-function variants (**Table S23**) and occurred within 43 different genes, including two known genes associated with ID and four other OMIM morbid genes (**Table S24**).

An additional 378 genes, of which 26 have an established disease-association in OMIM, were only partly covered by SRS (no coverage for more than 10% of the coding sequence; **Table S25**) but were well-covered in LRS. As expected, we found that these genes had a higher GC content than genes that were well-covered by SRS (48.8% compared 46.3%, p-value < 2.2e-16 Welch two-sample T-test).

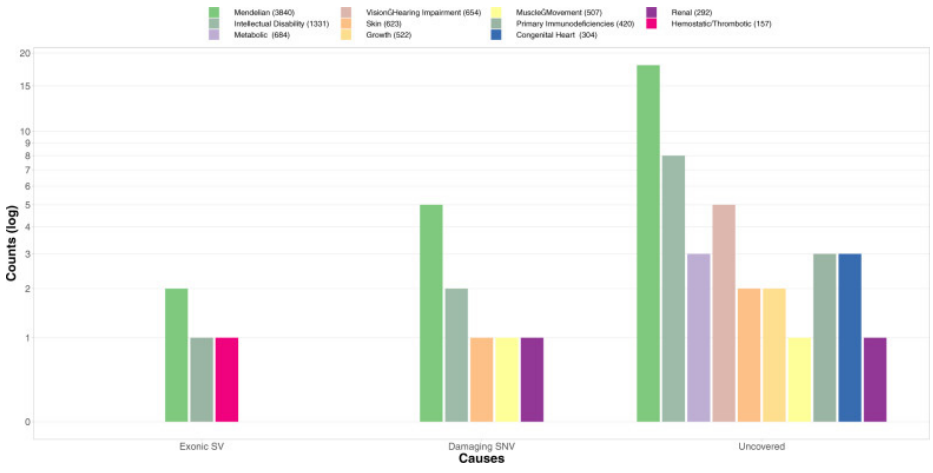

Figure 3. Interpretation of variation identified by LRS but not SRS. From left to right showing three groups: the number of genes of which coding regions are affected by an SV, the number genes affected by a putatively damaging SNV identified only in LRS, the number of genes of which coding regions are covered less than 10% in SRS but that do have coverage in LRS. Individual bars indicate different types of disorders based on diagnostic gene panels as used by Genome Diagnostics Nijmegen (https://www.radboudumc.nl/en/patientenzorg/onderzoeken/exome-sequencing-diagnostics/information-for-referrers/exome-panels), genetic testing laboratory. Numbers in the legend indicate the total number of genes in each of the gene panels.

3.5 Discussion

We performed long-read sequencing for 5 trios with unresolved intellectual disability and identified thousands of structural variants. We identified an average of 28,292 SVs, of which up to 93% was shown to adhere to Mendelian inheritance in 5 trios. We found considerable overlap with the HG002 published dataset, but slightly less overlap for insertions identified by Audano *et al*. We expect that this may be because Audano *et al*. used a different variant calling algorithm (SMRT-SV) whereas variants from HG002 were also called using PBSV. In addition, Audano *et al*. sequenced mostly individuals of non-European descent.

Although a substantial part of these SVs could not be identified by SRS using three different and commonly used calling algorithms, the significant overlap with existing LRS datasets and Mendelian inheritance concordance indicates that most of these events are likely true events. These SVs were not only present in the most repetitive regions of the genome, but also affected genes and coding regions. We also compared the SV datasets from the three different calling algorithms for SRS to each other, and found that the concordance among these algorithms is disappointingly low (**Table S26**). This is illustrative of the complexity of SV detection using SRS data, as also observed by others (Cameron et al., 2019) and suggests that LRS technology may be a prerequisite for reliable SV detection. In line with this we estimated that LRS has almost five times higher sensitivity for the detection of SVs than SRS. One striking observation is that the number of detected inversions in our cohort is relatively low, with 28 events in 5 trios compared to other studies that identified 156 inversions per genome (M. J. P. Chaisson et al., 2019). This suggests that sensitivity for these events could be improved, either with improved detection algorithms or alternative sequencing approaches such as Strand-seq (Sanders et al., 2017) or Bionano technology (Chan et al., 2018).

Notwithstanding the higher raw sequencing error rate of LRS, the relatively high sequence coverage for our samples, allowed us to call SNVs as well. Surprisingly, our SNV calls were of relatively high quality with MIEs as low as 3% for Trio 5, even without the use of Circular Consensus Sequencing technology (CCS) (Wenger et al., 2019). Most impressive is that all previously identified *de novo* SNVs through SRS were also identified as potential *de novo* SNVs in the LRS data, albeit with varying quality scores. SNV calling accuracy would have likely been improved significantly had we been able to

use CCS technology (Wenger et al., 2019). This shows the potential of LRS to comprehensively identify all genome variation in a single experiment.

We hypothesized a *de novo* structural variant, previously missed by microarray, exome and genome sequencing may be the cause for the disease in the five individuals with ID sequenced here (De Ligt et al., 2012; Gilissen et al., 2014). However, in this study we did not identify any *de novo* structural variants that could be confirmed by alternative techniques. There are several possibilities for our lack of confirmed *de novo* SV events. First of all, the original event in the proband may have been missed due to lack of sequence coverage. Based on an *in silico* titration and repeated SV calling, we find that although the quality and quantity of the SV yield increases with coverage, beyond 10x the increase in yield diminishes (**Table S27**).

Secondly, we found that the analyses for the identification of SVs is still being actively developed and results may change considerably depending on the calling algorithm, its version and settings and the reference genome version that is used. Improvements in read alignment and SV calling algorithms for LRS constitute a developing field and future reanalysis of our data may still identify genuine *de novo* SVs. However, even with such improvements some events may remain too complex to be reliably identified and different technologies may be required.

Thirdly, it is also possible that we have in fact identified *de novo* SVs but that the methods that were used to validate these events are not reliable enough to confirm such complex forms of genetic variation; for four events, no conclusions could be drawn, showing the need for robust validation methods in these complex genome regions. Lastly, the lack of an identified *de novo* SV may simply be because *de novo* mutation rate for SVs is very low. For instance, the current estimate for *de novo* mutations of large Copy Number Variants (CNVs), is as low as 0.2 events per genome per generation (Veltman & Brunner, 2012). For other SV types such estimates are less well-established, but it is not unlikely that these are as low as CNVs. Larger cohorts of trio-based LRS are required to fully capture the per generation *de novo* mutation rates of other structural events.

Alternatively, our initial hypothesis of a previously undetected SV as the cause for ID in these patients, may be incorrect, and the disorder is caused by other types of variants that have so far eluded detection. These could for example

be small insertion/deletion events, repeat expansions, or mosaic variation. Moreover, we may have identified the causative variants but have not been able to interpret them correctly.

Although LRS may indeed identify more variation, it is still unknown how much of this genome variation is clinically relevant, and thus how much of an advantage LRS offers over SRS for clinical WGS. Our results show that LRS identifies more SVs across the genome than SRS, some of which affect coding regions, and provides sequence coverage in difficult regions of the genome that harbor protein coding genes. Using LRS, we are also able to identify SNVs in these regions, and genes within these regions are part of virtual disease gene panels used in routine diagnostic labs for clinical exome interpretation. Whereas we did not identify clinically relevant SNVs in these genes, it is not unreasonable to speculate that there are patients out there that will benefit from variant calling in these regions only accessible by LRS. With the technological advances of CCS marketed as HiFi reads (Wenger et al., 2019) shortly, further enhancing robustness of SNV calling in LRS data, one may expect that genome-wide LRS may allow comprehensive analysis of all variant types per individual genome for clinical and genetic heterogenous disorders such as intellectual disability in the future.

Availability of Data and Materials

Data for all samples is accessible at EGA under accession number EGAS00001004319.

Declaration of Interests

AMW, SC, and PB are employees and shareholders of Pacific Biosciences, a company commercializing DNA sequencing technologies.

Acknowledgments

We thank the patients and their families for the participation in the study. We thank the Radboud Genomics Technology Center, Radboud University Medical Center Nijmegen, for their technical assistance.

Funding

This work is supported by grants of the Dutch Research Council (Aspasia 015.014.066 to LELMV) and ZonMW (846002003 to LELMV, 917-17-353 to CG). This work was supported by the Solve-RD project. 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 was part of the Netherlands X-omics Initiative and partially funded by NWO, project 184.034.019.

Web Resources

UCSC genome browser: https://genome.ucsc.edu/

Integrative Genomics Viewer (IGV): https://software.broadinstitute.org/software/igv/

Supplementary Information

Supplementary figures and tables can be accessed via this DOI:
https://doi.org/10.1038/s41431-020-00770-0

4. Comprehensive *de novo* mutation discovery with HiFi long-read sequencing

Erdi Küçük[1,2*], Bart P.G.H. van der Sanden[1,3*], Luke O'Gorman[1], Michael Kwint[1], Ronny Derks[1], Aaron M. Wenger[4], Christine Lambert[4], Shreyasee Chakraborty[4], Primo Baybayan[4], William J. Rowell[4], Zev Kronenberg[4], Han G. Brunner[1,3,5,6], Lisenka E.L.M. Vissers[1,3], Alexander Hoischen[1,2,7,*], Christian Gilissen[1,2,*]

[1] Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands

[2] Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands

[3] Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands

[4] Pacific Biosciences, Menlo Park, CA, USA

[5] Department of Clinical Genetics, Maastricht University Medical Center, Maastricht, The Netherlands

[6] GROW School for Oncology and Developmental Biology, Maastricht University Medical Center, Maastricht, The Netherlands

[7] Radboud University Medical Center for Infectious Diseases (RCI), Department of Internal Medicine, Radboud University Medical Center, Nijmegen, the Netherlands

[*] These authors contributed equally

4.1 Abstract

Background: Long-read sequencing (LRS) techniques have been very successful in identifying structural variants (SVs). However, the high error rate of LRS made the detection of small variants (substitutions and short indels <20bp) more challenging. The introduction of PacBio HiFi sequencing makes LRS also suited for detecting small variation. Here we evaluate the ability of HiFi reads to detect *de novo* mutations (DNMs) of all types, which are technically challenging variant types and a major cause of sporadic, severe, early-onset disease.

Methods: We sequenced the genomes of eight parent-child trios using high coverage PacBio HiFi LRS (~30-fold coverage) and Illumina short-read sequencing (SRS) (~50-fold coverage). *De novo* substitutions, small indels, short tandem repeats (STRs) and SVs were called in both datasets and compared to each other to assess the accuracy of HiFi LRS. In addition, we determined the parent-of-origin of the small DNMs using phasing.

Results: We identified a total of 672 and 859 *de novo* substitutions/indels, 28 and 126 *de novo* STRs, and 24 and 1 *de novo* SVs in LRS and SRS respectively. For the small variants there was a 92% and 85% concordance between the platforms. For the STRs and SVs the concordance was 3.6% and 0.8%, and 4% and 100% respectively. We successfully validated 27/54 LRS-unique small variants, of which 11 (41%) were confirmed as true *de novo* events. For the SRS-unique small variants we validated 42/133 DNMs and 8 (19%) were confirmed as true *de novo* event. Validation of 18 LRS-unique *de novo* STR calls confirmed none of the repeat expansions as true DNM. Confirmation of the 23 LRS-unique SVs was possible for 19 candidate SVs of which 10 (52.6%) were true *de novo* events. Furthermore, we were able to assign 96% of DNMs to their parental allele with LRS data, as opposed to just 33% with SRS data.

Conclusions: HiFi LRS can now produce the most comprehensive variant dataset obtainable by a single technology in a single laboratory, allowing accurate calling of substitutions, indels, STRs and SVs. The accuracy even allows sensitive calling of DNMs on all variant levels, and also allows for phasing, which helps to distinguish true positive from false positive DNMs.

Keywords: Long-read sequencing, HiFi reads, *de novo* mutations

4.2 Background

A comprehensive characterization of variation of individual human genomes is of great importance to gain insight into genetic traits and diseases (Lupski et al., 2011). For rare disease studies it is especially important to identify the full spectrum of all variant types, including substitutions, indels, short tandem repeats (STRs) and structural variants (SVs). A particular challenge for the accuracy of genomic technologies are *de novo* mutations (DNMs) (Acuna-Hidalgo et al., 2016; Veltman & Brunner, 2012), which have been shown to be a major cause of sporadic, severe, early onset disease (3,4). DNMs are mutations that arise in the germline of one of the parents during gamete formation and are transmitted to the offspring. Every human genome contains roughly between 40 and 90 DNMs on average (Veltman & Brunner, 2012). They are however also amongst the most challenging variants to identify, as DNM call sets typically contain large number of false positive calls due to sequencing artifacts, mapping artifacts, differences in sequence coverage and mosaicism (Acuna-Hidalgo et al., 2015; Kaplanis et al., 2020; Koboldt, 2020; Noyes et al., 2022; Pauper et al., 2021). Therefore, comprehensive detection of DNMs of all types demands the highest quality sequencing data.

Whereas short-read sequencing (SRS) can be used to accurately call small variants, such as substitutions and small indels (<20bp), the sensitivity to detect large STRs, copy number variants (CNVs) and SVs is limited as this can only be done by inference from systematic deviations in read coverage or read alignments (Alkan et al., 2011). Long-read sequencing (LRS) technologies typically generate sequencing reads of 10 to 100 kilo bases (kb) in size which offers many advantages compared to short-read sequencing (Mantere et al., 2019). Long-reads can interrogate regions of the human genome that are inaccessible by SRS and can encompass complete SV events thereby improving their detection (Logsdon et al., 2020; Marwaha et al., 2022). LRS has therefore been used extensively for *de novo* assembly of human genomes and for the characterization of structural genome variation that remains undetected by SRS (Mantere et al., 2019; Merker et al., 2018; Pauper et al., 2021; Reiner et al., 2018). However, LRS technologies have traditionally suffered from low accuracy at single base pair (bp) resolution, with a raw error rate of 8 to 15%, which did not allow them to reliably detect variants smaller than 50 bp. This reduced accuracy conserves the need to combine LRS with SRS to accurately detect the entire spectrum of *de novo* variation, which is accompanied by additional costs and time (Logsdon et al., 2020).

With improvements in LRS technology and specifically the recent introduction of PacBio HiFi reads, it is now possible to obtain high base call accuracy. With HiFi sequence reads, DNA templates of 10-30 kb in length are subjected to circular consensus sequencing (CCS) allowing to derive a consensus sequence of both strands of the insert region from multiple passes of the polymerase over a single template molecule (Logsdon et al., 2020; Vollger et al., 2020; Wenger et al., 2019).

The number of passes determines the accuracy of the consensus reads, since each pass allows for better error correction in the consensus sequence. HiFi reads are defined as reads with an accuracy of at least 99% (Phred quality score 20), theoretically resulting in the detection of substitutions and small indels being on-par with SRS technology (Wenger et al., 2019). HiFi technology has already been used to identify SVs in patients suffering from different genetic disorders, including synpolydactyly, syndromic intellectual disability, choroideremia and teratoid rhabdoid tumors (Fadaie et al., 2021; Melas et al., 2022; Mizuguchi et al., 2021; Sabatella et al., 2021). The increased base accuracy of HiFi sequencing should be especially advantageous for the detection of small DNMs and could even allow for improved sensitivity compared to SRS. Here, we investigated whether HiFi sequencing is sufficiently accurate to allow for the comprehensive detection of all types of *de novo* variation in parent-child trio genomes, which would remove the necessity to complement LRS by SRS and result in most comprehensive genomes.

4.3 Methods

4.3.1 Patient selection

We performed LRS on the Pacific Biosciences Sequel II instrument and SRS on the Illumina Novaseq instrument for 8 trios (24 samples; Supplementary Table 1). These trios are part of a previous study to identify disease-causing variants by exome sequencings and/or short-read genome sequencing to explain the neurodevelopmental disorder in the proband (Sanden et al., n.d.). In none of the 8 probands, a disease-causing variant was identified by SRS (van der Sanden et al., 2023). All participants or their legal representatives gave written informed consent. This study was approved by the Medical Review Ethics Committee Oost-Nederland and Radboudumc Institutional Review Board under 2020-6853, as part of 2018-4985 and 2014-1254.

4.3.2 Long-read sequencing and variant calling

For the LRS we targeted 30x HiFi coverage by using at least 3 SMRT Cells per sample (**Supplementary Table 1**). All samples were processed in the same fashion, according to the manufacturer's instructions (PacBio, Menlo Park, CA, USA). In brief, 5 μg DNA was sheared on Megaruptor 3 (Diagenode, Liège, Belgium) to a target size of 18 kb, libraries were prepared with SMRTbell express template prep kit 2.0 (PacBio, Menlo Park, CA, USA), size-selected >10 kb on the PippinHT (Sage Science, Beverly, MA, USA), and sequenced for 30 hours on the Sequel II system using Chemistry 2.0. HiFi reads were generated with CCS 4.2.0, and then processed using our in-house script which is available at https://github.com/PacificBiosciences/pb-human-wgs-workflow-snakemake.

Sequencing reads were aligned to the GRCh38/Hg38 genome with pbmm2 (version 1.4.0), using default parameters. Small variant (substitution and indel) calling was performed using DeepVariant (version 1.1.0) with default settings. No threshold for maximum size of the indels was applied and all indel calls were used for further analyses. STR calling was performed using Tandem Repeat Genotyper (TRGT; version 0.3.3) at 171,146 highly polymorphic repeat loci that are described in a tandem repeat catalog that is available together with the TRGT tool (https://github.com/pacificBiosciences/trgt/). SV calling was performed using PBSV (version 2.4.0) default settings with a minimum SV size of 20bp.

From the total variant call set we filtered *de novo* mutations (i.e. substitutions, indels and structural variants; DNMs) using slivar (Pedersen et al., 2021) (0.2.7) with two different sets of filter criteria. For the SVs we only applied the strict filtering criteria.

For the LRS strict and lenient lists we applied the following filters:

Parameter	Strict Filtering	Lenient Filtering
Proband genotype	0/1	0/1
Parental genotype	0/0	0/0
Parental alternative allele depth	0	<2 total
Proband allele depth	>5	NA
Reference allele depth	>10	NA
Total depth	<50	NA
Quality score	>30	NA
Genotype quality	>20	>10
Allele count in gnomAD and HPRC	<5	<5

For the STRs the output files were first filtered for loci for which all family members had both alleles genotyped. Subsequently, *de novo* STR expansions and contractions were selected using the number of repeat units of the two genotyped alleles. When the number of repeat units in one or both alleles of the patient were ≥2 repeat units longer or shorter than both parents, the repeat locus was considered *de novo*. Subsequently, we excluded *de novo* STR calls that were present in more than one patient of this cohort. Additionally, the repeat length had to be an outlier when compared to the alleles of all 23 other samples using the 1.5*interquartile range (IQR) rule. Finally, we excluded the *de novo* STR calls where one or both alleles had a TRGT quality score ≤ 0.8 (LRS) or where the repeat length of one or both alleles was outside the confidence intervals of ExpansionHunter.

4.3.3 Short-read Sequencing and Variant Calling

Short-read WGS was performed as described by the manufacturer (Illumina, San Diego, CA, USA), and in detail reported in van der Sanden et al. (van der Sanden et al., 2023). In brief, 1 µg DNA, isolated from whole blood, was used for library preparation using the Illumina TruSeq DNA PCR-free protocol, with an average insert size of 450 bp. To allow pooling of samples, barcoded indexing was included in the library preparation. Samples were pooled equimolarly on an S2 or S4 flowcell, prior to sequencing on an Illumina NovaSeq instrument to an anticipated genome-wide coverage of 50-fold, with a minimum of 45-fold.

After sequencing, FASTQ files were processed through our in-house pipeline for short-read genomes. Reads were mapped to the human reference genome (GRCh38/Hg38) using BWA (v.0.78) and the quality of the resulting BAM file was assessed using Qualimap (v.2.2.1). Variant calling was performed using various tools to optimize sensitivity per variant type. For calling small variants (substitutions and indels), GATK was used and no threshold for maximum size of the indels was applied. SVs were called using Manta Structural Variant Caller (v.1.1.0; Illumina), following a paired-end and split-read approach for SVs identification. No minimal size threshold was applied. CNVs were called using Canvas Copy Number Variant Caller (v.1.40.0; Illumina) using default parameters. STRs were called using ExpansionHunter (Dolzhenko et al., n.d.) using the same tandem repeat catalog containing 171,146 loci as for LRS.

For SRS we used two independent *de novo* callers, namely an in-house developed method based on Samtools mpileups and DeNovoCNN (Khazeeva et al., 2022). For the in-house method we first discarded all inherited variants and

variants with a gnomAD allele frequency >0.1% or GATK score <50. Remaining variants were run through the *de novo* caller which annotated the variants as inherited or possible *de novo* based on the Samtools mpileups. Subsequently, we only selected variants with ≥20% alternative allele depth, ≥5 alternative reads and ≤1% in-house allele frequency. For DeNovoCNN, inherited variants were also discarded, and the tool was run on the remainder of the variant list using default parameters, resulting in variants with a DeNovoCNN probability score >0.5. Variants called by both *de novo* callers were considered a true SRS DNM (SRS list). Variants only called by one *de novo* caller were listed separately (i.e., in-house unique list and DeNovoCNN unique list).

4.3.4 Variant annotation

Small variants from both LRS and SRS were annotated by an in-house pipeline. This variant annotation was performed using the Variant Effect Predictor (VEP V.91) and Gencode V.34 basic gene annotations. Frequency information was added from GnomAD V.2.1.1 and from an in-house database. In-house gene panel information was added for those genetic variants within a known disease gene.

SVs and CNVs were annotated using an in-house developed pipeline. This pipeline was based on ANNOVAR(K. Wang et al., 2010) and Gencode V.34 basic gene annotations. Additional frequency information was added from GnomAD V.2.1, 1000G V.8 and GoNL SV database.

4.3.5 Comparison of inherited variants in LRS and SRS

For comparing substitutions and indels between LRS and SRS, we used bcftools isec (version 1.8) (Danecek et al., 2021) to generate the intersection of two call sets. For the detection of Mendelian inheritance errors we used vcftools (version 0.13) with –mendel option (Danecek et al., 2011). For comparing structural variant call sets between two platforms, we used the Truvari (version 3.5) "bench" command using default parameters (English et al., n.d.).

4.3.6 Comparison of small *de novo* mutations in LRS and SRS

For the comparison of the small DNMs, we performed two analyses. First, we compared LRS small DNMs with SRS small DNMs. DNMs on the LRS strict list were overlapped with small DNMs on the SRS list. These mutations were marked as "overlap". Then, mutations were overlapped with the in-house unique and DeNovoCNN unique lists. Resulting variants were marked as "LRS+", while remaining mutations were marked as "LRS-unique".

Subsequently, we compared the SRS variants with the LRS variants. DNMs on the SRS list were overlapped with the small DNMs on the strict LRS list. These variants were marked as "overlap". Then, variants were overlapped with the lenient LRS list and resulting variants were marked "SRS+". Finally, remaining variants were marked as "SRS-unique".

4.3.7 Clustered small *de novo* mutations

During LRS small DNM analysis, we identified clusters of non-overlapping variants that fall in the same gene with approximately the same coverage and variant allele frequency. Since these variants were all unique to LRS and appeared inherited when checking the read alignment in IGV, we decided to systematically remove these clustered DNMs. In order to do this, we first selected LRS-unique variants separated per trio and ordered by the chromosome and genomic position. Then, variants in resulting lists were marked if the gene name was the same as the gene name of the previous and/or next variant on the list. The same was done for intergenic variants. Subsequently, clusters of DNMs were defined when two or more variants fell within one average read length from each other (Supplementary Table 2A). Clustered DNMs were excluded from further analyses.

4.3.8 Substitution and indel validation

All 54 LRS-unique, and 42 of the 133 SRS-unique variants were attempted to be validated using Sanger sequencing of proband, mother, and father. Primers were designed using Primer3Input. For 27 of the LRS-unique DNMs we were not able to design a primer set, and these were not further validated. PCRs for the remaining 27 LRS-unique and 42 SRS-unique small DNMs were performed using Amplitaq Gold 360 Master Mix (Thermo Fisher Scientific) according to the manufacturer's protocol. PCR products were enzymatically cleaned using Exonuclease I and FastAP, after which samples were Sanger sequenced. Finally, Sanger sequencing traces were analyzed using the SnapGene software package (version 5.2.2; GSL Biotech).

4.3.9 STR validation

For 18 LRS-unique and 18 SRS-unique STR calls we attempted validation using Sanger sequencing by the same approach as for the substitution and indel validations.

4.3.10 Structural variant validation

All 23 LRS unique SVs were validated using long-range PCR followed by sequencing on a PacBio Sequel IIe system. Primers were designed using

Primer3Input and PCR was performed using NEB LongAmp Hot Start Taq 2x Master Mix. For each PCR product 500ng was used as input for the library preparation and the normalized library was prepared according to manufacturer's instructions using the SMRTbell barcoded adapter complete prep kit. Finally, the library, with a loading concentration of 80 pm, was sequenced on a PacBio Sequel IIe system using a Sequel II SMRT Cell 8M (PacBio, Menlo Park, CA, USA) with a movie time of 30 hours and 0.7 hours pre-extension time.

4.3.11 Titration analysis

For the comparison of our data to LRS data with lower coverage, downsampling was performed using SAMTools v1.10 (H. Li, Handsaker, Wysoker, Fennell, Ruan, Homer, Marth, Abecasis, Durbin, & Genome Project Data Processing, 2009). Downsampling reduced the coverage of the samples from around 30x to 20x and 10x. On these samples with reduced coverage, *de novo* calling was then repeated as described above. We then compared these *de novo* calls to a truth set consisting of variants validated by either SRS or Sanger sequencing.

4.3.12 Phasing of small *de novo* mutations

For LRS, phasing was performed using WhatsHap (Patterson et al., 2015), using the default options with the '--indels' flag. Phased variants were considered informative for a *de novo* mutation if they are in the same phase block and were present in only one of the parents, while the other parent has homozygous reference call. Based on these informative variants, DNMs were classified as either paternal, maternal, or unknown, according to the following rules:

1. If fewer than 3 informative variants were present on a haplotype of the candidate DNM, the DNM was considered unknown.

2. If 3 or more informative variants were inherited from the same parent, then the DNM was assigned that respective parental origin. If more than 90% of the informative loci supported the same parental origin, the call was additionally classified as high-quality.

For SRS sequencing data, we used GATK Haplotypecaller (Poplin et al., 2018) to produce gVCFs. These were then combined with GATK CombineGVCFs tools and then genotyped with GATK GenotypeGVCFs. WhatsHap was run on the combined VCFs with default settings. DNMs were then classified according to the same rules as for LRS (described above).

4.4 Results

In order to demonstrate the utility of LRS for *de novo* mutation detection, we sequenced eight parent-child trios using high coverage PacBio HiFi sequencing. We previously performed Illumina short-read whole genome sequencing (SRS) for all eight trios (van der Sanden et al., 2023), which allowed us to compare the performance of LRS and SRS on the detection of all types of variation in these trios, with a particular focus on *de novo* mutations (**Figure 1**).

4.4.1 Sequencing characteristics

For the PacBio HiFi LRS we obtained average read lengths of 17 kb. Over 99.0% of the 5.7 million reads per sample aligned to the reference genome with an average mapping quality of 46.5 (Supplementary Table 2A). The base error rate, computed as the edit distance over total number of mapped bases, was 1.4% per sample (ranging from 1.2 to 1.5%), which is in agreement with what has been published before (Hon et al., 2020). This resulted in an average coverage depth of 31x for all 24 genomes, which was as expected based on targeted coverage of 30x. 92.6% of the genome had at least 10x coverage depth. The average read mapping rate for the SRS was 99.6% with an error rate of 0.9% (ranging from 0.8% to 1.0%; Supplementary Table 2B). The average coverage depth was 73x (ranging from 51x to 103x), and across all samples 83.4% of all bases were covered ≥50x, while 92.0% were covered ≥10x.

4.4.2 Variants overview

Variant calling from LRS with DeepVariant and from SRS with GATK both yielded on average 4.1 million substitutions per sample (Supplementary Table 3A). On average 3.8 million substitutions per sample were shared between two platforms, which corresponds to 94.0% concordance for both the LRS and SRS call sets (Supplementary Table 3A). Of the substitutions that are unique to LRS, about half of all LRS-unique variants (average 110,000), was detected in regions for which SRS had no read coverage (Supplementary Table 3A). We found that LRS provides sequence coverage in about 240Mb of the genome where SRS does not. We found that in these regions the rate of Mendelian inheritance errors for LRS is only 2.1% suggesting that the majority of variant calls are real (Supplementary Table 3A).

For indels, the same callers yielded on average 1.0 million variants for LRS compared to an average of 0.9 million indels per sample with SRS (Supplementary Table 3B). The concordance was only 63.1% for SRS and

58.0% for the LRS indel call set (Supplementary Table 3B). For indels unique to LRS, around 25% were detected in regions that SRS had no read coverage (Supplementary Table 3B). The MIE ratio of LRS-unique indels (8.9%) was lower than that of SRS-unique indels (13.0%), indicating a slightly better ability of LRS for detecting indels (Supplementary Table 3B).

4.4.3 *De novo* small variant detection

Performing both LRS and SRS on the same samples allowed us to identify all variant types including substitutions. In this study we focused on assessing the accuracy of LRS HiFi for comprehensively calling small variants and SVs. A sensitive way of doing this is to detect and assess *de novo* mutations, since this type of variation has proven to be an important factor in the disease etiology of severe, early-onset, rare disease.

During LRS small DNM analysis, we identified clusters of LRS-unique variants that fall in the same gene with approximately the same coverage and variant allele frequency. These variants all appeared not *de novo* upon visual inspection and were removed from further analyses as described in more detail in the Methods. In total, 672 small DNMs were identified using strict filtering criteria, with on average 84 (range 73-92) small *de novo* mutations per child using PacBio HiFi LRS (Figure 1B; Supplementary Table 4A and 5), being in line with previously reported number of *de novo* substitutions per genome (Acuna-Hidalgo et al., 2016; Veltman & Brunner, 2012). On average, 75 of these 84 variants were single base substitutions, while there were 4 insertions and 5 deletions between 2-50 bp. Only two insertions >50bp were called using DeepVariant and these were also retained. Comparison of small DNMs called from LRS data versus substitutions called from SRS data showed 92.0% concordance (Figure 2; Supplementary Table 6A). Of the overlapping substitutions, 94.3% were called by both SRS DNM callers in the overlap set and the other 5.7% by only one of the SRS DNM callers in the LRS+ set (See Methods). Among all LRS DNM call sets, eleven were located in the coding regions of the genome and were all detected by both LRS and SRS (Supplementary Table 7). For SRS we found 859 small DNMs, with on average 107 (range 91 - 141) small DNMs per patient (Figure 1B; Supplementary Table 4B and 8), including 95 substitutions, 4 insertions and 8 deletions. The concordance for SRS small *de novo* mutations versus those called from LRS data, was 84.5% (Supplementary Table 6B). Of the overlapping small DNMs, 80.3% were called using the stringent LRS *de novo* filtering in the direct overlap set and the other 19.7% using the lenient LRS *de novo* filtering in the

SRS+ set (see Methods). The concordance for coding small DNMs was 100% (13/13 variants; Supplementary Table 7), albeit that 2 of the 13 coding variants were only identified in LRS after lenient filtering.

4.4.4 Small *de novo* mutation validation

In order to assess the sensitivity of LRS for the detection of small DNMs, we first attempted to validate all 54 LRS-unique small *de novo* mutation (**Supplementary Table 5 and 9A**; **Supplementary Figure 1**). For these 54, we aimed to design standard primer pairs suitable for Sanger sequencing. Due to the complex genomic regions of these variants, we only succeeded to design primers for 27 (50%) of the 54 variants. Of the 27 variants with successful primer design, 11 (40.7%) were confirmed as a true DNM, 11 were true variants but inherited from one of the parents and five were not confirmed in the child and therefore considered false positives (**Figure 2; Supplementary Table 5**). Small DNM quality scores were on average significantly higher for the confirmed DNM calls than for the inherited and false positive calls with quality scores of 55.5 and 54.4 for confirmed vs. 36.6 and 36.5 for inherited variants vs. 31.2 and 30.6 for false positive variants (P=2.8e-7, P=8.3e-6, P=8.3e-6 and P=3.0e-4; t-test) (**Supplementary Figure 2**). When looking at the specific locations of the 11 confirmed LRS-unique small DNMs in the SRS data, we found that all DNMs showed coverage at the specific genomic position and that the mutations were called. However, ten of these mutations were assessed by the SRS *de novo* mutation callers as being potentially inherited due to a small number of alternative base calls in one of the parents, and one was assessed as low quality DNM because of a small number of alternative base calls in one of the parents (**Supplementary Table 10**).

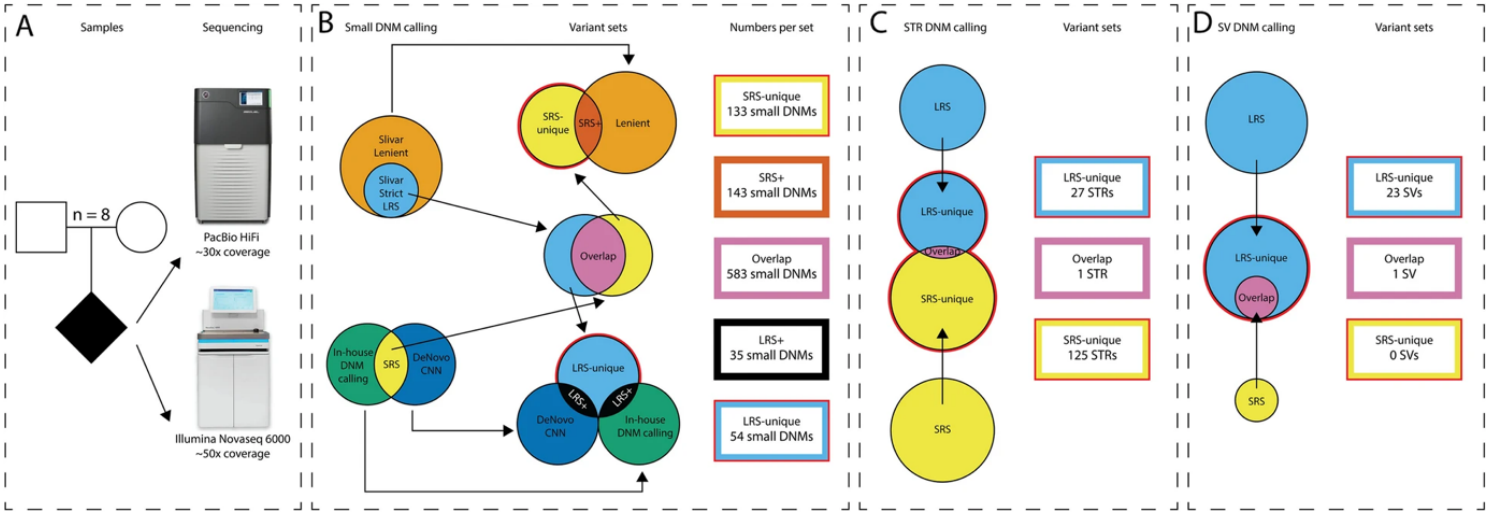

Figure 1. Study overview: We performed PacBio HiFi long-read sequencing and Illumina short-read sequencing for eight parent-child trios. (A). *De novo* substitutions, indels, STRs and SVs were called using dedicated bioinformatic tools. (B) For the small DNMs (substitutions and indels) the different call sets are depicted including the workflow of comparing the LRS call set to the SRS call set and the other way around. (C) The workflow for the comparison of STRs between LRS and SRS. (D) The workflow for the comparison of SVs between LRS and SRS. The circles in this figure are not drawn to scale.

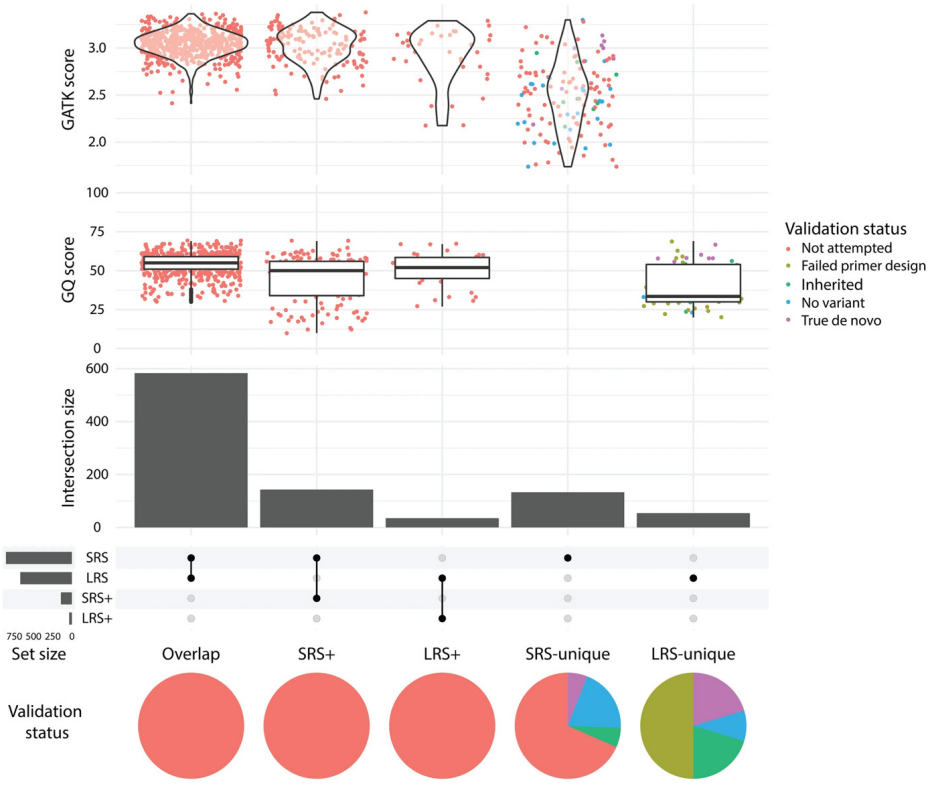

Figure 2. Detailed overview of all small DNMs: Upset plot of small DNMs detected by SRS and LRS.

The X-axis shows the concordance across different call sets. The Y-axis shows, from bottom to top, number of DNMs in each group, the DeepVariant Genotype Quality (GQ) scores of these small DNMs from the LRS data and the log-scaled GATK quality scores of these DNMs from the SRS data. Colors indicate the validation status and pie charts show the validation status of DNMs in each group.

For the 133 SRS-unique DNMs, visual inspection of the reads at the specific genomic position identified seven of them as high-confidence candidate small DNMs. Others were identified as low-confidence candidate small DNMs, either due to low read support or repetitive reference context. For validation, we selected a set of 42 SRS-unique small DNMs including all seven high-confidence candidate small DNMs as well as 35 randomly selected low-confidence candidate small DNMs (**Supplementary Table 9B**). For the total of 42 small DNMs we designed primers to determine whether these are true *de novo* calls. Of the 42 variants, eight (19%) were confirmed as true *de novo* events and eight other variants (19%) appeared to be inherited from one of the

parents (Supplementary Figure 3; Supplementary Table 8). For the remaining 26 variants (62%) Sanger sequencing failed to confirm the event called by Illumina SRS and these were therefore considered false positive calls. Five of the seven high-quality DNM candidates we initially selected were confirmed as true positive, while one had primer design sequencing failed and other one was false positive. Five of the eight SRS-unique true positive small DNMs also appeared *de novo* when inspecting the LRS alignment files (**Supplementary Table 11**). Of the three remaining SRS-unique small DNMs, two had low alternate allele depth and one had insufficient coverage for both alleles.

4.4.5 Differences between substitutions and indels

The predominant error mode in long-read sequencing is short insertions and deletions (Mantere et al., 2019). We therefore investigated whether there was a difference for the detection of substitutions and indels for both platforms. The 27 LRS-unique small DNMs for which we were able to perform validations consisted of 13 substitutions, 10 insertions and 4 deletions. None of the insertions and deletions were confirmed as true DNMs, while 11 substitutions were confirmed true *de novo*. For the insertions and deletions, 70% and 75% were inherited from one of the parents while 30% and 25% were false positive variant calls respectively. In addition, the average quality scores for the substitutions, insertions and deletions were divergent (53.5 and 52.5 vs. 32.9 and 32.5 vs. 36.3 and 36.5) (Supplementary Figure 1). For the SRS-unique variants, the 42 validated variants consisted of 34 substitutions, 2 insertions and 6 deletions. Only one deletion and seven substitutions were confirmed as true DNMs. Both insertions were false positive calls. In general, the SRS-unique variants were enriched for false positive calls, since 68% of the substitutions, 100% of the insertions and 17% of the deletions were false positive variant calls (Supplementary Figure 3). Furthermore, 9.5% of the substitutions were inherited while this was 0% for the insertions and 83% for the deletions.

4.4.6 De novo STRs

For STRs we used a tandem repeat catalog, containing 171,146 highly polymorphic repeat loci, as input for both TRGT and ExpansionHunter for LRS and SRS, respectively. On average we genotyped both alleles of all three family members for 171,038 (99.93%) loci for LRS and 171,113 (99.98%) for SRS (Supplementary Table 12). In total, we identified 28 (mean 4; range 1 – 6; Figure 1C and Supplementary Table 13) and 126 (mean 16; range 5 – 31; Figure 1C and Supplementary Table 14) repeat loci in LRS and SRS where one or both alleles in the child were ≥2 repeat units longer or shorter than the

number of repeat units in both parents and met our quality metrics (Methods). Therefore, these repeat calls were considered high quality candidate *de novo* STRs. Of these *de novo* repeats, only one call (3.6% for LRS and 0.8% for SRS) was concordant between the two platforms (Supplementary Table 15). We attempted to validate 18 LRS-unique and 18 SRS-unique high quality *de novo* STR calls. For the LRS-unique calls none were confirmed as true *de novo* repeat expansion. Of the 18 STR calls, 14 were false positive calls and four were true but not *de novo* because the repeat length was the same in one or both parents (Supplementary Table 16). For the SRS-unique STRs also none of the 18 high quality *de novo* STR calls were confirmed as true *de novo* as 13 calls were false positive and five were true but inherited from one or both of the parents (Supplementary Table 16).

4.4.7 *De novo* SVs

In addition to substitutions, indels and STRs, we also identified *de novo* structural variants for our patients using PBSV for LRS and Manta for SRS. In total, we identified 24 *de novo* candidate SVs with LRS and one *de novo* candidate SV with SRS (Figure 1D; Supplementary Table 17). The one SV in SRS that overlapped with LRS and was considered a true *de novo* event. The remaining 23 LRS-unique variants consisted of 13 insertions, 8 deletions, and 2 duplications (size range 21 – 991 bp). We aimed to systematically validate the *de novo* SVs using (long-range) PCR and subsequent targeted sequencing on a PacBio Sequel IIe system. For four of 23 variants, validation experiments repeatedly failed due to difficulties with designing suitable PCR primers. However, two out of the 23 variants were confirmed as genuine *de novo* SV events (Figure 3). In addition, eight SVs (five insertions and three deletions) in the size range of 21 to 991 bp were located within repeat regions and could be considered as *de novo* repeat expansions and contractions (Supplementary Figure 4). Therefore, in total 10 SVs were confirmed as a *de novo* event. Out of the 9 events that were not *de novo*, two SVs were inherited from one or both parents, while seven events were not detected at all (Supplementary Table 17). When analyzing the alignment files of both LRS and SRS at the genomic positions of all 24 SVs, it turned out that, besides the one overlapping SV, seven different LRS-unique SVs could be visually detected in hindsight (Supplementary Table 17; Supplementary Figure 5). Of these, five were validated as *de novo* event and one was inherited, while for one multiple validation attempts failed. For the remaining 16 SVs, we did not observe any patterns reminiscent of an SV in the SRS data (Supplementary Table 17; Supplementary Figure 5).

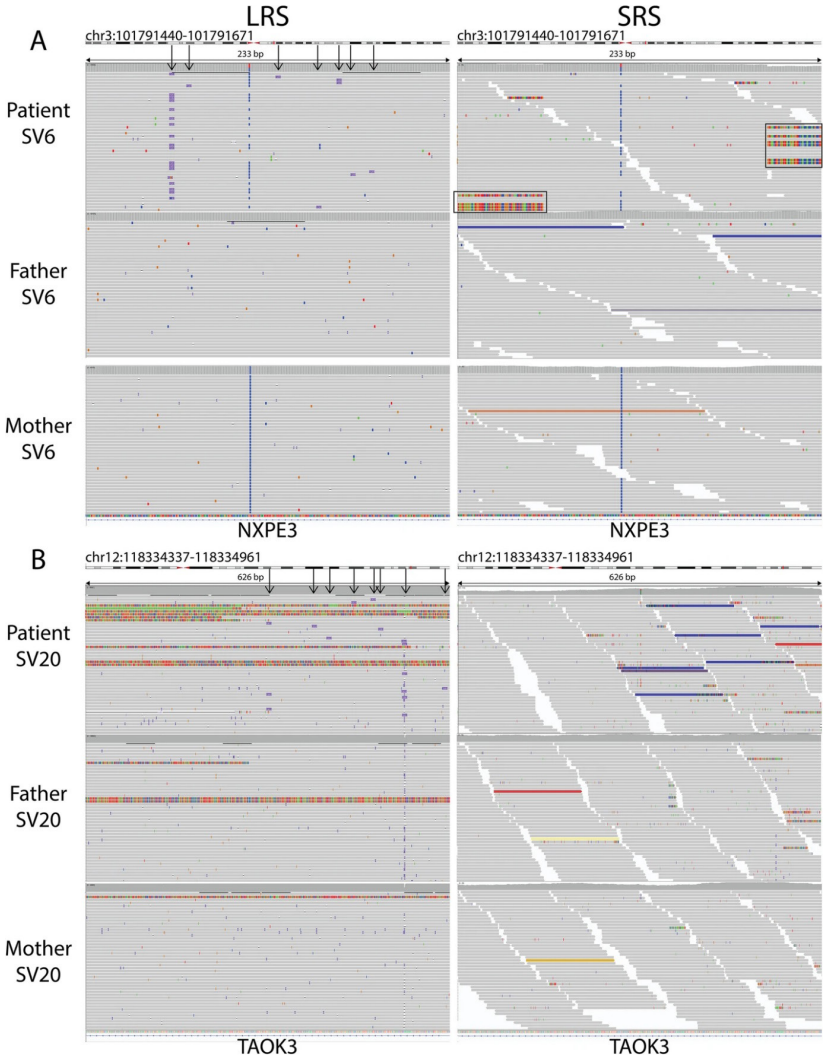

Figure 3. Two confirmed true *de novo* SVs only detect by LRS: For two variants the gel image and Integrative Genomics Viewer (IGV) screenshot for LRS and SRS is presented. (A) SV6 is a 123 bp duplication in an intron of the gene *NXPE3*. The position of the duplication call in the child's LRS reads are indicated with a black arrow. In the SRS data there are clipped reads, indicated with the two red boxes, hinting towards an SV event. However, the SV was not called in SRS. (B) SV17 is a 303 bp insertion in an intron of the gene *TAOK3*. The position of the insertion call in the child's LRS reads are indicated with a black arrow. In the SRS data some blue reads are visible, which represent reads with a smaller insert size than expected indicating a possible insertion. However, the SV was not called from the SRS data.

4.4.8 Titration

Our LRS samples were sequenced to relatively high coverage depth of 30x. When down-sampling to an average 20x and 10x coverage depth for the child and parents we observed on average 8 and 37 out of 75 validated DNMs could no longer be detected in the proband respectively (Supplementary Table 18; Supplementary Figure 6). In addition, the number of potential small DNMs increased considerably to on average 234 and 1,120 calls at 20x and 10x coverage depth respectively. This suggests that with the current LRS technology obtaining 30x average coverage depth is required for optimal detection of DNMs.

4.4.9 Phasing of small *de novo* mutations

One potential advantage for the detection of small DNMs using long-reads is the possibility to phase the DNMs and determine the parent-of-origin based on inherited variants (i.e., markers; Figure 4). Using the LRS reads for phasing resulted in haplotype blocks with a mean length of 570 kb, with an average of 800 small variants per block. When we used the SRS reads, mean length of haplotype blocks was 1.2 kB, with 14 small variants per block. With LRS we were able to assign 96% of DNMs to a haplotype block and subsequently all DNMs with a haplotype block were assigned to a parental allele (Supplementary Table 19A). For more than 80% of the phased DNMs from LRS there was >90% agreement between markers. With SRS we were able to assign 46% of the DNMs to a haplotype block. Because of the relatively small size of the phase blocks we could only assign 33% of the total DNMs to a parental allele (Supplementary Table 19B). Comparing the successfully phased DNMs we found >90% concordance between SRS and LRS (Supplementary Table 20). We found that all three discordant DNMs (100%) were found in repeated and low complexity regions of the genome. We found that 72.3% of the phased small LRS-detected DNMs and 78.4% of small SRS-detected DNMs were paternal, which was expected based on other studies of DNMs (Goldmann et al., 2016; Noyes et al., 2022; Peters et al., 2015) (Supplementary Table 19A).

One of the advantages of performing phasing is that we expect that true DNMs to be phased with high quality, while false positives caused by sequencing artifacts would not fit into a haplotype. Therefore, we checked the phasing of true positive and false positive DNMs from the LRS call set. All 11 validated LRS-unique small DNMs were phased. For the false positive and inherited DNM calls, 13 of 16 were not successfully assigned to a haplotype block by Whatshap (Supplementary Table 21). These results show that phasing can help to distinguish true positive from false positive DNMs (Fisher's exact test, P=3.39e-5).

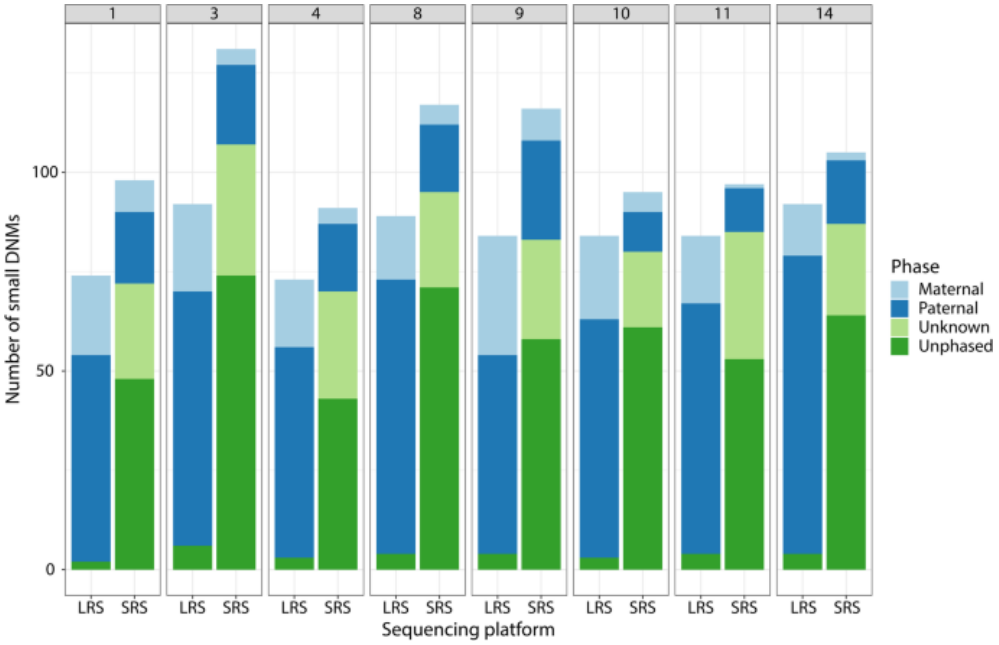

Figure 5. Phasing of DNMs: Number of DNMs per trio in a stacked bar graph, with colors for phasing results. With SRS and LRS next to each other grouped per trio (on average 89 phased small DNMs in LRS versus 21 phased small DNMs in SRS). Status of parentally phased DNMs for each trio. X-axis shows the sequencing platform, while Y-axis shows the number of DNMs. Colors indicate the assigned parental origin.

4.5 Discussion

Here we investigated whether HiFi PacBio sequencing offers sufficient base call quality to allow for comprehensive *de novo* mutation detection. This is important for several reasons: LRS is currently considered for samples that were unresolved by exome or genome sequencing with the intention to identify so far 'hidden', undetected, SVs. However, several studies have shown that short-read exomes or genomes may have missed genetic variants due to limitations of the technology or experimental design at that time (Mantere et al., 2019; Merker et al., 2018; Reiner et al., 2018; Seo et al., 2016). The potential ability to comprehensively detect all types of variation, now enabling the technically challenging *de novo* mutations, paves the way for LRS to replace SRS as the standard technology for genetic analyses as soon as costs become comparable. This possibly enables testing of 'all' rare disease patients with a suspected genetic cause with a single comprehensive test. However, depending on platform and local infrastructure, LRS is currently 3-6 fold more expensive than SRS.

When comparing all substitutions between the platforms we find that there is a 94% concordance and that the MIE rate for variants unique to LRS is very low at 2.1%. If we assume that all substitutions with correct Mendelian inheritance are real, the average sensitivity of HiFi LRS is 99.84% compared to 99.96% for SRS, and for indels 95.03% and 93.44% respectively. When comparing DNMs from LRS with SRS, we find that the overlap for LRS was 92.0% and for SRS 84.5%. LRS detected 54 unique variants while SRS detected 133 unique variants. With SRS, we found that these specific variants are mostly false positive calls in the proband (i.e., sequencing and mapping artifacts), whereas with LRS most (68.8%) were inherited. The same observation was made by Noyes et al. in their study that used multiple long and short-read technologies to establish a most comprehensive set of DNMs in a parent-child quad (Noyes et al., 2022). They observed that 71% of false DNM calls were due to a missed call in one of the parents. This is promising for LRS technology since improved coverage, or improved evenness of coverage, will likely reduce this type of false positives. If we would disregard this type of false positives from our study, LRS performs very favorably to SRS. In this case our *de novo* validation rate for LRS, considering only successful PCRs, would increase from 40.7% (11/27) to 68.8% (11/16), while the validation rate for SRS DNMs would only increase from 19.0% (8/42) to 23.5% (8/34).

We experimentally validated 11 LRS-unique and 8 SRS-unique DNMs. For the LRS-unique variants, we found that all 11 were called by SRS as well. One DNM was assessed as low-quality DNM in SRS due to mapping quality issues at the position of this event. The other ten were considered as potentially inherited mutations since they had fractional support (3-6 reads with alternative allele) in the parents. This might be due to sequencing artifacts or might be due to parental mosaicism. For two of these ten variants, Sanger sequencing traces only showed a small variant peak in one of the parents. Therefore, we still considered these as true DNMs, but the deeper sequence coverage of SRS had a slightly higher chance to identify low parental mosaicism in these two. For the SRS-unique small DNMs, four out of the eight were not called *de novo* in LRS due to poor genotyping in one of the parents. Two small DNMs were not called *de novo* in LRS due to the presence of alternative allele reads in one of the parents. The remaining two small DNMs were not called in LRS due to insufficient coverage and a low variant allele frequency at the position of the event in the proband.

Compared to the study by Noyes et al., our per sample *de novo* mutation numbers and concordance ratio (between SRS and LRS) are very similar (Noyes

et al., 2022). Noyes et al. called an average of 81 *de novo* substitutions and 6 indels in their probands. Our per sample averages for these types of variants were 75 and 9 respectively. On the other hand, their SRS call set consists of on average 82 *de novo* mutations in the probands, which is somewhat smaller than our per-sample average of 107. This is mostly because *de novo* mutations in repetitive regions were removed from their call set. Since their SRS call set is more restrictive, the concordance of their LRS call set compared to their SRS call set drops under 80%, compared to our finding of 92.0%.

In a previous study, we applied LRS to 5 trios, using a PacBio Sequel system and without the use of circular consensus calling to improve base pair accuracy (Pauper et al., 2021). When considering the single trio from this study that was sequenced at similar coverage (30x) we previously identified 655 small DNM candidates compared to 84 small DNMs per trio in our current study. Even though, due to the higher error rate in the sequencing data of our previous study (Pauper et al., 2021), the selection criteria for small DNMs were much more stringent, the number of small DNM candidates was still considerably higher. In our comparison with SRS we identified an overlap of only 58.3% whereas in our current study this is 92.0%. Fifty percent of false positive small DNMs in our previous study were due to false positive insertion calls in the proband, whereas this is a substantially smaller proportion (30%) in the current study. The differences between these studies illustrate the improvements that have been made with the introduction of HiFi sequencing.

Besides showing that small DNM calling using HiFi LRS is on par with SRS, we have also analyzed four traditional benefits of LRS over SRS. First, we show the more accurate detection of *de novo* SVs. We confirmed ten out of 23 *de novo* SVs in total, of which eight in Repbase (Bao et al., 2015) annotated repetitive DNA elements. The validation rate of only 43.5% may be explained in part due to the proximity of all 23 *de novo* SVs to repetitive reference contexts which made these challenging to confirm. Out of the 13 SVs that were not confirmed, for four the validation did not refute the *de novo* event itself but was mostly inconclusive. Combining the ten confirmed *de novo* SVs with the one SV detected by both sequencing platforms this comes down to an average of 1.375 *de novo* SV per genome. This number is markedly higher than current estimates based on short-read WGS data of 0.02 to 0.286 *de novo* CNVs and SVs (>50bp) per genome (Belyeu et al., 2021; Collins et al., 2020; Veltman & Brunner, 2012). This is likely due to detection of the eight SVs that could be considered repeat expansions/contractions. When only considering the other

three SVs, our study also suggests that *de novo* SVs are rare with a *de novo* SV rate of 0.375. The rare nature and the possibility to identify those reliably with LRS, without enriching for false positive SV calls, confirms the accuracy of LRS for this variant class. It also confirms that a *de novo* SV hypothesis for rare and severe disease works and strongly reduces the number of candidate variants per offspring.

Secondly, we performed STR detection from both LRS and SRS. In total we detected 28 high quality *de novo* STR expansions and contractions for LRS and 126 for SRS. Of these only one overlapped between the two technologies. Moreover, none of the LRS-unique and SRS-unique repeat expansion, for which we attempted validations, was confirmed to be true *de novo*. This shows that although LRS seems to outperform SRS in this area with a slightly higher specificity, there is still a lot of room for improvement in the detection of STRs. For a fair comparison we restricted our analysis 171,146 known highly polymorphic repeat regions. However, to show the full potential of the detection of STRs with LRS, a genome-wide analysis would be more appropriate. This is illustrated by the fact that the SV calling on LRS data identified another eight *de novo* structural variants that upon closer examination turned out to be STR expansions/contractions.

Thirdly, LRS allows improved phasing of the DNMs and determine the parent-of-origin based on surrounding inherited variants. Comparing only phased small DNMs we find a more than 90% concordance of the assigned parental allele between LRS and SRS. However, almost all DNMs (>96%) could be phased with LRS compared to only 33% with SRS. Phasing also supports the quality of our DNM results in LRS. The fact that the DNMs are not artifacts is supported by the consistency of the phasing by multiple single nucleotide polymorphisms, all supporting the same parental allele. In LRS, additional validation of a DNM with an orthogonal technology could be omitted when additional support from phasing results based on a reasonable number of SNPs is available. Benefits of phasing in future studies not only entail this increase in DNM specificity, but could also increase the specificity for post-zygotic and somatic DNMs (Acuna-Hidalgo et al., 2015; King et al., 2017) and allow better studies of DNM biology (Goldmann et al., 2016; A. Kong et al., 2012).

Finally, with LRS more of the human genome is accessible and, for the first time, variants can be called in these regions that remained inaccessible with other sequencing technologies. With LRS we found on average 240Mb of

uniquely covered regions per sample, compared to 133Mb per sample for SRS. This is also in agreement with previous literature about the dark regions of the genome (Ebbert et al., 2019).

Despite these advantages of LRS over SRS, the cost of sequencing is an important disadvantage of HiFi LRS. The current price of a HiFi genome at 30-fold coverage is 3-6 times higher than a genome achievable with SRS at 30-fold coverage. With future iterations of the HiFi LRS platform the costs for a 30-fold coverage genome will drop up to 3-fold, but also SRS will be available at half its current price. To address whether the benefits of LRS are worth the additional costs, more extensive clinical utility studies are required, which is beyond the scope of this current study.

4.6 Conclusions

HiFi LRS can now produce a very comprehensive WGS datasets obtainable by a single technology in a single laboratory, allowing accurate calling of substitutions, indels, STRs and SVs. This possibly enables for truly generic testing of 'all' rare disease patients with a suspected genetic cause with a single comprehensive test. The accuracy of HiFi LRS even allows sensitive calling and phasing of DNMs, which are a major cause of severe early-onset disease, on all variant levels.

Declarations

Ethics approval and consent to participate

All participants or their legal representatives gave written informed consent. This study was approved by the Medical Review Ethics Committee Oost-Nederland and Radboudumc Institutional Review Board under 2020-6853, as part of 2018-4985 and 2014-1254.

Consent for publication

All participants or their legal representatives gave written informed consent.

Availability of data and materialsThe datasets supporting the conclusions of this article are available in the EGA repository under accession number EGAS00001006479.

Competing interests

AMW, CL, SC, PB, WJR and ZK are employees and shareholders of Pacific Biosciences, a company commercializing DNA sequencing technologies.

Funding

Financial support was obtained from grants from the Netherlands Organization for Health Research and Development (ZonMw; 843002608, 846002003 and 015014066 to LELMV) and from the Netherlands Organization for Scientific Research (917-17-353 to CG). Drs Gilissen, Vissers, Brunner, and Hoischen are supported by the Solve-RD project. The Solve-RD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 779257. Drs Brunner and Vissers are part of ERN-ITHACA and Dr Hoischen is part of ERN-RITA.

Authors' contributions

Conceptualization: HGB, LELMV, AH, CG; Data curation: EK, BPGHS, LO, AMW, WJR, ZK; Formal analysis: EK, BPGHS, RD, AMW, CL, SC, PB, WJR, ZK; Funding acquisition: HGB, LELMV, AH, CG; Investigation: EK, BPGHS; Methodology: EK, BPGHS, LO, AMW, WJR, ZK; Project administration: AH, CG; Resources: HGB, LELMV, AH, CG; Software: EK, LO, AMW, WJR, ZK; Supervision: AH, CG; Validation: BPGHS, MK, RD; Visualization: EK, BPGHS; Writing-original draft: EK, BPGHS, AH, CG; Writing-review & editing: EK, BPGHS, AH, CG. All authors have contributed to the manuscript and have approved the final version.

Acknowledgements

We thank Radboudumc Genome Technology Center for sequencing and performing the SV validations.

List of abbreviations

AQ	Allele Quality
BAM	Binary Alignment Map
CCS	Circular Consensus Sequencing
CNV	Copy Number Variant
DNM	*De Novo* Mutation
GATK	Genome Analysis Toolkit
GQ	Genotype Quality
HiFi	High Fidelity
IGV	Integrative Genome Viewer
Indel	Insertion or Deletion
LRS	Long-Read Sequencing
PBSV	Pacific Biosciences Structural Variant
PCR	Polymerase Chain Reaction
SMRT	Single-Molecule Real-Time
SRS	Short-Read Sequencing
SV	Structural Variant
TRGT	Tandem Repeat GenoTyper
WGS	Whole Genome Sequencing

Supplementary Information

Supplementary figures and tables can be accessed at following DOI:
https://doi.org/10.1186/s13073-023-01183-6

5. Discussion

In this thesis I showed how advancements in technology enable a more thorough examination of the human genome, which has the potential to enhance both patient diagnoses and medical research. The main focus was on two technological advancements: improved exome sequencing with the Twist exome capture assay (Chapter 2), and improved variant detection with Pacific Biosciences long-read sequencing technology (Chapters 3 and 4).

5.1 Summary and Implications of Key Results

5.1.1 Twist exome capture kit provides more uniform coverage and consistent variant calling in the coding regions over the state-of-the-art

Whole exome sequencing (WES) has become a standard tool in clinical diagnostics, thanks to the advancements in targeted sequencing approaches. Various exome capture kits employ different target enrichment strategies, leading to differences in coverage uniformity and capture efficiency. In Chapter 2, I compared the coding sequence coverage and SNV detection sensitivity of the novel Twist exome capture kit to widely used Agilent V5 and V7 exome kits, as well as to SR- and LR-WGS. The Twist enrichment kit not only had a higher coverage ratio of coding regions (99.4% vs 96.7%), but also showed it had a more even coverage profile compared to Agilent kits. Whereas a2.7% improvement may appear relatively modest, this encompasses about 1 Mb of coding sequence. Crucially, I demonstrated that the sensitivity of variant detection using TWIST exome capture kit at 70x coverage was comparable to other kits at 150x, with only 0.5% of variants missed. I estimate that up to 40% decrease in sequencing costs could be possible with the reduced coverage. This means that even though the cost of SR-WGS is decreasing rapidly, exome sequencing can still be substantially more cost-effective for routine diagnostics and large-scale research projects.

5.1.2 LRS revealed the true scope of previously undetected structural variation

Structural variants are difficult to detect using traditional short-read sequencing technologies, due to their size and sequence complexity (M. J. P. Chaisson et al., 2019; Huddleston et al., 2017). In Chapter 3, I showed that when compared to whole genome SRS (at 27x-33x), LRS found 5-fold more structural variants, with 28,292 SVs per sample. The detected SVs affected around 12 Mb of sequence. We confirmed the reliability of detected SVs by

showing a low number of Mendelian inheritance errors and significant overlap with published datasets. The increased sensitivity for SV detection is partially due to the fact that long reads are able to interrogate complex regions of the genome, where short-read techniques typically fall short. We found around 35 Mb of the human reference genome (hg38) exclusively covered by LRS, which contained on average 3,874 SVs and 32,440 SNVs per sample. We also found that genes located in those so-called "dark" regions had a higher GC-content than the genome average. Contrary to SRS, LRS is not prone to GC-bias, and therefore was able to uncover variation in those regions.

The relatively high sequencing error rate of LRS (8-15%) compared to SRS (~1%) is a limitation to the identification of single nucleotide variants. We show that we were able to remedy this problem by increasing the average sequencing coverage of an LRS sample to 30x, thereby obtaining a substantially concordant (85%) SNV call set compared to SRS. This confirms that LRS error rates are mostly random and that, at higher sequence coverage, LRS technology may be able to replace SRS in many applications.

5.1.3 HiFi reads improves the accuracy for detection of point mutations and small indels, producing a comprehensive and accurate variant call set

One of the drawbacks of the LRS technologies that was apparent from the previous study is their high base calling error rate, which especially affects the specificity of SNV detection. In Chapter 4, I demonstrated that an improved version of PacBio LRS, so-called HiFi long-reads, improve the accuracy of the detection of point mutations and small indels in the genome to a level comparable to whole genome SRS. We compared LRS to SRS and found that there is a high concordance for inherited variants, at around 94%. Considering that around 2% of the LRS-detected regions were not covered with SRS and therefore had no chance of overlap, this is a remarkably high level of concordance.

The high accuracy of HiFi reads means that they can be utilized in technically challenging applications. One such application is the detection of *de novo* mutations, which are important drivers of developmental disorders. The concordance between LRS and SRS for *de novo* variation was considerable, at around 85%. Interestingly, while false positives from the SRS call set were mostly due to sequencing errors and artifacts, false positives from the LRS call set came predominantly from inherited variants that were incorrectly

genotyped at positions of low-coverage in parental genomes. This indicates that simply increasing coverage for LRS samples will likely further reduce false positive calls.

Another potential advantage of LRS is the ability to phase variants, i.e. derive whether variants occur on the same allele. LRS was able to parentally phase 98% of the DNMs, while for SRS this was around 20%. We also found that successful phasing is an indication of quality of the DNM call. 13 of 16 of our false positive DNM calls could not be phased, whereas none of the 11 true DNMs could not be phased. Therefore, we believe phasing can be a useful quality metric for variant calls in LRS data.

A recent long-read trio study by Noyes et al. (2025) in 73 children from 42 simplex autism families (157 individuals) combined PacBio HiFi for discovery with ONT and Illumina validation, assaying ~2.77 Gb of callable genome and yielding on average 95 de novo mutations (DNMs) per child (≈87.5 SNVs and 7.8 indels). Long reads increased DNM discovery by 20–40% over prior Illumina analyses of the same families and more than doubled detectable postzygotic mutations (PZMs). In total, 6,030 DNMs (SNVs) and 533 indel DNMs were validated; 15.1% of SNVs were PZMs. Parent-of-origin could be assigned for 98.0% (germline) and 96.1% (PZMs), revealing a strong paternal bias for germline DNMs (3.98:1) and a modest paternal skew for PZMs (1.15:1), with a paternal age effect of +1.32 SNVs/year. These results are widely in line with my findings in Chapter 4, confirming its results on number of DNMs per proband, increased yield with LRS and paternal bias.

Overall, these results showed that LRS is now capable of producing an extensive and accurate variant call set that includes substitutions, insertions and deletions, short tandem repeats, and structural variants. For instance, in a recent study researchers looked at 100 samples with 145 clinically relevant, hard-to-detect known mutations. They were able detect 80% of these variants with 30x LRS, including 90% of the SVs (Höps et al., 2025). The ability to detect different kinds of variation with a single technology potentially opens the door to the use of LRS as a comprehensive first test for all suspected rare disease patients. However, the routine application of LRS still requires overcoming significant technical challenges and limitations discussed below, namely the cost of larger cohort sizes, availability of software and best practices for the analysis and the difficulty of interpreting the LRS detected variants.

5.2 Limitations of Our Approach

In recent years, several studies have demonstrated the potential of LRS to identify the genetic cause in unresolved patients (Conlin et al., 2022). For instance, using targeted LRS, a recent study (Miller et al., 2021) identified missing disease-causing genetic variants in 6 of 10 undiagnosed patients with suspected Mendelian disorders. Notably, LRS was able to resolve complex structural rearrangements in 8 cases in that study, leading to novel diagnoses in 6 cases. Similarly, in a study focusing on Werner syndrome, an adult-onset recessive progeroid disorder, applying targeted Oxford Nanopore sequencing detected the "missing" second pathogenic variant in 8 of 9 patients who previously had only one mutant allele identified (Miller et al., 2022). These included four cryptic intronic splice-site mutations that had eluded short-read tests. Such results highlight how LRS can fill critical gaps in molecular diagnoses by pinpointing variants that were not accessible before. For example, variants in so-called "NGS dead zones": homopolymers, genes with high GC-bias, or those that have pseudogene counterparts, but also repeat expansions and complex structural rearrangements (Mandelker et al., 2016).

In my studies of long-read sequencing, I have mainly focused on the technical evaluation of technological improvements, but not that much on the effect of these improvements for medical research and diagnostics. For our studies, we selected patients with unresolved intellectual disability, for which short-read sequencing had not identified a genetic cause. We were expecting to identify new genetic causes of disease, specifically undetected structural variants. Although, for these patients, we generated a more comprehensive catalog of genomic variants compared to the standard, disappointingly, in our studies we did not identify any genuine candidate variants in these unsolved patients. Here I briefly examine several reasons why I believe our studies were partially unsuccessful.

5.2.1 Larger cohort studies may be required to establish true clinical utility of LRS

At the start of my studies, the cost of long-read sequencing was still prohibitive for including large numbers of samples. Because we wanted to focus on *de novo* mutations, the costs for sequencing tripled per patient. Therefore, in Chapter 3 we were only able to examine 5 patients, whereas in Chapter 4 we examined a mere 8 patients. From microarray studies we know that in an unselected cohort of ID patients, about 1 in 6 patients carries a pathogenic *de*

novo CNV (Vinas-Jornet, 2018). If we ignore the fact that our samples were already pre-screened using WES and WGS, assuming a similar rate of LRS SVs as for microarray CNVs, this means that in our 13 samples we had a 92% probability of not finding any pathogenic *de novo* SVs. This notion is somewhat supported by a study of Kobayashi *et al*. (Sanford Kobayashi et al., 2022) who used LR-WGS on 26 patients with undiagnosed, severe developmental disorders. They were not able to solve any of their cases despite extensive evaluation. Their approach was very similar to Chapters 3 and 4 of this work, which had 5 and 8 patients respectively. The failure of Kobayashi *et al.* to find new diagnoses in their larger pool of patients suggests that the cohort sizes in Chapters 2 and 3 were not sufficient to evaluate the true clinical utility of LRS. It might be argued that this shows that the immediate utility of LRS is minimal.

However, recent studies with increased sample sizes do show an improvement in diagnostic yield. Redfield et al. evaluated PacBio HiFi long-read genome sequencing in 19 pediatric SNHL probands who were nondiagnostic after prior testing (100% exome sequencing; 94.7% short-read genome sequencing), achieving a 21.1% diagnostic yield (4/19) (Redfield et al., 2024). At an even larger scale, the Solve-RD study re-analyzed a cohort of 114 undiagnosed rare disease families using LRS. As a result, it identified new diagnoses for 12 families and identified potential pathogenic candidate variants for another 5 patients (Steyaert et al., 2025). In a similar vein, a recent study of 96 trios with congenital adrenal hyperplasia (CAH) – a condition notoriously challenging for genetic diagnosis due to a highly homologous pseudogene interfering with short-read mapping – showed that long-read sequencing provided more precise molecular diagnoses for 15% of patients across this large cohort (Wang et al., 2025). By fully resolving the complex CYP21A2 gene and its pseudogene, LRS achieved a level of accuracy and completeness in variant detection that was previously unattainable with short reads.

Another consideration is how samples were selected. For our study we selected cases that were negative by exome and short-read genome sequencing. These cases were however not specifically selected based on phenotypes but all presented with general ID. Samples from the study of Steyaert et al., where however only partially screened by exome and or genome and typically with older generations of these technologies. Similarly, Wang et al. examined a disease-specific cohort of congenital adrenal hyperplasia, where the causal gene CYP21A2 resides in a highly homologous locus with the CYP21A1P pseudogene; such architecture favors LRS over short reads for accurate variant detection. These

differences in pre-test probability and extent of prior testing can partially explain why our study did not identify the genetic cause in any of the tested samples.

In general, in larger cohorts, long-read whole-genome sequencing has demonstrated a consistent added diagnostic yield in rare disease cohorts that were negative by short-read methods. Across recent LRS studies of unresolved patients, where samples varied in how they were prescreened and selected, long-read WGS provided an additional ~7–17% of diagnoses (**Table 1**).

Table 1. Diagnostic yield of long-read sequencing studies in rare disease.

Study	Year	Technology	Samples	Phenotype	Sample Selection	Additional Yield (%)
Pauper et al.	2021	PacBio CLR	5 trios	Intellectual disability (ID)	WES/WGS-negative	0%
Kobayashi et al.	2022	PacBio HiFi (30×)	30 single patients	Severe developmental disorders	WGS-negative	~3.3%
Miller et al.	2022	ONT (targeted)	9 single patients	Werner syndrome (adult progeroid)	Cases with one known allele (WES)	~89%
AlAbdi et al.	2023	PacBio HiFi (10×)	34 single patients	Various autosomal recessive diseases	WES-negative	38%
Kucuk et al.	2023	PacBio HiFi (30×)	8 trios	Intellectual disability (ID)	WES/WGS-negative	0%
Hiatt et al.	2024	PacBio HiFi (~30×)	96 single patients	NDD/MCA	WGS-negative	7.3%
Redfield et al.	2024	PacBio HiFi (24–32×)	19 single patients	Sensorineural hearing loss	WES/WGS-negative	21.1%
Fabián-Morales et al.	2024	PacBio HiFi (30×)	3 single patients	Inherited retinal dystrophies (IRD)	WES-negative	100%
Steyaert et al.	2025	PacBio HiFi (~10×)	~100 trios	Mixed neuromuscular/ epilepsy phenotypes	WES/WGS negative	~13%
Wang et al.	2025	PacBio HiFi (targeted)	96 trios	Congenital adrenal hyperplasia (CAH)	Targeted SRS and PCR	~15%

Studies are grouped by platform and reported as additional yield–i.e., the proportion of previously unsolved cases (typically WES/WGS-negative) that received a diagnosis specifically due to long-read sequencing.

Applying LRS to larger cohorts is becoming more feasible due to recent advances in throughput and decreasing costs. For instance, the newly released Pacific Biosciences Revio system has brought down the cost of sequencing a whole human genome with 30x coverage (the standard in many clinical applications) to around $1000 (Pacific Biosciences, 2023) in consumables. Even though this is still three times the cost of a comparable SRS human genome (around $500), the price point is affordable for major centers with high throughput projects. Therefore, long read sequencing is being adopted for population-scale biomedical research, such as NIH's All of Us and Center for Alzheimer's and Related Dementias (CARD) programs (De Coster et al., 2021). The first stage of the All of Us program involves the sequencing 6000 whole genomes with LRS. Similarly large studies with clinical genomes will be needed to determine the additional diagnostic yield of LRS compared to SRS more accurately. In addition, this scale will help in ascertaining patients with phenotypes that are most likely to benefit from LRS in the future.

As population-scale LRS becomes the norm, estimating the true burden of pathogenic SVs will be possible. Based on the past whole genome SRS studies, it is estimated that between 4% to 12% of high-impact coding alleles are SVs (Lappalainen et al., 2019) with between 0.1% to 0.3% of individuals in the general population carrying a clinically relevant SV (Collins et al. 2020). These estimates based on the coding regions are unlikely to change dramatically with LRS. Even though LRS studies typically find 3 to 5 times more SVs than SRS, 75 to 80% of these novel SVs are in non-coding repetitive regions. In a recent study of samples from 1000 Genomes project, Nanopore sequencing yielded 167,291 SVs, considerably advancing the state of the art compared to the SRS studies (Schloissnig et al., 2024). LRS not only yields more SVs, it also allows more precise detection of SV breakpoints. A national pilot from Genomic Medicine Sweden evaluated PacBio Revio long-read genome sequencing for "digital karyotyping" of clinically indicated chromosomal rearrangements across 16 samples from 13 families collected nationwide, each previously known to harbor an SV (Eisfeldt et al., 2024). LRS detected 14/16 rearrangements, resolved 13 to nucleotide-level breakpoints, and identified one additional complex event by read-depth; the two undetected events involved Chromosome 21 (one mosaic), consistent with acrocentric/low-mappability challenges. As a result, authors of that study now propose a 5-year implementation plan to scale LRS in Swedish rare disease diagnostics.

For clinical relevance, it is also important to note that Beyter et al. (2021), using LRS data from 3,622 Icelanders, found 5,238 SVs that are in strong LD with

variants that are associated with a disease or a phenotype in the GWAS catalog (Beyter et al., 2021). According to these results the number of structural variants (SVs) linked to disease-associated variants in the GWAS catalog more than doubled when using long-read sequencing (LRS) compared to short-read sequencing (SRS), underscoring the added resolution LRS provides in detecting clinically relevant SVs. Even though they might be in non-coding regions, LRS detected SVs can contain a significant amount of information related to physical traits and disease, which might be missed by SRS. As I will discuss below, as our ability to interpret non-coding variation improves, the clinical utility of LRS detected SVs will increase.

5.2.2 Streamlined and scalable analysis of LRS data is constrained by lack of standard software and best practices

In genomics research, the significance of software cannot be underestimated, as it plays a crucial role in analyzing the vast amount of data generated from sequencing technologies. The analysis of sequencing data involves separate tools for preprocessing and alignment of the reads to the reference genome, as well as calling and filtering different types of genomic variants. However, the development of software often lags behind the emergence of new sequencing platforms, which necessitates the creation of new tools to effectively handle and interpret the novel types of data. For instance, even though PacBio LRS technology has been accessible since early 2011, it wasn't until 2022 that the company launched an official software tool designed for extracting methylation signals from the data. This significant delay in software availability is the primary reason I did not explore methylation patterns in my samples. The availability of new tools is also hampered by the fact that developing and maintaining high quality scientific software requires significant amount of time, computational expertise and resources that are often limited in academic centers. Although there is a trend toward the commercial development of bioinformatics software and services, the market remains very small and specialized (Gullapalli, 2020).

These limitations are particularly evident in the case of LRS, especially HiFi reads, where specialized software is still under heavy development. An overview of 170 currently available bioinformatic tools developed for LRS shows that 30% of them focused on error correction and polishing of the reads, and a further 30% and 20% were dedicated to *de novo* assembly and alignment, respectively (Amarasinghe et al., 2020). This reveals the lack of specialized software dedicated to more downstream analysis such as variant calling.

Furthermore, due to its comparative advantage, most variant calling tools developed for LRS are focused on SV detection (Ho et al., 2020.; Mahmoud et al., 2019). For the purposes of this study, only two tools were available to use for SNV and small indel calling from HiFi reads, namely DeepVariant and Longshot (Edge & Bansal, 2019; Yun et al., 2021). In contrast, a recent meta-analysis of 68 benchmarking studies revealed 498 published tools available for SRS data (Gardner et al., 2022). Furthermore, most reviews of SRS variant calling can list up to a dozen tools each for small variants, SVs and somatic calling (Barbitoff et al., 2022; Koboldt, 2020). By employing different algorithms and approaches, these tools increase the utility of SRS for different use cases.

Conversely, the lack of mature computational methods for LRS means its full potential is often not realized. For example, LRS can be a quite powerful tool to detect complex genomic inversions. However, the approach in Chapter 3 yielded on average 10 to 20 inversions per sample, while recent studies revealed that there can be 50 to 150 inversions in an individual human genome (Porubsky et al., 2022). Since there are no dedicated inversion callers for LRS data as of writing, an important part of human genetic variation remains underexplored. Another obvious direction for new methods is to make use of the ability to perform haplotype phasing. In Chapter 4, I have combined published tools and custom scripts to demonstrate the substantial advantage of LRS for this task. In the future, a routine phasing step can be integrated into variant callers for LRS data. Furthermore, even though I was able to achieve phasing ranges at an average of 800 Kbp, a haplotype-resolved *de novo* assembly of the data resulted in phased blocks of 80 Mbp in length on average (Cheng et al., 2021). This means similar phasing ranges can be obtained if this approach can be integrated into routine variant calling for LRS reads. For SRS, this concept is already implemented in GATK Haplotypecaller, which uses local *de novo* assembly of haplotypes to call variants.

However, the mere availability of the scientific software is not enough for optimal performance. Software developers and the users have to work together to determine and disseminate the best practices. In the absence of mature software, many centers create custom pipelines for analyzing data, with varying results and quality that makes cross-studies comparison difficult. An example would be Chapter 3 of my thesis, where we used two different sets of filtering criteria when dealing with *de novo* variation, to make sure that we were not arbitrarily disregarding any true positive results. Furthermore, when we compared our results to Noyes et al., we had to account for their different

filtering criteria. Standardization of pipelines through comprehensive documentation, validation of results, and automation of as many steps as possible (including quality control) allows researchers to avoid such hurdles.

It is therefore not unlikely that future re-analysis with new or improved methods for variant detection, applied to the LRS data from Chapters 2 and 3 will still identify plausible candidate variants. This is especially true for certain types of SVs, such as repeat expansions and small indels (<50 bp), due to the dearth of mature tools to detect these variants from HiFi data.

5.2.3 LRS results are not easily translated to clinical outcomes due to lack of annotations and databases

For very novel technologies it is not only the development of software that may lag behind but also genome annotations that allow for the interpretation of identified variants. For coding single nucleotide variation, there are many resources available that are necessary to interpret the clinical relevance of variants. For example, the gnomAD database contains data from 76,156 individual genomes and is used to quickly filter-out variants that are too common in the population to be disease-causing (Karczewski et al., 2020). Such resources are (at the time of writing) not yet available for LRS detected structural and non-coding variants, even though gnomAD recently incorporated population frequencies for SRS detected SVs. In Chapters 2 and 3 we filtered SVs based on their frequency within our own cohort in order to include common SVs. However, this reduced the number of SVs only by 15% to about 25 thousand per case. Further determining pathogenicity of SVs then remains a manual task that is done on a case-by-case basis. For instance, recently released guidelines for CNV reporting and interpretation take a rule-based approach that relies heavily on the specific opinions of domain experts (Amarasinghe et al., 2020). This approach is not easily scalable to LRS, which can detect on average 29 thousand high-quality SVs per sample, which can be filtered down to 25 thousand based on cohort frequencies, a set that includes on average 1000 SVs overlapping with genic regions. It is important to note that more than 70% of our SVs did not substantially overlap (>50%) with any variant from the GnomAD-SV catalog. This is expected as GnomAD-SV is based on SRS and therefore lacking in insertions and repetitive SVs that constitute the majority of our LRS call set.

Aforementioned difficulties forced this study to mostly focus on *de novo* SVs, rather than examination of inherited SVs. The future availability of better

population databases with SVs from LRS data, which will take advantage of the more comprehensive and accurate variant detection, may allow us to re-examine inherited SVs identified in our patients possibly resulting in new candidate variants. The recent release of the human pangenome, which is based on highly accurate (99%) and highly comprehensive (99% of expected bases) genome assemblies of 47 individuals with diverse ancestries (Liao et al., 2023), besides providing a global representation of genomic variants, will also facilitate the discovery of complex SVs that cannot be easily mapped to the current human reference genome (T. Wang et al., 2022).

Furthermore, most of the current guidelines and tools focus on protein-coding regions, while this work and many others have shown that true advantage of LRS lies in uncovering variants in more complex regions of the genome, which are typically non-coding regions (M. J. Chaisson et al., 2015; Huddleston et al., 2017). These regions often remain poorly annotated in terms of the regulatory features, sequence constraint and allele frequencies of known variants, making variants found in these loci difficult to interpret (Ellingford et al., 2022). This is a challenge very relevant to LRS, as one of its main advantages is the ability to uncover variation in these regions.

In theory, LRS data allows for different approaches like using *de novo* assembly or telomere-to-telomere (T2T) reference genomes to discover novel variation in such regions. In practice, classifying these variants in a diagnostics context is quite difficult, due to a lack of proper annotations with relevant information. Variant interpretation is dependent upon precise location of structural elements such as coding sequences, introns, splice sites, promoters, regulatory motifs and repeats. Furthermore, our clinical diagnostics pipeline uses external databases like GnomAD and ClinVar, which requires variants to be mapped to the human reference genome. This mapping step in itself is challenging, especially for complex SVs where mapping coordinates are often not precise. As it currently stands, even though LRS increases sensitivity of SV detection, prioritizing and pinpointing the small subset of clinically relevant SVs remains a challenge.

5.3 Future Directions

This thesis and many other works have demonstrated the potential of LRS for improving clinical diagnostics by providing a comprehensive catalog of human variation. However, I have also pointed out some potential reasons

why our studies were unsuccessful with respect to finding a genetic cause in these patients, and the future developments that may help us to finally resolve the genetic cause in these patients. However, there are also some other developments that are necessary in order to allow LRS to become a mainstream technology for genetic diagnostics.

5.3.1 We need gold-standard LRS data sets for benchmarking and evaluation

Adopting a new technology for clinical research and diagnostics presents several challenges. Often the first concern is the reliability of results. An ideal clinical test should have high sensitivity to ensure no pathogenic variant is missed and high specificity to prevent false diagnoses. To give an idea of the scale, RadboudUMC has sequenced the exomes of more than 50 thousand patients over a decade. That means a test with a 5% false positive rate will yield almost 3 thousand false diagnoses for patients, which would put a substantial burden on the healthcare system. This is equally true for large scale research projects. This is why limitations of any new technology must be thoroughly understood and documented before it can be scaled up.

One way to address this challenge is the dissemination of best practices through publications, which greatly facilitates the adoption process. These publications can help to establish standard experimental and computational methods, as well as guidelines for evaluating the accuracy of new technologies. An important part of this process is the establishment of publicly available gold standard datasets, which are essential for benchmarking the performance of new sequencing technologies and computational methods. These datasets contain extensively sequenced reference samples, as well as high quality, validated variant call sets. For instance, Genome in a Bottle (GIAB) datasets served this purpose for SRS (Zook et al., 2016). Recently, researchers used CEPH/Utah pedigree from GIAB dataset and relied on Mendelian inheritance to filter variants obtained from Illumina, Nanopore and Pacbio datasets. This process resulted in a very comprehensive truth set for GRCh38, including tandem repeats and SVs (Kronenberg et al., 2024). Furthermore, the recently completed SEQ2 study made mult-platform and cross-validated reference materials that include two commercially available LRS genomes. However, they did not perform any clinical validation with patient samples (Mercer et al., 2021). As it stands, these individual studies provide reasonably comprehensive evaluations, but they still represent starting points that will be continually improved as sequencing technologies and variant callers mature.

A recent development of the benchmarking sets is the extension of the Platinum Genomes dataset (a high-confidence call set for the well-studied NA12878 family), which has long served as a gold standard for method development. In the context of long-read sequencing, Platinum Genomes has been extended to a "Platinum Pedigree" benchmark using PacBio HiFi and ONT data for the entire 17-member CEPH1463 family (Kronenberg et al., 2025). This effort dramatically expanded the variant truth set into difficult genomic regions: the long-read Platinum Pedigree added ~200 Mb of high-confidence sequence not covered by earlier benchmarks and introduced the first rigorous truth calls for tandem repeats and structural variants in NA12878's genome (Kronenberg et al., 2025). The comprehensive variant map now includes ~4.7 million SNVs, 767k indels, >537k repeat variants, and ~24k SVs, allowing researchers to assess both precision and recall of variant callers even in previously "hidden" genomic regions. This is especially relevant for long-read technologies, which aim to call all variant types across the genome. By providing a trusted standard, the Platinum Genomes/Pedigree benchmark enables objective evaluation of long-read variant calling accuracy and informs pipeline improvements. For instance, retraining DeepVariant on the new long-read benchmark data reduced genotyping errors by ~34%, underscoring how such high-quality truth sets drive better performance (Kronenberg et al., 2025). In summary, Platinum Genomes, now augmented with long-read data, remains a cornerstone for benchmarking and has been pivotal in measuring the gains of long-read sequencing in variant detection and ensuring that bioinformatics methods keep pace with the technology's capabilities.

5.3.2 We need orthogonal technologies to experimentally validate difficult variants

The evaluation of novel technologies and the establishment of gold standard datasets are contingent upon the experimental validation of newly discovered variants. In particular, the validation of structural variants poses a persistent challenge (Liu et al., 2022). These variants are frequently located within repetitive or low complexity regions of the genome, rendering conventional methods such as Sanger sequencing and long-range PCR inadequate for their validation. To illustrate, during our validation attempts in Chapter 3, for 4 out of the 23 *de novo* structural variants in repetitive regions we encountered repeated failures due to the unavailability of functional primers for the PCR process.

To address these issues, the exploration of new orthogonal technologies, such as optical mapping, has been shown to be promising (Chan et al., 2018;

Miller et al., 2021). Optical mapping presents a highly cost-efficient solution by overcoming the limitations of current sequencing approaches in complex regions. However, the current optical genome mapping (OGM) technologies suffer from a coarse resolution, as an example SVs detected by Bionano platform typically range from 5 Kbp to several megabase pairs. Although this may allow the validation of SV events from LRS, OGM is not a sequencing technology and is unable to detect the exact breakpoints of an SV event. This is especially relevant in the case of short tandem repeats variation. Furthermore, due to its limited resolution, this technology cannot be used to validate very small variants in complex genome regions.

An alternative to OGM are linked reads generated by platforms like 10x Genomics, which emerged as a cost-efficient alternative to LRS. They offer the capability to pair reads across distances of up to 150 kb (Zheng et al., 2016). Several bioinformatics approaches have been developed to detect structural variations (SVs) using the perturbations in linked reads. These methods typically possess a specific resolution for detecting SVs. Notable techniques employed for analyzing linked reads encompass LongRanger (Stransky et al., 2014), which exhibits a minimum resolution of 50 bp for deletions and 30 kb for rearrangements. Another method called GROC-SVs (Spies et al., 2017) relies on localized assembly and offers a minimum resolution of 10 kbp. As with optical mapping, these ranges mean linked reads cannot be easily utilized for validating SVs at any size. Nevertheless, both of these techniques offer valuable support for the validation of challenging-to-detect variants and have the potential to enhance the accuracy of variant calling.

5.3.3 We need tools and databases to help the interpretation of new variants

The interpretation of newly discovered variants is a significant challenge in diagnostic practices, as previously discussed. Evaluating the clinical relevance of structural variants (SVs) is notably challenging, partly due to the technical obstacles involved in their detection, which often demand hands-on examination of the sequencing data. Considering the substantial quantity of SVs identified, it is essential to employ an initial filtering strategy that focuses on whether they intersect with genes or the coding parts of the genome known as exons. However, this method does not extend to SVs in non-coding regions, despite these constituting the majority of SVs found. This means that many potentially significant SVs may not be captured if we only consider those overlapping with genic or exonic areas.

In future investigations, an alternative approach involving the integration of multiple types of data could aid in interpreting structural variants in non-coding regions. For example, functional genomics studies have revealed an extensive catalog of regulatory elements in the genome that impact gene expression. By mapping these elements to structural variants, their influence on gene expression can be unveiled (Collins et al., 2022), including SVs in the non-coding regions. The recently published PGG.SV database of structural variants contains annotations from approximately 1,000 long-read sequencing genomes, serving as a significant step forward (Wang et al., 2023). The ultimate goal is to utilize these annotations in conjunction with gene expression databases to gain a comprehensive understanding of the phenotype.

Recently, a draft human pangenome reference has been published (Liao *et al*, 2023) providing yet another resource for improving structural variants. Unlike the standard human reference genome which has a linear sequence, human pangenome incorporates genomic variants from 47 diverse individuals in a graph structure. With a pangenome, sequencing reads have a higher likelihood of aligning accurately because the reference includes multiple variant forms, reducing biases and improving the detection of SVs.

The mere availability of various databases, though significant, is insufficient on its own. In order to truly automate the variant interpretation process, the different sources external information needs to be integrated in a framework and assessed together. In light of the rapid advancements in deep learning techniques, Artificial Intelligence (AI) models have emerged as promising tools for this task (Dias & Torkamani, 2019). These models have already been employed to determine the functional effects of noncoding variants, such as spliceAI for predicting splicing variants (Jaganathan et al., 2019) and DeepSEA which is trained on transcription factor binding sites, DNAse 1 hypersensitivity sites and histone mark profiles (Zhou & Troyanskaya, 2015). As clinical decision support tools, AI models can be trained on known disease-causing variants and then can be used to predict the pathogenicity of variants of unknown significance. One recently published model is Fabric GEM, which is trained on genomic variants (SNVs, indels and SVs) and disease phenotypes to suggest a list of causal genes for each patient. When benchmarked on a set of 177 probands with Mendelian disorders, Fabric GEM was able to suggest the correct gene in the top two in 90% of the cases (De La Vega et al., 2021). This is an encouraging result that suggests deep learning-based artificial intelligence models, trained on extensively annotated and validated sets of pathogenic

variants, hold promise for automating the assessment of the pathogenicity of newly discovered variants. One of the challenges in this regard is curation and validation of large variant sets required for training deep learning models.

5.4 Conclusion

In 2022, Nature Methods chose LRS as the method of the year, recognizing its great promise in advancing research and diagnostics. As the technology matures, there is no doubt that LRS will become an indispensable tool in research, with applications like detecting complex SVs in cancer genomes, identifying full-length transcripts and alternative splicing events in transcriptomes and characterizing microbial populations by resolving individual genomes in metagenomics. In diagnostics, the capability of LRS to detect methylation as well as the full spectrum of genetic variants makes it perhaps the holy grail of first-tier tests. However, routine application of LRS in diagnostics and research requires a fundamental shift in how we analyze genomic data. *De novo* genome assembly of each sample, with automated annotation and interpretation of variants, should be the norm in the long term. This will finally allow an in-depth and unbiased exploration of human genetic diversity that is not constrained by the limitations of existing reference genomes, which are inherently biased towards known variations and well-studied populations.

6. Summary

The identification of pathogenic genetic variation remains a central challenge in the diagnosis of rare diseases, particularly for genetically heterogeneous disorders such as intellectual disability (ID). While next-generation sequencing (NGS) has substantially improved diagnostic yield over the past decade, limitations in coverage uniformity, structural variant (SV) detection, and resolution in repetitive or GC-rich regions continue to hinder comprehensive variant discovery.

This thesis investigates the potential of emerging sequencing technologies to improve the sensitivity and resolution of genomic diagnostics. In Chapter 2, the performance of three exome capture kits and both short-read and long-read whole genome sequencing (WGS) was evaluated. The Twist exome capture platform demonstrated superior coverage uniformity and completeness across coding regions, enabling reliable detection of single nucleotide variants (SNVs) and copy number variants (CNVs) even at reduced average sequencing depths.

In Chapter 3, long-read sequencing (LRS) using Pacific Biosciences HiFi technology was applied to a cohort of five trios with unresolved ID despite prior short-read WES and WGS. LRS enabled access to ~35 Mb of genomic regions inaccessible to short reads and identified a large number of previously undetected SVs. Although no pathogenic de novo SVs were identified, LRS uncovered coding variants in known disease genes located within these inaccessible regions, demonstrating its added diagnostic potential.

Chapter 4 assessed whether the improved base-level accuracy of high-fidelity LRS enables robust detection of de novo SNVs. By sequencing eight trios with both short- and long-read platforms, the study revealed high concordance in SNV calls and Mendelian inheritance, indicating that LRS can now reliably detect de novo point mutations, further supporting its application in clinical genomics.

Collectively, the findings of this thesis highlight the diagnostic value of novel sequencing technologies. Improved exome capture and long-read sequencing platforms increase variant detection sensitivity and broaden the accessible mutational landscape. Future implementation in clinical practice will require standardized analytical pipelines, expanded reference datasets, and further validation of the clinical relevance of novel findings. Nonetheless, these technologies represent a critical advance toward more comprehensive and accurate genetic diagnosis of rare diseases.

7. Samenvatting

In het verleden heeft het gebruik van nieuwe technologieën ons begrip van genetische aandoeningen aanzienlijk verbeterd en is het aantal patiënten met een moleculaire diagnose toegenomen. Toch brengen deze nieuwe technologieën ook uitdagingen met zich mee, omdat elke methode unieke eigenschappen heeft die een specifieke aanpak vereisen voor data-analyse en interpretatie. Dit proefschrift onderzoekt hoe recente ontwikkelingen, zoals verbeterde exoomverrijking en long-read sequencing (LRS), kunnen bijdragen aan de detectie van genetische variatie bij patiënten met zeldzame aandoeningen.

In **hoofdstuk 2** vergeleken we drie verschillende exoomverrijkingstechnieken met genoomsequencing. De prestaties van de nieuwe Twist-exoomverrijkingstechniek werden geëvalueerd op basis van dekking en variantdetectie. We concludeerden dat Twist superieur is aan bestaande exoomkits, met name door de meer uniforme dekking, wat resulteert in vergelijkbare gevoeligheid voor variantdetectie bij lagere sequentiediepte dan bij oudere technieken.

In hoofdstuk 3 onderzochten we of long-read sequencing (LRS) structurele varianten kan identificeren die gemist worden met short-read sequencing (SRS). Door LRS toe te passen op vijf ouder-kind-trio's vonden we dat LRS een groter deel van het genoom dekt en aanzienlijk meer structurele varianten detecteert, waaronder inserties en deleties, wat de toegevoegde waarde van deze techniek voor onopgeloste genetische gevallen bevestigt.

Hoofdstuk 4 evalueerde de nauwkeurigheid van LRS bij het opsporen van basepaarvariatie, met name *de novo* puntmutaties. Acht trios werden zowel met LRS als met SRS sequenced en de resultaten werden vergeleken. We concludeerden dat LRS in toenemende mate vergelijkbare of zelfs betere prestaties levert dan SRS bij het detecteren van puntmutaties, waardoor het potentieel heeft om SRS in diagnostische toepassingen te vervangen.

Hoofdstuk 5 biedt een overkoepelende bespreking van de belangrijkste bevindingen. We benadrukken dat technologische vooruitgang in sequencing aanzienlijk bijdraagt aan de moleculaire diagnostiek van zeldzame ziekten. Tegelijkertijd bespreken we ook beperkingen, zoals het gebrek aan gestandaardiseerde analysetools voor LRS-data, de noodzaak van grotere cohorten om de klinische relevantie van bevindingen te onderbouwen, en de complexiteit van variantinterpretatie bij onvoldoende geannoteerde regio's. Toekomstig onderzoek zal zich moeten richten op standaardisatie van data-

analyse, betere interpretatiemethoden en integratie van nieuwe technologieën in klinische workflows.

Samenvattend toont dit proefschrift aan dat verbeterde exoomverrijking en LRS substantiële voordelen bieden voor de detectie van genetische varianten, wat belangrijke implicaties heeft voor zowel onderzoek als klinische diagnostiek van zeldzame genetische aandoeningen.

7

Appendices

References

Acuna-Hidalgo, R., Bo, T., Kwint, M. P., Van De Vorst, M., Pinelli, M., Veltman, J. A., Hoischen, A., Vissers, L. E. L. M., & Gilissen, C. (2015). Post-zygotic Point Mutations Are an Underrecognized Source of *de Novo* Genomic Variation. *American Journal of Human Genetics*, *97*(1), 67–74. https://doi.org/10.1016/j.ajhg.2015.05.008

Acuna-Hidalgo, R., Veltman, J. A., & Hoischen, A. (2016). New insights into the generation and role of *de novo* mutations in health and disease. *Genome Biology*, *17*(1). https://doi.org/10.1186/S13059-016-1110-1

Aiuti, A., Cattaneo, F., Galimberti, S., Benninghoff, U., Cassani, B., Callegaro, L., Scaramuzza, S., Andolfi, G., Mirolo, M., Brigida, I., Tabucchi, A., Carlucci, F., Eibl, M., Aker, M., Slavin, S., Al-Mousa, H., Al Ghonaium, A., Ferster, A., Duppenthaler, A., … Roncarolo, M.-G. (2009). Gene Therapy for Immunodeficiency Due to Adenosine Deaminase Deficiency. *New England Journal of Medicine*, *360*(5), 447–458. https://doi.org/10.1056/NEJMOA0805817/SUPPL_FILE/NEJM_AIUTI_447SA1.PDF

AlAbdi, L., Shamseldin, H. E., Khouj, E., Helaby, R., Aljamal, B., Alqahtani, M., … Alkuraya, F. S. (2023). Beyond the exome: Utility of long-read whole genome sequencing in exome-negative autosomal recessive diseases. Genome Medicine, 15(1), Article 114. https://doi.org/10.1186/s13073-023-01270-8

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet*, *12*(5), 363–376. https://doi.org/10.1038/nrg2958

Allen, A. S., Berkovic, S. F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E. E., Epstein, M. P., Glauser, T., Goldstein, D. B., Han, Y., Heinzen, E. L., Hitomi, Y., Howell, K. B., Johnson, M. R., Kuzniecky, R., Lowenstein, D. H., Lu, Y. F., Madou, M. R. Z., Marson, A. G., … Winawer, M. R. (2013). *De novo* mutations in epileptic encephalopathies. *Nature*, *501*(7466), 217–221. https://doi.org/10.1038/NATURE12439

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology 2020 21:1*, *21*(1), 1–16. https://doi.org/10.1186/S13059-020-1935-5

Arts, P., Simons, A., Alzahrani, M. S., Yilmaz, E., Alidrissi, E., Van Aerde, K. J., Alenezi, N., Alghamdi, H. A., Aljubab, H. A., Al-Hussaini, A. A., Almanjomi, F., Alsaad, A. B., Alsaleem, B., Andijani, A. A., Asery, A., Ballourah, W., Bleeker-Rovers, C. P., Van Deuren, M., Van Der Flier, M., … Hoischen, A. (2019). Exome sequencing in routine diagnostics: A generic test for 254 patients with primary immunodeficiencies. *Genome Medicine*, *11*(1). https://doi.org/10.1186/s13073-019-0649-3

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., Warren, W. C., Magrini, V., McGrath, S. D., Li, Y. I., Wilson, R. K., & Eichler, E. E. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, *176*(3), 663-675 e19. https://doi.org/10.1016/j.cell.2018.12.019

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 1–6. https://doi.org/10.1186/S13100-015-0041-9/TABLES/2

Barbitoff, Y. A., Abasov, R., Tvorogova, V. E., Glotov, A. S., & Predeus, A. V. (2022). Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. *BMC Genomics*, *23*(1), 1–17. https://doi.org/10.1186/S12864-022-08365-3/FIGURES/5

Barbitoff, Y. A., Polev, D. E., Glotov, A. S., Serebryakova, E. A., Shcherbakova, I. V., Kiselev, A. M., Kostareva, A. A., Glotov, O. S., & Predeus, A. V. (2020). Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Scientific Reports*, *10*(1), 1–13. https://doi.org/10.1038/s41598-020-59026-y

Barbosa, M., Joshi, R. S., Garg, P., Martin-Trujillo, A., Patel, N., Jadhav, B., Watson, C. T., Gibson, W., Chetnik, K., Tessereau, C., Mei, H., De Rubeis, S., Reichert, J., Lopes, F., Vissers, L. E. L. M., Kleefstra, T., Grice, D. E., Edelmann, L., Soares, G., … Sharp, A. J. (2018). Identification of rare *de novo* epigenetic variations in congenital disorders. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-04540-x

Belova, V., Shmitko, A., Pavlova, A., Afasizhev, R., Cheranev, V., Tabanakova, A., Ponikarovskaya, N., Rebrikov, D., & Korostin, D. (2022). Performance comparison of Agilent new SureSelect All Exon v8 probes with v7 probes for exome sequencing. *BMC Genomics*, *23*(1), 1–8. https://doi.org/10.1186/S12864-022-08825-W/TABLES/1

Belyeu, J. R., Brand, H., Wang, H., Zhao, X., Pedersen, B. S., Feusier, J., Gupta, M., Nicholas, T. J., Brown, J., Baird, L., Devlin, B., Sanders, S. J., Jorde, L. B., Talkowski, M. E., & Quinlan, A. R. (2021). *De novo* structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *American Journal of Human Genetics*, *108*(4), 597–607. https://doi.org/10.1016/J.AJHG.2021.02.012

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, *40*(10), e72. https://doi.org/10.1093/NAR/GKS001

Bennett, J., Wellman, J., Marshall, K. A., McCague, S., Ashtari, M., DiStefano-Pappas, J., Elci, O. U., Chung, D. C., Sun, J., Wright, J. F., Cross, D. R., Aravand, P., Cyckowski, L. L., Bennicelli, J. L., Mingozzi, F., Auricchio, A., Pierce, E. A., Ruggiero, J., Leroy, B. P., … Maguire, A. M. (2016). Safety and durability of effect of contralateral-eye administration of AAV2 gene therapy in patients with childhood-onset blindness caused by RPE65 mutations: a follow-on phase 1 trial. *The Lancet*, *388*(10045), 661–672. https://doi.org/10.1016/S0140-6736(16)30371-3

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. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature Genetics 2021 53:6*, *53*(6), 779–786. https://doi.org/10.1038/s41588-021-00865-4

Billant, O., Léon, A., Le Guellec, S., Friocourt, G., Blondel, M., & Voisset, C. (2016). The dominant-negative interplay between p53, p63 and p73: A family affair. *Oncotarget*, *7*(43), 69549. https://doi.org/10.18632/ONCOTARGET.11774

Biosciences, P. (2023). *More HiFi reads, more samples, more discovery*.

Bustos, F., Espejo-Serrano, C., Segarra-Fas, A., Toth, R., Eaton, A. J., Kernohan, K. D., Wilson, M. J., Riley, L. G., & Findlay, G. M. (2021). A novel RLIM/RNF12 variant disrupts protein stability and function to cause severe Tonne–Kalscheuer syndrome. *Scientific Reports 2021 11:1*, *11*(1), 1–9. https://doi.org/10.1038/s41598-021-88911-3

Cameron, D. L., Di Stefano, L., & Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-11146-4

Carvalho, C. M., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet*, *17*(4), 224–238. https://doi.org/10.1038/nrg.2015.25

Chaisson, M. J., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J., & Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*(7536), 608–611. https://doi.org/10.1038/nature13907

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). https://doi.org/10.1038/s41467-018-08148-z

Chan, S., Lam, E., Saghbini, M., Bocklandt, S., Hastie, A., Cao, H., Holmlin, E., & Borodkin, M. (2018). Structural variation detection and analysis using bionano optical mapping. In *Methods in Molecular Biology* (Vol. 1833, pp. 193–203). Humana Press Inc. https://doi.org/10.1007/978-1-4939-8666-8_16

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Kallberg, 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*, *32*(8), 1220–1222. https://doi.org/10.1093/bioinformatics/btv710

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R., & Hall, I. M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods*, *12*(10), 966–968. https://doi.org/10.1038/nmeth.3505

Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Chen, R., Euskirchen, G., Butte, A. J., & Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, *29*(10), 908–916. https://doi.org/10.1038/nbt.1975

Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., Kathiresan, S., Kenny, E. E., Lindgren, C. M., MacArthur, D. G., North, K. N., Plon, S. E., Rehm, H. L., Risch, N., Rotimi, C. N., Shendure, J., Soranzo, N., & McCarthy, M. I. (2020). A brief history of human disease genetics. *Nature*, *577*(7789), 179. https://doi.org/10.1038/S41586-019-1879-7

Coe, B. P., Witherspoon, K., Rosenfeld, J. A., van Bon, B. W., Vulto-van Silfhout, A. T., Bosco, P., Friend, K. L., Baker, C., Buono, S., Vissers, L. E., 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. *Nat Genet*, *46*(10), 1063–1071. https://doi.org/10.1038/ng.3092

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature 2020 581:7809*, *581*(7809), 444–451. https://doi.org/10.1038/s41586-020-2287-8

Collins, R. L., Glessner, J. T., Porcu, E., Lepamets, M., Brandon, R., Lauricella, C., Han, L., Morley, T., Niestroj, L. M., Ulirsch, J., Everett, S., Howrigan, D. P., Boone, P. M., Fu, J., Karczewski, K. J., Kellaris, G., Lowther, C., Lucente, D., Mohajeri, K., … Talkowski, M. E. (2022). A cross-disorder dosage sensitivity map of the human genome. *Cell*, *185*(16), 3041-3055.e25. https://doi.org/10.1016/J.CELL.2022.06.036

Conlin, L. K., Aref-Eshghi, E., McEldrew, D. A., Luo, M., & Rajagopalan, R. (2022). Long-read sequencing for molecular diagnostics in constitutional genetic disorders. *Human Mutation*, *43*(11), 1531–1544. https://doi.org/10.1002/HUMU.24465

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., MacArthur, D. G., MacDonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., … Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, *464*(7289), 704. https://doi.org/10.1038/NATURE08516

Cooper, G. M., Nickerson, D. A., & Eichler, E. E. (2007). Mutational and selective effects on copy-number variants in the human genome. *Nat Genet*, *39*(7 Suppl), S22-9. https://doi.org/10.1038/ng2054

Cooper, G. M., Zerr, T., Kidd, J. M., Eichler, E. E., & Nickerson, D. A. (2008). Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet*, *40*(10), 1199–1203. https://doi.org/10.1038/ng.236

Corominas, J., Smeekens, S. P., Nelen, M. R., Yntema, H. G., Kamsteeg, E. J., Pfundt, R., & Gilissen, C. (2022). Clinical exome sequencing–Mistakes and caveats. *Human Mutation*, *43*(8), 1041. https://doi.org/10.1002/HUMU.24360

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K. J., Gall, A., Girón, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., … Flicek, P. (2019). Ensembl 2019. *Nucleic Acids Research*, *47*(D1), D745–D751. https://doi.org/10.1093/nar/gky1113

Cutting, G. R. (2014). Cystic fibrosis genetics: from molecular understanding to clinical application. *Nature Publishing Group*. https://doi.org/10.1038/nrg3849

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., Mcvean, G., Durbin, R., & Project, G. (2011). The variant call format and VCFtools. *BIOINFORMATICS APPLICATIONS NOTE*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

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), 1–4. https://doi.org/10.1093/GIGASCIENCE/GIAB008

De Coster, W., Weissensteiner, M. H., & Sedlazeck, F. J. (2021). Towards population-scale long-read sequencing. *Nature Reviews. Genetics*, *22*(9), 572. https://doi.org/10.1038/S41576-021-00367-3

De La Vega, F. M., Chowdhury, S., Moore, B., Frise, E., McCarthy, J., Hernandez, E. J., Wong, T., James, K., Guidugli, L., Agrawal, P. B., Genetti, C. A., Brownstein, C. A., Beggs, A. H., Löscher, B. S., Franke, A., Boone, B., Levy, S. E., Õunap, K., Pajusalu, S., … Kingsmore, S. F. (2021). Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. *Genome Medicine*, *13*(1). https://doi.org/10.1186/S13073-021-00965-0

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. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. *New England Journal of Medicine*, *367*(20), 1921–1929. https://doi.org/10.1056/NEJMoa1206524

der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media.

Deutschbauer, A. M., Jaramillo, D. F., Proctor, M., Kumm, J., Hillenmeyer, M. E., Davis, R. W., Nislow, C., & Giaever, G. (2005). Mechanisms of Haploinsufficiency Revealed by Genome-Wide Profiling in Yeast. *Genetics*, *169*(4), 1915. https://doi.org/10.1534/GENETICS.104.036871

Dias, R., & Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine*, *11*(1), 1–12. https://doi.org/10.1186/S13073-019-0689-8/FIGURES/1

Díaz-de Usera, A., Lorenzo-Salazar, J. M., Rubio-Rodríguez, L. A., Muñoz-Barrera, A., Guillen-Guio, B., Marcelino-Rodríguez, I., García-Olivares, V., Mendoza-Alvarez, A., Corrales, A., Íñigo-Campos, A., González-Montelongo, R., & Flores, C. (2020). Evaluation of Whole-Exome Enrichment Solutions: Lessons from the High-End of the Short-Read Sequencing Scale. *Journal of Clinical Medicine*, *9*(11), 3656. https://doi.org/10.3390/jcm9113656

Diaz-Horta, O., Bademci, G., Tokgoz-Yilmaz, S., Guo, S., Zafeer, F., Sineni, C. J., Duman, D., Farooq, A., Tekin, M., & Macdonald Foundation, J. T. (2019). *Novel variant p.E269K confirms causative role of PLS1 mutations in autosomal dominant hearing loss*. https://doi.org/10.1111/cge.13626

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., Ibá ~ Nez, K., Tucci, A., Lajoie, B. R., Veldink, J. H., Raymond, F. L., … Eberle, M. A. (n.d.). *ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions*. https://doi.org/10.1093/bioinformatics/btz431

Dong, X., Liu, B., Yang, L., Wang, H., Wu, B., Chen, X., Yu, S., Chen, B., Wang, S., Xu, X., Zhou, W., & Lu, Y. (2020a). Clinical exome sequencing as the first-tier test for diagnosing developmental disorders covering both CNV and SNV: a Chinese cohort Diagnostics. *J Med Genet*, *57*, 558–566. https://doi.org/10.1136/jmedgenet-2019-106377

Dong, X., Liu, B., Yang, L., Wang, H., Wu, B., Chen, X., Yu, S., Chen, B., Wang, S., Xu, X., Zhou, W., & Lu, Y. (2020b). Clinical exome sequencing as the first-tier test for diagnosing developmental disorders covering both CNV and SNV: a Chinese cohort Diagnostics. *J Med Genet*, *57*, 558–566. https://doi.org/10.1136/jmedgenet-2019-106377

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), 1–23. https://doi.org/10.1186/s13059-019-1707-2

Edge, P., & Bansal, V. (2019). Longshot: accurate variant calling in diploid genomes using single-molecule long read sequencing. *BioRxiv*.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138. https://doi.org/10.1126/science.1162986

Eisfeldt, J., Ameur, A., Lenner, F., Ten Berk de Boer, E., Ek, M., Wincent, J., Vaz, R., Ottosson, J., Jonson, T., Ivarsson, S., Thunström, S., Topa, A., Stenberg, S., Rohlin, A., Sandestig, A., Nordling, M., Palmebäck, P., Burstedt, M., Nordin, F., … Lindstrand, A. (2024). A national long-read sequencing study on chromosomal rearrangements uncovers hidden complexities. Genome Research, 34(11), 1774–1784. https://doi.org/10.1101/gr.279510.124

Ellingford, J. M., Ahn, J. W., Bagnall, R. D., Baralle, D., Barton, S., Campbell, C., Downes, K., Ellard, S., Duff-Farrier, C., FitzPatrick, D. R., Greally, J. M., Ingles, J., Krishnan, N., Lord, J., Martin, H. C., Newman, W. G., O'Donnell-Luria, A., Ramsden, S. C., Rehm, H. L., … Whiffin, N. (2022). Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Medicine*, *14*(1), 1–19. https://doi.org/10.1186/S13073-022-01073-3/FIGURES/3

English, A. C., Menon, V. K., Gibbs, R., Metcalf, G. A., & Sedlazeck, F. J. (n.d.). *Truvari: Refined Structural Variant Comparison Preserves Allelic Diversity*. https://doi.org/10.1101/2022.02.21.481353

Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: Description, history and methods to detect structural variation. *Briefings in Functional Genomics*, *14*(5), 305–314. https://doi.org/10.1093/bfgp/elv014

Fabián-Morales, G. E., Ordoñez-Labastida, V., … Zenteno, J. C. (2024). Resolving the diagnostic odyssey in inherited retinal dystrophies through long-read genome sequencing (preprint). medRxiv. https://doi.org/10.1101/2024.08.28.24312668

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). https://doi.org/10.1016/J.XHGG.2021.100046

Farek, J., Hughes, D., Mansfield, A., Krasheninina, O., Nasser, W., Sedlazeck, F. J., Khan, Z., Venner, E., Metcalf, G., Boerwinkle, E., Muzny, D. M., Gibbs, R. A., & Salerno, W. (2018). xAtlas: Scalable small variant calling across heterogeneous next-generation sequencing experiments. *BioRxiv*.

Farwell, K. D., Shahmirzadi, L., El-Khechen, D., Powis, Z., Chao, E. C., Tippin Davis, B., Baxter, R. M., Zeng, W., Mroske, C., Parra, M. C., Gandomi, S. K., Lu, I., Li, X., Lu, H., Lu, H. M., Salvador, D., Ruble, D., Lao, M., Fischbach, S., … Tang, S. (2015). Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: Results from 500 unselected families with undiagnosed genetic conditions. *Genetics in Medicine*, *17*(7), 578–586. https://doi.org/10.1038/gim.2014.154

Fernandez-Novo, S., Pazos, F., & Chagoyen, M. (2016). Rare disease relations through common genes and protein interactions. *Molecular and Cellular Probes*, *30*(3), 178–181. https://doi.org/10.1016/J.MCP.2016.03.004

Gardner, P. P., Paterson, J. M., McGimpsey, S., Ashari-Ghomi, F., Umu, S. U., Pawlik, A., Gavryushkin, A., & Black, M. A. (2022). Sustained software development, not number of citations or journal choice, is indicative of accurate bioinformatic software. *Genome Biology*, *23*(1), 1–13. https://doi.org/10.1186/S13059-022-02625-X/FIGURES/2

Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., & Muller, J. (2018). AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*, *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

Gilissen, C., Hoischen, A., Brunner, H. G., & Veltman, J. A. (2011). Unlocking Mendelian disease using exome sequencing. *Genome Biology*, *12*(9). https://doi.org/10.1186/GB-2011-12-9-228

Girirajan, S., Dennis, M. Y., Baker, C., Malig, M., Coe, B. P., Campbell, C. D., Mark, K., Vu, T. H., Alkan, C., Cheng, Z., Biesecker, L. G., Bernier, R., & Eichler, E. E. (2013). Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *American Journal of Human Genetics*, *92*(2), 221–237. https://doi.org/10.1016/J.AJHG.2012.12.016

Goldmann, J. M., Wong, W. S. W., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A. B., Glusman, G., Vissers, L. E. L. M., Hoischen, A., Roach, J. C., Vockley, J. G., Veltman, J. A., Solomon, B. D., Gilissen, C., & Niederhuber, J. E. (2016). Parent-of-origin-specific signatures of *de novo* mutations. *Nature Genetics*, *48*(8), 935–939. https://doi.org/10.1038/NG.3597

Gullapalli, R. R. (2020). Evaluation of Commercial Next-Generation Sequencing Bioinformatics Software Solutions. *The Journal of Molecular Diagnostics : JMD*, *22*(2), 147–158. https://doi.org/10.1016/J.JMOLDX.2019.09.007

Hamdan, F. F., Srour, M., Capo-Chichi, J. M., Daoud, H., Nassif, C., Patry, L., Massicotte, C., Ambalavanan, A., Spiegelman, D., Diallo, O., Henrion, E., Dionne-Laporte, A., Fougerat, A., Pshezhetsky, A. V., Venkateswaran, S., Rouleau, G. A., & Michaud, J. L. (2014). *De novo* mutations in moderate or severe intellectual disability. *PLoS Genetics*, *10*(10). https://doi.org/10.1371/JOURNAL.PGEN.1004772

Hartman, P., Beckman, K., Silverstein, K., Yohe, S., Schomaker, M., Henzler, C., Onsongo, G., Lam, H. C., Munro, S., Daniel, J., Billstein, B., Deshpande, A., Hauge, A., Mroz, P., Lee, W., Holle, J., Wiens, K., Karnuth, K., Kemmer, T., … Thyagarajan, B. (2019). Next generation sequencing for clinical diagnostics: Five year experience of an academic laboratory. *Molecular Genetics and Metabolism Reports*, *19*, 100464. https://doi.org/10.1016/J.YMGMR.2019.100464

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/J.YGENO.2015.11.003

Hiatt, S. M., Lawlor, J. M. J., Handley, L. H., Latner, D. R., Bonnstetter, Z. T., Finnila, C. R., … Cooper, G. M. (2024). Long-read genome sequencing and variant reanalysis increase diagnostic yield in neurodevelopmental disorders. Genome Research, 34(11), 1747–1762. https://doi.org/10.1101/gr.279227.124

Hiller, M., Chrysostomakis, I., Arantes, L. S., et al. (2025). Long-read sequencing and genome assembly of challenging specimens. Genome Biology, 26: 25

Ho, S. S., Urban, A. E., & Mills, R. E. (n.d.). *Structural variation in the sequencing era*. https://doi.org/10.1038/s41576-019-0180-9

Hoischen, A., Krumm, N., & Eichler, E. E. (2014). Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nat Neurosci*, *17*(6), 764–772. https://doi.org/10.1038/nn.3703

Hon, T., Mars, K., Young, G., Tsai, Y. C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P., & Rank, D. R. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data 2020 7:1*, *7*(1), 1–11. https://doi.org/10.1038/s41597-020-00743-4

Höps, W., Weiss, M. M., Derks, R., Galbany, J. C., Ouden, A. den, 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. (2025). HiFi long-read genomes for difficult-to-detect, clinically relevant variants. *American Journal of Human Genetics*, *112*(2), 450–456. https://doi.org/10.1016/J.AJHG.2024.12.013

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 Res*, *27*(5), 677–685. https://doi.org/10.1101/gr.214007.116

Iadarola, B., Xumerle, L., Lavezzari, D., Paterno, M., Marcolungo, L., Beltrami, C., Fortunati, E., Mei, D., Vetro, A., Guerrini, R., Parrini, E., Rossato, M., & Delledonne, M. (2020). *Shedding light on dark genes: enhanced targeted resequencing by optimizing the combination of enrichment technology and DnA fragment length*. https://doi.org/10.1038/s41598-020-66331-z

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., Smith, J. D., Paeper, B., Nickerson, D. A., Dea, J., Dong, S., Gonzalez, L. E., Mandell, J. D., Mane, S. M., Murtha, M. T., … Wigler, M. (2014). The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature*, *515*(7526), 216–221. https://doi.org/10.1038/NATURE13908

Jain M., Koren S., Miga K.H., Quick J., Rand A.C., Sasani T.A., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology 36(4): 338–345. doi:10.1038/nbt.4060.

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

Jamal, S. M., Yu, J.-H., Chong, J. X., Dent, K. M., Conta, J. H., Tabor, H. K., & Bamshad, M. J. (2013). Practices and Policies of Clinical Exome Sequencing Providers: Analysis and Implications. *American Journal of Medical Genetics Part A*, *161*(5), 935–950. https://doi.org/https://doi.org/10.1002/ajmg.a.35942

Javadzadeh, S., Adamson, A., Park, J., Jo, S.-Y., Ding, Y.-C., Bakhtiari, M., et al. (2025). Analysis of targeted and whole genome sequencing of PacBio HiFi reads for comprehensive genotyping of VNTRs. PLoS Computational Biology, 21(4): e1012885

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 2020 586:7831*, *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., … Daly, M. J. (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

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, *32*(Database issue), D493–D496. https://doi.org/10.1093/nar/gkh103

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*. https://doi.org/10.1093/NAR/GKAC511

Kim, J., Cho, S., Han, Y., et al. (2024). Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. Human Genome Variation, 11: 33

King, D. A., Sifrim, A., Fitzgerald, T. W., Rahbari, R., Hobson, E., Homfray, T., Mansour, S., Mehta, S. G., Shehla, M., Tomkins, S. E., Vasudevan, P. C., & Hurles, M. E. (2017). Detection of structural mosaicism from targeted and whole-genome sequencing data. *Genome Research*, *27*(10), 1704–1714. https://doi.org/10.1101/GR.212373.116

Kobayashi, E. S., Batalov, S., Wenger, A. M., Lambert, C., Dhillon, H., Hall, R. J., ... Bainbridge, M. N. (2022). Approaches to long-read sequencing in a clinical setting to improve diagnostic rate. Scientific Reports, 12(1), 16945. https://doi.org/10.1038/s41598-022-20113-x

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine 2020 12:1*, *12*(1), 1–13. https://doi.org/10.1186/S13073-020-00791-W

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W. S. W., Sigurdsson, G., Walters, G. B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D. F., Helgason, A., Magnusson, O. T., ... Stefansson, K. (2012). Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature*, *488*(7412), 471–475. https://doi.org/10.1038/NATURE11396

Kong, S. W., Lee, I. H., Liu, X., Hirschhorn, J. N., & Mandl, K. D. (2018). Measuring coverage and accuracy of whole-exome sequencing in clinical context. *Genetics in Medicine*, *20*(12), 1617–1626. https://doi.org/10.1038/gim.2018.51

Kronenberg, Z., Nolan, C., Porubsky, D., Mokveld, T., Rowell, W. J., Lee, S., Dolzhenko, E., Chang, P.-C., Holt, J. M., Saunders, C. T., Olson, N. D., McGee, S., Guarracino, A., Koundinya, N., Harvey, W. T., Watkins, W. S., Munson, K. M., Hoekzema, K., Chua, K. P., ... Eberle, M. A. (2024). The Platinum Pedigree: A long-read benchmark for genetic variants. *BioRxiv*, 2024.10.02.616333. https://doi.org/10.1101/2024.10.02.616333

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., ... Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature 2001 409:6822*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Lappalainen, T., Scott, A. J., Brandt, M., & Hall, I. M. (2019). Genomic Analysis in the Age of Human Genome Sequencing. *Cell*, *177*(1), 70–84. https://doi.org/10.1016/J.CELL.2019.02.032

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*, *15*(6), R84. https://doi.org/10.1186/gb-2014-15-6-r84

Lek, M., Karczewski, K. J., Minikel, E. V, samocha, K. E., Banks, E., Fennell, timothy, O, anne H., Ware, J., Hill, andrew J., cummings, B. B., tukiainen, taru, Birnbaum, D. P., Kosmicki, J., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., & Berghout, J. (2016). Lorena Orozco 29 , gina M. Peloso 2,27,28 , Ryan Poplin 18 , Manuel a. Rivas 2 , Valentin Ruano-Rubio 18 , samuel a. Rose 6 , Douglas M. *Nature Publishing Group*, *19*(6), 57. https://doi.org/10.1038/nature19057

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/10.1002/humu.22813

Li, H. (2016). Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics*, *32*(14), 2103–2110. https://doi.org/10.1093/bioinformatics/btw152

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, 1000 Genome Project Data Processing. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, W. (2023). Rare-variant genetic architecture. *Nature Genetics*, *55*(3). https://doi.org/10.1038/s41588-023-01354-6

Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., … Paten, B. (2023). A draft human pangenome reference. *Nature 2023 617:7960*, *617*(7960), 312–324. https://doi.org/10.1038/s41586-023-05896-x

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

Lupski, J. R. (2007). Genomic rearrangements and sporadic disease. *Nat Genet*, *39*(7 Suppl), S43-7. https://doi.org/10.1038/ng2084

Lupski, J. R., Belmont, J. W., Boerwinkle, E., & Gibbs, R. A. (2011). Clan genomics and the complex architecture of human disease. *Cell*, *147*(1), 32–43. https://doi.org/10.1016/J.CELL.2011.09.008

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology*, *20*(1). https://doi.org/10.1186/S13059-019-1828-7

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, 837. https://doi.org/10.1038/s41467-024-44804-3

Maia, N., Nabais Sá, M. J., Melo-Pires, M., de Brouwer, A. P. M., & Jorge, P. (2021). Intellectual disability genomics: current state, pitfalls and future challenges. *BMC Genomics*, *22*(1). https://doi.org/10.1186/S12864-021-08227-4

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 2016 18:12*, *18*(12), 1282–1289. https://doi.org/10.1038/gim.2016.58

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

Martinez-Granero, F., Blanco-Kelly, F., Sanchez-Jimeno, C., Avila-Fernandez, A., Arteche, A., Bustamante-Aragones, A., Rodilla, C., Rodríguez-Pinilla, E., Riveiro-Alvarez, R., Tahsin-Swafiri, S., Trujillo-Tiebas, M. J., Ayuso, C., Rodríguez De Alba, M., Lorda-Sanchez, I., & Almoguera, B. (2021). Comparison of the diagnostic yield of aCGH and genome-wide sequencing across different neurodevelopmental disorders. *Npj Genomic Medicine*, *6*(1), 25. https://doi.org/10.1038/s41525-021-00188-7

Marwaha, S., Knowles, J. W., & Ashley, E. A. (2022). A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Medicine*, *14*(1). https://doi.org/10.1186/S13073-022-01026-W

Mathieson, I., & Reich, D. (2017). *Differences in the rare variant spectrum among human populations*. https://doi.org/10.1371/journal.pgen.1006581

McNeill, A. (2022). Exome sequencing-one test to rule them all? *European Journal of Human Genetics : EJHG*, *30*(8), 869. https://doi.org/10.1038/S41431-022-01145-3

Meienberg, J., Zerjavic, K., Keller, I., Okoniewski, M., Patrignani, A., Ludin, K., Xu, Z., Steinmann, B., Carrel, T., Röthlisberger, B., Schlapbach, R., Bruggmann, R., & Matyas, G. (2015). New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Research*, *43*(11). https://doi.org/10.1093/nar/gkv216

Melas, M., Kautto, E. A., Franklin, S. J., Mori, M., McBride, K. L., Mosher, T. M., Pfau, R. B., Hernandez-Gonzalez, M. E., McGrath, S. D., Magrini, V. J., White, P., Samora, J. B., Koboldt, D. C., & Wilson, R. K. (2022). Long-read whole genome sequencing reveals HOXD13 alterations in synpolydactyly. *Human Mutation*, *43*(2), 189–199. https://doi.org/10.1002/HUMU.24304

Mendell, J. R., Al-Zaidy, S., Shell, R., Arnold, W. D., Rodino-Klapac, L. R., Prior, T. W., Lowes, L., Alfano, L., Berry, K., Church, K., Kissel, J. T., Nagendran, S., L'Italien, J., Sproule, D. M., Wells, C., Cardenas, J. A., Heitzer, M. D., Kaspar, A., Corcoran, S., … Kaspar, B. K. (2017). Single-Dose Gene-Replacement Therapy for Spinal Muscular Atrophy. *New England Journal of Medicine*, *377*(18), 1713–1722. https://doi.org/10.1056/NEJMOA1706198/SUPPL_FILE/NEJMOA1706198_DISCLOSURES.PDF

Mercer, T. R., Xu, J., Mason, C. E., & Tong, W. (2021). The Sequencing Quality Control 2 study: establishing community standards for sequencing in precision medicine. *Genome Biology*, *22*(1), 1–7. https://doi.org/10.1186/S13059-021-02528-3/FIGURES/1

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

Miller, D. E., Lee, L., Galey, M., Kandhaya-Pillai, R., Tischkowitz, M., Amalnath, D., … Oshima, J. (2022). Targeted long-read sequencing identifies missing pathogenic variants in unsolved Werner syndrome cases. Journal of Medical Genetics, 59(11), 1087–1094. https://doi.org/10.1136/jmedgenet-2022-108485

Mizuguchi, T., Okamoto, N., Yanagihara, K., Miyatake, S., Uchiyama, Y., Tsuchida, N., Hamanaka, K., Fujita, A., Miyake, N., & Matsumoto, N. (2021). Pathogenic 12-kb copy-neutral inversion in syndromic intellectual disability identified by high-fidelity long-read sequencing. *Genomics*, *113*(1 Pt 2), 1044–1053. https://doi.org/10.1016/J.YGENO.2020.10.038

Mokry, M., Feitsma, H., Nijman, I. J., de Bruijn, E., van der Zaag, P. J., Guryev, V., & Cuppen, E. (2010). Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. *Nucleic Acids Research*, *38*(10). https://doi.org/10.1093/nar/gkq072

Neveling, K., Feenstra, I., Gilissen, † Christian, Lies, †, Hoefsloot, H., Kamsteeg, E.-J., Mensenkamp, A. R., Rodenburg, R. J. T., Yntema, H. G., Spruijt, L., Vermeer, S., Rinne, T., Van Gassen, K. L., Bodmer, D., Lugtenberg, D., De Reuver, R., Buijsman, W., Derks, R. C., Wieskamp, N., … Nelen, M. (2013). A Post-Hoc Comparison of the Utility of Sanger Sequencing and Exome Sequencing for the Diagnosis of Heterogeneous Diseases. *OFFICIAL JOURNAL Www.Hgvs*, *34*(12), 1721–1726. https://doi.org/10.1002/humu.22450

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., & Bamshad, M. J. (2009). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics 2009 42:1*, *42*(1), 30–35. https://doi.org/10.1038/ng.499

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*, *0*(0). https://doi.org/10.1016/J.AJHG.2022.02.014

Noyes, M. D., Sui, Y., Kwon, Y., Koundinya, N., Wong, I., Munson, K. M., Hoekzema, K., Kordosky, J., Garcia, G. H., Knuth, J., Lewis, A. P., & Eichler, E. E. (2025). Long-read sequencing of trios reveals increased germline and postzygotic mutation rates in repetitive DNA. bioRxiv. https://doi.org/10.1101/2025.07.18.665621

Nurk, S., et al. (2022). The complete sequence of a human genome. Science, 376(6588): 44–53

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), btv566. https://doi.org/10.1093/bioinformatics/btv566

Pacific Biosciences of California, Inc. (2023, January 9). PacBio announces record orders… https://www.pacb.com/press_releases/pacbio-announces-record-orders-including-orders-for-76-revio-systems-received-in-the-fourth-quarter-of-2022/

Parla, J. S., Iossifov, I., Grabill, I., Spector, M. S., Kramer, M., & McCombie, W. R. (2011). A comparative analysis of exome capture. *Genome Biology*, *12*(9), R97. https://doi.org/10.1186/GB-2011-12-9-R97

Patterson, M. D., Marschall, T., Pisanti, N., Van Iersel, L., Stougie, L., Klau, G. W., & Schönhuth, A. (2015). WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *22*(6), 498–509. https://doi.org/10.1089/CMB.2014.0157

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., L M Vissers, L. E., & Gilissen, C. (2021). Long-read trio sequencing of individuals with unsolved intellectual disability. *European Journal of Human Genetics*, *29*, 637–648. https://doi.org/10.1038/s41431-020-00770-0

Pedersen, B. S., Brown, J. M., Dashnow, H., Wallace, A. D., Velinder, M., Tristani-Firouzi, M., Schiffman, J. D., Tvrdik, T., Mao, R., Best, D. H., Bayrak-Toydemir, P., & Quinlan, A. R. (2021). Effective variant filtering and expected candidate variant yield in studies of rare human disease. *Npj Genomic Medicine 2021 6:1*, *6*(1), 1–8. https://doi.org/10.1038/s41525-021-00227-3

Pendleton, M., Sebra, R., Pang, A. W., Ummat, A., Franzen, O., Rausch, T., Stutz, A. M., Stedman, W., Anantharaman, T., Hastie, A., Dai, H., Fritz, M. H., Cao, H., Cohain, A., Deikus, G., Durrett, R. E., Blanchard, S. C., Altman, R., Chin, C. S., … Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods*, *12*(8), 780–786. https://doi.org/10.1038/nmeth.3454

Peters, B. A., Kermani, B. G., Alferov, O., Agarwal, M. R., McElwain, M. A., Gulbahce, N., Hayden, D. M., Tang, Y. T., Zhang, R. Y., Tearle, R., Crain, B., Prates, R., Berkeley, A., Munné, S., & Drmanac, R. (2015). Detection and phasing of single base *de novo* mutations in biopsies from human in vitro fertilized embryos by advanced whole-genome sequencing. *Genome Research*, *25*(3), 426. https://doi.org/10.1101/GR.181255.114

Pfundt, R., del Rosario, M., ELM Vissers, L., Kwint, M. P., Janssen, I. M., de Leeuw, N., Yntema, H. G., Nelen, M. R., Lugtenberg, D., Kamsteeg, E.-J., Wieskamp, N., Stegmann, A. P., Stevens, S. J., Rodenburg, R. J., Simons, A., Mensenkamp, A. R., Rinne, T., Gilissen, C., Scheffer, H., … Hehir-Kwa, J. Y. (2017). Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Nature Publishing Group*. https://doi.org/10.1038/gim.2016.163

Pilipenko, V. V., He, H., Kurowski, B. G., Alexander, E. S., Zhang, X., Ding, L., Mersha, T. B., Kottyan, L., Fardo, D. W., & Martin, L. J. (2014). Using Mendelian inheritance errors as quality control criteria in whole genome sequencing data set. *BMC Proceedings*, *8*(Suppl 1), S21. https://doi.org/10.1186/1753-6561-8-S1-S21

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

Porubsky, D., Höps, W., Ashraf, H., Hsieh, P. H., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maria Maggiolini, F. A., Harvey, W. T., Henning, B., Audano, P. A., Gordon, D. S., Ebert, P., Hasenfeld, P., Benito, E., Zhu, Q., Lee, C., Antonacci, F., … Korbel, J. O. (2022). Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*, *185*(11), 1986-2005.e26. https://doi.org/10.1016/J.CELL.2022.04.017/ATTACHMENT/C9CD2D6C-8C00-462D-AF5E-804403AAF113/MMC9.PDF

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., … Ostell, J. M. (2014). RefSeq: An update on mammalian reference sequences. *Nucleic Acids Research*, *42*(D1), 756–763. https://doi.org/10.1093/nar/gkt1114

Przybyla, L., & Gilbert, L. A. (2021). A new era in functional genomics screens. *Nature Reviews Genetics 2021 23:2*, *23*(2), 89–103. https://doi.org/10.1038/s41576-021-00409-w

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

Rabbani, B., Tekin, M., & Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, *59*, 5–15. https://doi.org/10.1038/jhg.2013.114

Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339. https://doi.org/10.1093/bioinformatics/bts378

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. https://doi.org/10.1101/2024.09.30.24314377

Reiner, J., Pisani, L., Qiao, W., Singh, R., Yang, Y., Shi, L., Khan, W. A., Sebra, R., Cohen, N., Babu, A., Edelmann, L., Jabs, E. W., & Scott, S. A. (2018). Cytogenomic identification and long-read single molecule real-time (SMRT) sequencing of a Bardet-Biedl Syndrome 9 (BBS9) deletion. *NPJ Genomic Medicine*, *3*(1). https://doi.org/10.1038/S41525-017-0042-3

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

Royer-Bertrand, B., Cisarova, K., Superti-Furga, A., Niel-Butschi, F., Mittaz-Crettol, L., & Fodstad, H. (2021). CNV detection from exome sequencing data in routine diagnostics of rare genetic disorders: Opportunities and limitations. *Genes*, *12*(9). https://doi.org/10.3390/genes12091427

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., & Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, *46*(9), 944–950. https://doi.org/10.1038/ng.3050

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

Sanden, B. P. G. H. van der, Schobers, G., Galbany, J. C., Koolen, D. A., Sinnema, M., Reeuwijk, J. van, Stumpel, C. T. R. M., Kleefstra, T., Vries, B. B. A. de, Ruiterkamp-Versteeg, M., Leijsten, N., Kwint, M., Derks, R., Swinkels, H., Ouden, A. den, Pfundt, R., Rinne, T., Leeuw, N. de, Stegmann, A. P., … Vissers, L. E. L. M. (n.d.). The performance of genome sequencing as a first-tier test for neurodevelopmental disorders. *In Revision*.

Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C. J., & Lansdorp, P. M. (2017). Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nature Protocols*, *12*(6), 1151–1176. https://doi.org/10.1038/nprot.2017.029

Sanford Kobayashi, E., Batalov, S., Wenger, A. M., Lambert, C., Dhillon, H., Hall, R. J., Baybayan, P., Ding, Y., Rego, S., Wigby, K., Friedman, J., Hobbs, C., & Bainbridge, M. N. (2022). Approaches to long-read sequencing in a clinical setting to improve diagnostic rate. *Scientific Reports 2022 12:1*, *12*(1), 1–7. https://doi.org/10.1038/s41598-022-20113-x

Schloissnig, S., Pani, S., Rodriguez-Martin, B., Ebler, J., Hain, C., Tsapalou, V., Söylev, A., Hüther, P., Ashraf, H., Prodanov, T., Asparuhova, M., Hunt, S., Rausch, T., Marschall, T., & Korbel, J. O. (2024). Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project. *BioRxiv*, 2024.04.18.590093. https://doi.org/10.1101/2024.04.18.590093

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*, *19*(6), 329–346. https://doi.org/10.1038/s41576-018-0003-4

Sekhar, C., Chilamakuri, R., Lorenz, S., Madoui, M.-A., Vodák, D., Sun, J., Hovig, E., Myklebost, O., & Meza-Zepeda, L. A. (2014). Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics*, *15*(449), 1–13.

Seo, J. S., Rhie, A., Kim, J., Lee, S., Sohn, M. H., Kim, C. U., Hastie, A., Cao, H., Yun, J. Y., Kim, J., Kuk, J., Park, G. H., Kim, J., Ryu, H., Kim, J., Roh, M., Baek, J., Hunkapiller, M. W., Korlach, J., … Kim, C. (2016). *De novo* assembly and phasing of a Korean human genome. *Nature*, *538*(7624), 243–247. https://doi.org/10.1038/nature20098

Shakked, B. P., Barel, O., Singer, A., Regev, M., Dayan, E. Z., Zeev, B. Ben, Mor, N., Kol, N., Nayshool, O., Shimshoviz, N., Joseph, I. B., Yagel, D. M., Javasky, E., Einy, R., Gal, M., Cohen, J. G., Shohat, M., & Dominissini, D. (2021). A single center experience with publicly funded clinical exome sequencing for neurodevelopmental disorders or multiple congenital anomalies. *Scientific Reports*, 1–8. https://doi.org/10.1038/s41598-021-98646-w

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, *550*(7676). https://doi.org/10.1038/NATURE24286

Shendure, J., Findlay, G. M., & Snyder, M. W. (2019). Genomic Medicine-Progress, Pitfalls, and Promise. *Cell*, *177*(1), 45–57. https://doi.org/10.1016/J.CELL.2019.02.003

Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S., Lintner, K. E., Ding, Q., Wang, Z., Hu, J., Wang, D., Wang, F., Wang, L., Lyon, G. J., Guan, Y., … Wang, K. (2016). Long-read sequencing and *de novo* assembly of a Chinese genome. *Nat Commun*, *7*, 12065. https://doi.org/10.1038/ncomms12065

Shigemizu, D., Momozawa, Y., Abe, T., Morizono, T., Boroevich, K. A., Takata, S., Ashikawa, K., Kubo, M., & Tsunoda, T. (2015). Performance comparison of four commercial human whole-exome capture platforms. *Scientific Reports 2015 5:1*, *5*(1), 1–8. https://doi.org/10.1038/srep12742

Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J. M., Salit, M., West, R. B., Batzoglou, S., & Sidow, A. (2017). Genome-wide reconstruction of complex structural variants using read clouds. *Nature Methods*, *14*(9), 915–920. https://doi.org/10.1038/nmeth.4366

Srivastava, S., Love-Nichols, J. A., Dies, K. A., Ledbetter, D. H., Martin, C. L., Chung, W. K., Firth, H. V., Frazier, T., Hansen, R. L., Prock, L., Brunner, H., Hoang, N., Scherer, S. W., Sahin, M., & Miller, D. T. (2019). Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genetics in Medicine*, *21*(11), 2413–2421. https://doi.org/10.1038/s41436-019-0554-6

Stawicka, E., Zielińska, A., Górka-Skoczylas, P., et al. (2024). SCN1A – Characterization of the gene's variants in a Polish cohort of patients with Dravet syndrome. Current Issues in Molecular Biology, 46(5), 4437–4451. DOI: 10.3390/cimb46050269.

Steyaert, W., Sagath, L., Demidov, G., Yépez, V. A., Esteve-Codina, A., Gagneur, J., Ellwanger, K., Derks, R., Weiss, M., den Ouden, A., van den Heuvel, S., Swinkels, H., Zomer, N., Steehouwer, M., O'Gorman, L., Astuti, G., Neveling, K., Schüle, R., Xu, J., … Hoischen, A. (2025). Unraveling undiagnosed rare disease cases by HiFi long-read genome sequencing. *Genome Research*. https://doi.org/10.1101/GR.279414.124

Stransky, N., Cerami, E., Schalm, S., Kim, J. L., & Lengauer, C. (2014). The landscape of kinase fusions in cancer. *Nature Communications*, *5*(1), 4846. https://doi.org/10.1038/ncomms5846

Sulonen, A. M., Ellonen, P., Almusa, H., Lepistö, M., Eldfors, S., Hannula, S., Miettinen, T., Tyynismaa, H., Salo, P., Heckman, C., Joensuu, H., Raivio, T., Suomalainen, A., & Saarela, J. (2011). Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biology*, *12*(9). https://doi.org/10.1186/GB-2011-12-9-R94

Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front Bioeng Biotechnol*, *3*, 92. https://doi.org/10.3389/fbioe.2015.00092

van der Sanden, B. P. G. H., Schobers, G., Corominas Galbany, J., Koolen, D. A., Sinnema, M., van Reeuwijk, J., Stumpel, C. T. R. M., Kleefstra, T., de Vries, B. B. A., Ruiterkamp-Versteeg, M., Leijsten, N., Kwint, M., Derks, R., Swinkels, H., den Ouden, A., Pfundt, R., Rinne, T., de Leeuw, N., Stegmann, A. P., … Vissers, L. E. L. M. (2023). The performance of genome sequencing as a first-tier test for neurodevelopmental disorders. *European Journal of Human Genetics : EJHG*, *31*(1), 81–88. https://doi.org/10.1038/S41431-022-01185-9

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

Vissers, L. E. L. M., Gilissen, C., & Veltman, J. A. (2016). Genetic studies in intellectual disability and related disorders. In *Nature Reviews Genetics* (Vol. 17, Issue 1, pp. 9–18). Nature Publishing Group. https://doi.org/10.1038/nrg3999

Vissers, L. E. L. M., Van Nimwegen, K. J. M., Schieving, J. H., Kamsteeg, E. J., Kleefstra, T., Yntema, H. G., Pfundt, R., Van Der Wilt, G. J., Krabbenborg, L., Brunner, H. G., Van Der Burg, S., Grutters, J., Veltman, J. A., & Willemsen, M. A. A. P. (2017). A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *19*(9), 1055–1063. https://doi.org/10.1038/GIM.2017.1

Vollger, M. R., Logsdon, G. A., Audano, P. A., Sulovari, A., Porubsky, D., Peluso, P., Wenger, A. M., Concepcion, G. T., Kronenberg, Z. N., Munson, K. M., Baker, C., Sanders, A. D., Spierings, D. C. J., Lansdorp, P. M., Surti, U., Hunkapiller, M. W., & Eichler, E. E. (2020). Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Annals of Human Genetics*, *84*(2), 125–140. https://doi.org/10.1111/AHG.12364

Vulto-van Silfhout, A. T., Hehir-Kwa, J. Y., van Bon, B. W. M., Schuurs-Hoeijmakers, J. H. M., Meader, S., Hellebrekers, C. J. M., Thoonen, I. J. M., de Brouwer, A. P. M., Brunner, H. G., Webber, C., Pfundt, R., de Leeuw, N., & De Vries, B. B. A. (2013). Clinical Significance of *De Novo* and Inherited Copy-Number Variation. *Human Mutation*, *34*(12), 1679–1687. https://doi.org/10.1002/humu.22442

Wang, J., Raskin, L., Samuels, D. C., Shyr, Y., & Guo, Y. (2015). Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics*, *31*(3), 318–323. https://doi.org/10.1093/bioinformatics/btu668

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, Q., Shashikant, C. S., Jensen, M., Altman, N. S., & Girirajan, S. (2017). Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Scientific Reports*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-01005-x

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., … Haussler, D. (2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature 2022 604:7906*, *604*(7906), 437–446. https://doi.org/10.1038/s41586-022-04601-8

Wang Y et al. (2025). High clinical utility of long-read sequencing for precise diagnosis of congenital adrenal hyperplasia in 322 probands. Hum. Genomics 19(1): 3[7][26].

Weiner, D. J., Nadig, A., Jagadeesh, K. A., Dey, K. K., Neale, B. M., Robinson, E. B., Karczewski, K. J., & O'Connor, L. J. (2023). Polygenic architecture of rare coding variation across 394,783 exomes. *Nature 2023 614:7948*, *614*(7948), 492–499. https://doi.org/10.1038/s41586-022-05684-z

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., & Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, *6*, 100. https://doi.org/10.12688/f1000research.10571.1

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet*, *14*(2), 125–138. https://doi.org/10.1038/nrg3373

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

Wood, K. A., & Goriely, A. (2022). The impact of paternal age on new mutations and disease in the next generation. Fertility and Sterility, 118(6), 1001–1012. DOI: 10.1016/j.fertnstert.2022.10.017.

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. *Https://Doi.Org/10.1056/NEJMoa2209046*. https://doi.org/10.1056/NEJMOA2209046

Wright, C. F., FitzPatrick, D. R., Firth, H. V., et al. (2023). Genomic diagnosis of developmental disorders in a large-scale sequencing study. New England Journal of Medicine, 389(13), 1237–1249. DOI: 10.1056/NEJMoa2209046.

Yang, L. (2020). A practical guide for structural variation detection in human genome. *Current Protocols in Human Genetics*, *107*(1), e103. https://doi.org/10.1002/CPHG.103

Yang, Y., Muzny, D. M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., Veeraraghavan, N., Hawes, A., Chiang, T., Leduc, M., Beuten, J., Zhang, J., He, W., Scull, J., Willis, A., … Eng, C. M. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA - Journal of the American Medical Association*, *312*(18), 1870–1879. https://doi.org/10.1001/jama.2014.14601

Yuan, B., Schulze, K. V., Assia Batzir, N., Sinson, J., Dai, H., Zhu, W., Bocanegra, F., Fong, C. T., Holder, J., Nguyen, J., Schaaf, C. P., Yang, Y., Bi, W., Eng, C., Shaw, C., Lupski, J. R., & Liu, P. (2022). Sequencing individual genomes with recurrent genomic disorder deletions: an approach to characterize genes for autosomal recessive rare disease traits. *Genome Medicine*, *14*(1), 113. https://doi.org/10.1186/S13073-022-01113-Y

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

Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., Mudivarti, P. A., Wyatt, P. W., Bharadwaj, R., Makarewicz, A. J., Li, Y., Belgrader, P., Price, A. D., Lowe, A. J., Marks, P., … Ji, H. P. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, *34*(3), 303–311. https://doi.org/10.1038/nbt.3432

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, *12*(10), 931. https://doi.org/10.1038/NMETH.3547

Zhou, J., Zhang, M., Li, X., Wang, Z., Pan, D., & Shi, Y. (2021). Performance comparison of four types of target enrichment baits for exome DNA sequencing. *Hereditas*, *158*(1). https://doi.org/10.1186/s41065-021-00171-3

Zhou, X., Feliciano, P., Shu, C., et al. (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. DOI: 10.1038/s41588-022-01148-2.

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., Henaff, E., McIntyre, A. B. R., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., … Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data 2016 3:1*, *3*(1), 1–26. https://doi.org/10.1038/sdata.2016.25

Acknowledgements

I am deeply grateful to everyone who made this thesis possible.

First and foremost, I thank my supervisors, Juliet Hampstead and Christian Gilissen, for their guidance, rigor, and above all patience. Their advice sharpened my thinking, their questions raised my standards, and their trust gave me the space to grow into an independent researcher. I am also incredibly grateful to Alexander Hoischen and Lisenka Vissers for their collaboration and enthusiasm throughout my PhD journey. I would like to extend a special thanks to Han Brunner, my original *promotor*, for his encouragement and critical insights during the initial phase of my PhD before his retirement.

This work would not have been possible without the collaboration of my fellow researchers, and I couldn't have asked for better ones. I owe a special thanks to my shared first authors with whom I worked closely on the core chapters of this thesis. I thank Burcu Yaldiz for our work together on the Twist exome capture evaluation, Marc Pauper for our collaboration on the long-read trio sequencing project, and Bart van der Sanden for our joint efforts on the comprehensive HiFi sequencing study.

This thesis also benefited from the broader community and infrastructure at the Department of Human Genetics, Radboud University Medical Center. Thank you to everyone who provided the essential assistance with data handling and analysis. I also wish to extend my gratitude to the team at Pacific Biosciences for their collaboration and technical expertise regarding SMRT and HiFi sequencing.

Finally, to my family and friends: thank you for your patience, encouragement, and unwavering faith during the long and arduous journey of my PhD. This thesis is as much yours as it is mine.

CV

Muhammet Erdi Küçük
Bischofplatz 16, Dresden 01097, Saxony, Germany
erdikucuk@gmail.com

Academic Appointments and Work Experience

NGS Bioinformatician – Max Planck Institute of Cell Biology and Genetics (Oct 2023-Current)

Doctoral Researcher – Radboud University Medical Center (RadboudUMC), Netherlands (Oct 2018 – Dec 2022)

- Conducted long-read sequencing studies to uncover novel *de novo* mutations in undiagnosed patients.
- Developed and optimized analytical pipelines for structural variant detection and phasing.
- Collaborated with clinicians, data scientists, and sequencing technology providers.
- Co-authored multiple peer-reviewed publications on clinical genomics.

Bioinformatics Analyst – Seven Bridges Genomics, Turkey (Oct 2017 – Sep 2018)

- Developed pipelines for the Turkish Genome Project.
- Delivered workshops and training on cloud-based NGS workflows.
- Supported clients with custom data analysis pipelines.

Research Assistant – Michael Smith Genome Sciences Centre, BC Cancer Agency, Canada (Aug 2013 – Apr 2017)

- Contributed to assembly and annotation of the North American Bullfrog genome.
- Developed Kollector, a tool for transcript-guided targeted gene assembly.
- Conducted downstream analyses for the Personalized Oncogenomics (POG) initiative.

Internships

- Friedrich Miescher Laboratory, Germany (2012): Characterized Ndc1 protein in nuclear pore assembly.
- Harvard-MIT HST Program, USA (2011): Developed cryopreservation protocol for murine sperm.

Education

PhD in Human Genetics – Radboud University Nijmegen, Netherlands (2022)
MSc in Bioinformatics – University of British Columbia, Canada (2017)
BSc in Molecular Biology and Genetics – Bilkent University, Turkey (2013)

Technical Expertise

Programming & Tools

- R, Python, Bash, MATLAB, Java
- HPC & cloud-based environments (Seven Bridges, SLURM)
- Bioinformatics pipeline development (Snakemake, Nextflow)

Data Analysis

- Variant calling (short and long reads)
- Structural variation and de novo mutation detection
- Genome and transcriptome assembly (Kollector, Canu, Flye)

Domains

- Clinical and rare disease genomics
- Long-read sequencing (PacBio, ONT)
- Statistical genetics and phenotype-genotype integration

Publications

- Kucuk, E.*, van der Sanden, B.*, O'Gorman, L., Kwint, M., Derks, R., ... & Gilissen, C. (2023). Comprehensive *de novo* mutation discovery with HiFi long-read sequencing. Genome Medicine.
- Pauper, M.*, Kucuk, E.*, Wenger, A. M., Chakraborty, S., ... & Gilissen, C. (2021). Long-read trio sequencing of individuals with unsolved intellectual disability. Eur J Hum Genet, 29(4), 637–648.
- Kucuk, E., Chu, J., Vandervalk, B. P., ... & Birol, I. (2017). Kollector: transcript-informed, targeted *de novo* assembly of gene loci. Bioinformatics, 33(12), 1782–1788.
- Hammond, S. A., Warren, R. L., ... Kucuk, E., & Birol, I. (2017). The North American bullfrog draft genome provides insight into hormonal regulation of long noncoding RNA. Nat Commun, 8:1433.
- Yaldiz, B.*, Kucuk, E.*, Hampstead, J., Hofste, T., ... & Gilissen, C. (2023). Twist Exome Capture Allows for Lower Average Sequence Coverage in Clinical Exome Sequencing. Human Genomics.

Honors and Awards

High Honor Scholarship – Bilkent University
International Student Award – University of British Columbia
Internationalization Grant – Radboud University

Languages

- Turkish – Native
- English – C2
- Dutch – A2
- German – A2

PhD portfolio of Muhammet Erdi Küçük

Department: Human Genetics
PhD period: 10/10/2018 – 30/06/2024
PhD Supervisor(s): Prof. Christian Gilissen
PhD Co-supervisor(s): Dr. Juliet Hampstead

Training activities	Hours
Courses	
• RIMLS - Introduction course "In the lead of my PhD" (2019)	15.00
• Radboudumc - Scientific integrity (2023)	20.00
Seminars	
Conferences	
• ESHG Poster Presentation (2023)	20.00
Other	
Teaching activities	
Lecturing	
• Genetics Beginner R Course (2019)	15.00
Supervision	
• Bachelor Intern	30.00
Total	**100.00**

Radboudumc
university medical center
IN DEI NOMINE FELICITER
Radboud University