Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

DATA SCIENCE AND COMPUTATION

Ciclo XXXIII

Settore Concorsuale: 06/A1 – GENETICA MEDICA

Settore Scientifico Disciplinare: MED/03 – GENETICA MEDICA

Integrating structural variant calling, annotation and prioritization into whole genome analysis workflows: a practical application in the molecular diagnosis of neurodevelopmental disorders

Presentata da: Emanuela Iovino

Coordinatore Dottorato

Daniele Bonaccorsi

Supervisor

Prof. Marco Seri

Co-Supervisor

Dr Tommaso Pippucci

Esame finale anno 2023

Table of Contents

Abstract

Background: WGS is increasingly used as a first-line diagnostic test for patients with rare genetic diseases such as neurodevelopmental disorders (NDD). Clinical applications require a robust infrastructure to support processing, storage and analysis of WGS data. The identification and interpretation of SVs from WGS data also needs to be improved. Finally, there is a need for a prioritization system that enables downstream clinical analysis and facilitates data interpretation. Here, we present the results of a clinical application of WGS in a cohort of patients with NDD.

Methods: We developed highly portable workflows for processing WGS data, including alignment, quality control, and variant calling of SNVs and SVs. A benchmark analysis of state-of-the-art SV detection tools was performed to select the most accurate combination for SV calling. A gene-based prioritization system was also implemented to support variant interpretation.

Results: Using a benchmark analysis, we selected the most accurate combination of tools to improve SV detection from WGS data and build a dedicated pipeline. Our workflows were used to process WGS data from 77 NDD patient-parent families. The prioritization system supported downstream analysis and enabled molecular diagnosis in 32% of patients, 25% of which were SVs and suggested a potential diagnosis in 20% of patients, requiring further investigation to achieve diagnostic certain

Conclusion: Our data suggest that the integration of SNVs and SVs is a main factor that increases diagnostic yield by WGS and show that the adoption of a dedicated pipeline improves the process of variant detection and interpretation.

Introduction

1.1 Next Generation Sequencing

Next-generation sequencing (NGS) has led to an increase in understanding of the human genome and its relation to disease. From the beginning, sequencing technologies have shown potential, both in the clinical setting, by enabling molecular diagnosis of genetic diseases, and in research, by improving the understanding of genetic backgrounds in the general population and cohort studies. Increasing advances in sequencing technology and bioinformatics led to the introduction of whole genome sequencing (WGS), ushering in a new era of genomics in which every region of the genome can be explored. WGS is not only a powerful tool in research but is also becoming an integral part of clinical diagnostics and is contributing to the clinical management of various genetic diseases.

Promising results from the Rare Diseases Pilot study of the 100,000 Genomes Project support the clinical utility of WGS as a first-tier diagnostic test, particularly in patients with rare diseases ("100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report," 2021). However, several challenges in data processing, management, and interpretation of variants remain to be overcome to enable widespread application in clinical practice and to realize the full potential of WGS.

WGS provides several advantages compared with other sequencing technologies. Until recently, data complexity and high cost have been critical bottlenecks that have limited its widespread use in favor of target-enriched sequencing approaches, such as whole-exome sequencing (WES) and custom or disease-specific sequencing panels.

Disease-specific sequencing panels (target sequencing, TS) are powerful and economical tools that restrict the sequencing to gene sets associated with genetic disorders (Dillon et al., 2018). Genes of interest are enriched and sequenced simultaneously with an especially high read

depth of targeted regions enabling the detection of variants with very low non-reference allele frequencies caused by germline mosaics. While this approach has proven useful in the clinical diagnosis of well-established phenotypes whose genetic basis is well-described, it has limited to pre-selected genes risks overlooking crucial variants outside the target regions. Furthermore, technical limitations due to the enrichment approach result in discontinuous coverage preventing the detection of all types of genomic variants, such as copy number variants (CNVs), and the correct sequencing of GC-rich regions, such as the first exons. Therefore, a negative TS result may be inconclusive and require additional tests, such as WES, delaying the diagnosis.

WES is a target sequencing approach that captures the protein-coding regions of the genome, known as the exome, including approximately only about 1-2% of the entire genome (Caspar et al., 2018). Because an estimated 85% of disease-causing variants are found in exome regions, WES is widely used in the clinical setting as a first-tier test instead of TS, especially for several rare genetic disorders (Delaney et al., 2016). In addition, since WES is not limited to the analysis of genes previously associated with a particular disease, it provides a more comprehensive view of genomic variation and enables the identification of novel disease-gene associations in a research setting. WES has dramatically increased diagnostic yield in individuals with suspected genetic disorders, and yields generally range from 10% to 58%, depending on the genetic disorders, the clinical characteristics of the population tested, and analytical strategies (Lalonde et al., 2020). For example, an analytical approach based on family trio sequencing improves diagnostic yield because it can eliminate hundreds of non-causative variants and facilitates the detection of de novo or compound heterozygous variants in protein-coding regions (Alfares et al., 2020) (Tan et al., 2019). Although WES is a more comprehensive test than TS, WES and TS have several limitations inherent to target sequencing techniques, and they may miss different types and regions of disease-causing genomic mutations.

WGS, which overcomes many technical limitations inherent to target enrichment approaches, is the most suitable tool to comprehensively assess all types of genomic variations across the genome in an unbiased manner (Lionel et al., 2018). First, the absence of target enrichment probes allows unbiased genome-wide sequencing to reveal the protein-coding variants that are hidden due to uneven capture or non-inclusive target design (Lelieveld et al., 2015). WGS also

allows to go beyond the boundaries of the protein-coding regions and identify novel non-coding mutations and their association with disease (Perenthaler et al., 2019). Additionally, WGS allows comprehensive detection of all forms of variation, including single nucleotide variants, small insertion or deletion (indel) variants, structural variants, short tandem repeats and mitochondrial variants in a single assay. (Table 1)

	TS	WES	WGS
Read length(bp)	~ 300	~ 150	~ 150
Read Depth	200-1000x	~ 100 x	~ 30x
Error rate (%)	0.1	0.1	0.1
Advantage	High read depth, easy interpretation, cost efficiency	Additional sequence information compared to TS, cost efficiency	Uniform, complete access to all type of variants
Limitation	Incomplete coverage due to high GC content, missing enrichment probes, and regions with mappability<1, unprecise detection of CNVs and mitochondrial variants, no detection of SV	Incomplete coverage due to high GC content, missing enrichment probes, and regions with mappability <1, unprecise detection of CNVs and mitochondrial variants, no detection of SV	Imprecise on low-mappability regions

Table 1. Comparison of variant detection in different NGS approaches.

TS, WES and WGS are usually performed using sequencing-by-synthesis technology, which generates short reads (SR) of 25 to 300 base pairs (bp). Among the SR sequencing approaches, Illumina sequencing technology is the most widely used (Bentley et al., 2008). Briefly, the Illumina sequencing workflow consists of library preparation by random fragmentation of DNA, adapter ligation, and massively parallel sequencing of adapter ligated fragments (Figure a2). SR have an extremely low error rate per base pair, equivalent to 0.1% per nucleotide, making this method particularly suitable for the detection of SNVs or CNVs. The major drawback of SR sequencing is the read lengths, which cause alignment problems in particularly complex regions characterized by repeated sequences longer than the read length. Due to sequence homology, SRs can be mapped to multiple regions, resulting in ambiguous alignment that potentially leads to variant detection errors (Mandelker et al., 2016). As a result,

entire sections of our genome (more than 15%) cannot be resolved using SR and remain inaccessible and insufficiently explored (Logsdon et al., 2020).

Third-generation sequencing technologies. Third-generation sequencing technologies, also known as long-read sequencing (LR), overcome these limitations of SR sequencing. LR technology performs direct single DNA molecule sequencing in real-time and generates reads of several thousand bp with uniform coverage across the genome (Merker et al., 2018). Therefore, LR can access locations that are difficult to reach for SR and accurately map highly complex, repetitive, or homologous regions minimizing ambiguous alignments and reducing false negative and false positive calls.

Third generation sequencing technologies include two sequencing approaches: Pacific Biosciences (PacBio) and Oxford Nanopore Technology (ONT). PacBio technology relies on Single Molecule Real Time sequencing (SMRT) to generate highly accurate and tens of kilobases long reads. With SMRT sequencing, a DNA molecule is ligated to hairpin adaptors to generate a circular molecule known as a SMRTbell. Once the SMRTbell is developed, it is bound by a DNA polymerase for sequencing (Figure 2b).

The ONT technology differs from PacBio in that it uses linear DNA molecules instead of circular ones. ONT sequencing is based on nanopores through which a linear DNA molecule passes, causing a small electric current (Figure 2c). Consequently, due to the different structure of nucleotides and their charge, the measured current can be translated into corresponding bases on DNA (Lin et al., 2021). Among LR technologies, ONT sequencing generates the longest continuous reads. ONT reads are on average 7 to 8 kb long, but they can reach a length of 1 million bases. LR sequencing has been used in particular to improve genome assembly and enabled the completion of the ambitious Telomere-to-Telomere (T2T) project, the first complete, gapless sequence of a human genome (Nurk et al., 2022). Despite having the advantage of enabling access to regions inaccessible to SR, the clinical application of LR is severely limited by lower throughput, higher cost, and as yet standardized practices for downstream analysis.

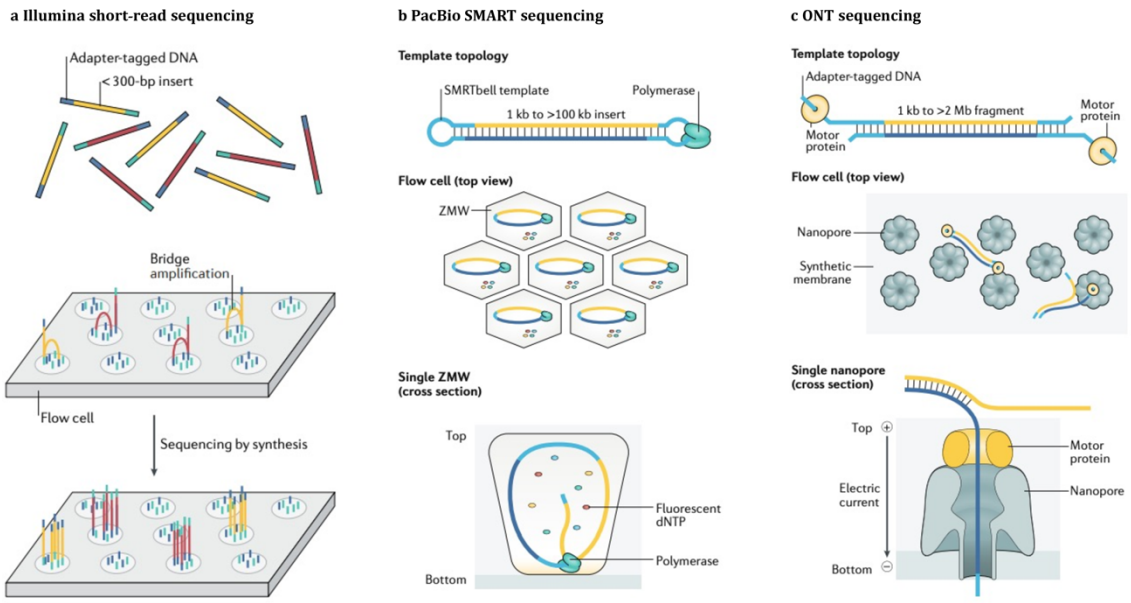


Figure 1. From (Nurk et al., 2022). Overview of short-read and long-read sequencing. a) Graphical representation of Illumina short-read sequencing: DNA fragments (yellow and red) are ligated with adapters (blue and aqua). The adapters contain two elements: unique molecular identifiers and sequences complementary to oligonucleotides bound to the surface of a flow cell. Initially, DNA was bound to the adapter and loaded into a flow cell to begin clustering. Thousands of copies of each fragment are generated through a bridge amplification process. Then, one strand refolds and the adapter at the other end binds to another oligonucleotide in the flow cell. A polymerase incorporates nucleotides to form the bridges of the double-stranded DNA molecules, which are then denatured to remain single-stranded. This process is repeated several times, resulting in several million double-stranded DNA. During sequencing by synthesis, fluorescently labeled deoxynucleoside triphosphates are incorporated into the newly synthesized DNA strand at each cycle. After incorporation, a laser excites the fluorophore on the strand and emits a characteristic fluorescent signal corresponding to each base. b) A graphical representation of PacBio: fragmented DNA (yellow for forward strand, dark blue for reverse strand) is ligated to hairpin adaptors (light blue) to form a topologically circular molecule called SMRTbell. The SMRTbell system is bound by a DNA polymerase and loaded into a cell for sequencing. Each cell contains millions of chambers called zero-mode waveguides (ZMWs), nanophotonic devices that confine light to a small observation volume. A single DNA molecule is immobilized on the bottom of the ZMWs. When the polymerase incorporates a nucleotide, light is emitted in a different color corresponding to each DNA base so that the incorporation of the nucleotide can be measured in real time. c) A graphical representation of ONT: The long DNA molecule is labeled with sequencing adapters (light blue) loaded into a motor protein at one or both ends. The DNA is combined with tethering proteins and loaded into the flow cell for sequencing. Thousands of nanopores are located in the flow cell, and the tethering proteins guide the DNA into proximity with these nanopores. The adapter is inserted into the nanopore opening, and the motor protein begins to unwind the double-stranded DNA.

The DNA is negatively charged, and when an electric current is applied, it moves through the pore, causing characteristic interruptions in the current.

1.2 Human Genetic Variation

A human genome typically differs from the reference genome at 4.1 to 5.0 million sites. However, most of the genetic variants identified in a typical human genome are common mutations shared by more than 0.5% of the population, and only 1 to 4 % of variants are rare (Auton et al., 2015). The vast majority of common and rare mutations are inherited from the parents, while a small proportion arise de novo from novel events in the parental gametes. *De novo* mutations (DNMs) are extremely rare events. The average DNM rate in the human genome is estimated to be about 1-1.3 × 10-8 mutations per base per generation, with considerable variation among families and classes of variation (Auton et al., 2015). This results in approximately 0.0154 CNV DNMs and 44 - 82 SNV DNMs, of which an average of 1.43 occur in the coding sequence. (Acuna-Hidalgo et al., 2016). DNMs are typically considered more deleterious than inherited variants because they are extremely rare events, are usually novel and not observed in the general population and have not yet been acted upon by natural selection (Veltman and Brunner, 2012).

De novo and inherited variants in humans occur in many forms and can range from mutations in single base pairs to changes in the structure of the DNA sequence. Every form of human variation is found throughout the genome, in both the protein-coding and non-coding sequences. The impact of mutations can range from benign to deleterious. In the broadest sense, genetic variants are typically divided into two distinct classes: single nucleotide variants (SNVs), which represent a qualitative class of mutations and involve the replacement of a single base, and the class of structural variants (SVs), which represent a quantitative class of mutations because they affect the dosage or copy number of genomic regions and their structure as the inversions. A schematic figure of genomic variants is shown in Figure 2.

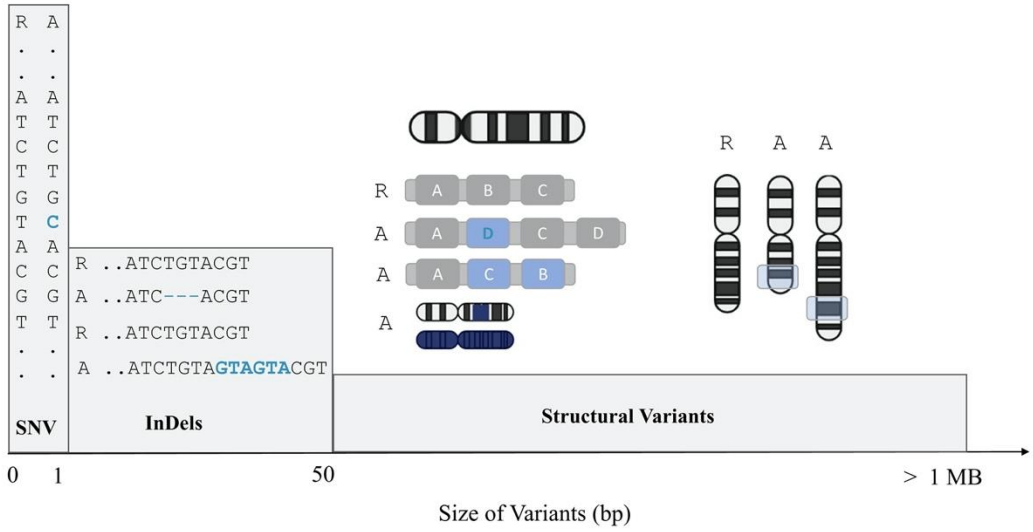


Figure 2. **The spectrum of human genetic variation**. The x-axis measures the number of nucleotides, from 1 bp to > 1 Mb. Above the axis, types of genetic variation are shown, with their size range and sequence compositions. SNV indicates a single substitution of one base in sequence; indels are small insertions or deletions in DNA sequence. Structural variants include changes in copy number or in the structure of DNA sequence.

SNVs constitute the most common genomic variation in the human genome and account for more than 90% of genomic variants. Most SNVs are located in the non-coding region and although they do not directly affect the biological function of a protein, they can influence gene regulation and splicing. SNVs found in the coding region are generally considered higher impact and more likely to be associated with disease because they may code for functional changes in amino acid structure (missense) or lead to premature protein truncation (nonsense). Also, the variants located at the canonical splice site within 2 bp of an exon-intron junction have larger effects and are known to be strong diagnostic candidates in loss-of-function disorders (Blakes et al., 2022). However, SNVs in coding regions can also cause a change that has no effect on the resulting protein sequence (synonymous) or on its function. Insertions and deletions (indels) of one or a few nucleotides (< 50 bp) in DNA sequence represent the second

most common type of variation in the genome. Indels occurring in coding regions can either leave the protein sequence unchanged or alter the reading frame of the transcript, thereby altering the protein sequence and leading to premature termination of the protein (Lin et al., 2017). Several technologies have been developed to study small variants such as SNVs and indels, from high-density SNP microarrays to NGS technologies. In the last two decades, advances in bioinformatics and the increasing use of WES and WGS have led to increasingly accurate detection of these small variants, allowing disease-causing genes or variants to be identified (Koboldt, 2020) (Zhao et al., 2020). In addition, major efforts in this area have led to the development of widely accepted best practices for small variant detection, which has greatly facilitated the conduct of sequencing studies to identify disease-causing variants (https://gatk.broadinstitute.org/hc/en-us).

SVs vary widely in size and type and can be balanced or unbalanced. Balanced SVs are changes in DNA structure, such as inversions (INV) of a genetic fragment or translocations (TRA) of DNA segments within or between chromosomes. Unbalanced SVs, also referred to as copy number variation (CNV), involve the gain or deletion of genetic material (Escaramís et al., 2015). SVs vary widely in size, ranging from 50 bp to well over megabases of sequence, and affect more nucleotides per genome than any other variant (Ho et al., 2020). Because of their size, they can have potentially significant phenotypic effects on gene expression or regulation compared to SNVs. SVs can affect gene expression in several ways. First, unbalance SVs can directly cause a gain by duplication (DUP) or a loss by deletion (DEL) of entire genes, resulting in a gene dosage effect by increasing or decreasing gene expression and the amount of encoded protein. SVs can also indirectly affect the regulatory architecture of the genome by altering the spatial relationship between regulatory elements and genes located even at great distances from the variant (Spielmann et al., 2018). Alternatively, SVs may also give rise to novel fusion genes involved in immune responses, particularly in cancer (Dubois et al., 2022). Since SVs are extremely diverse in type and size, they are much more difficult to identify and largely understudied, compared with SNVs. Over the past decades, numerous methods have been developed to identify different types of SVs, ranging from cytogenetic detection (e.g., karyotyping), array-based technologies (e.g., array and FISH), short-read WGS, and linked-read sequencing (e.g., 10X Genomics Chromium Technology) to long-read sequencing (e.g., PacBio and ONT). Continued advances in sequencing technologies have

steadily improved the detection of SVs and deepened our understanding of their role in disease etiology, regulation of gene expression and human diversity. For example, several neurological and neurodevelopmental disorders are associated with inherited and DNM SVs, such as Alzheimer's and Parkinson's disease, autism spectrum disorders (ASDs) and intellectual disability (ID) (Antaki et al., 2022) (D'haene and Vergult, 2021) (Lin et al., 2022). In addition, SVs are a key mutational process in various cancers, play an important role in autoimmune diseases, and recent studies suggest that they contribute to the susceptibility to COVID -19 or the progression of diseases (Sahajpal et al., 2022). Despite the great progress made in the discovery and characterization of SVs in the human genome, SV identification has not yet reached the reliability and quality of SNV calling. Due to the complexity of the genome, technical errors in sample preparation, and difficulties in identifying variants larger than reads, accurate and precise identification of SVs remains a challenge.

Lastly, another important class of genomic variations in the human genome often associated with SVs are short tandem repeats (STRs), also called microsatellites. STRs, like SVs, vary in size but are generally defined as a repeating motif 2-6 base pairs (bp) in size. STR Mutation rates are incredibly high compared to other variant classes, and it is estimated that each individual has about 100 de novo STRs and a human genome contains about 1.5 million STRs (Tankard et al., 2018). STRs are known to cause some Mendelian diseases known as repeat expansion disorders, such as Huntington's disease and hereditary ataxias, but there is growing evidence for a widespread role of common variations in STRs in complex traits such as gene expression. STRs can disrupt gene expression and cause aberrant protein folding or premature truncation, and currently about 30 diseases are associated with these variations. Diagnostic identification of STRs is challenging, and laboratory methods such as polymerase chain reaction (PCR) can be time-consuming and costly. NGS and especially WGS appear to be the most appropriate tool for characterizing STRs, and several tools are being developed to better understand these mutations in common traits and diseases (Rajan-Babu et al., 2021).

1.3 Structural Variants detection in short reads WGS data

Short-read-based WGS provides a unique opportunity for detecting SVs down to base pair resolution. While detection of SNVs/Indels has been standardized using "gold standard" tools universally used for sequencing and variant calling, such as BWA and GATK, best practices for identifying SV have yet to be defined. About 80 calling tools have been developed for identifying SV in short-read WGS, however there is no agreement on a single tool to use in variant calling also in clinical testing. The vast majority of these tools try to identify divergences between the reference genome and sample reads by examining the following features: read depth, paired-end, split and clip reads, and de novo assembly (Figure 3)(Escaramís et al., 2015).

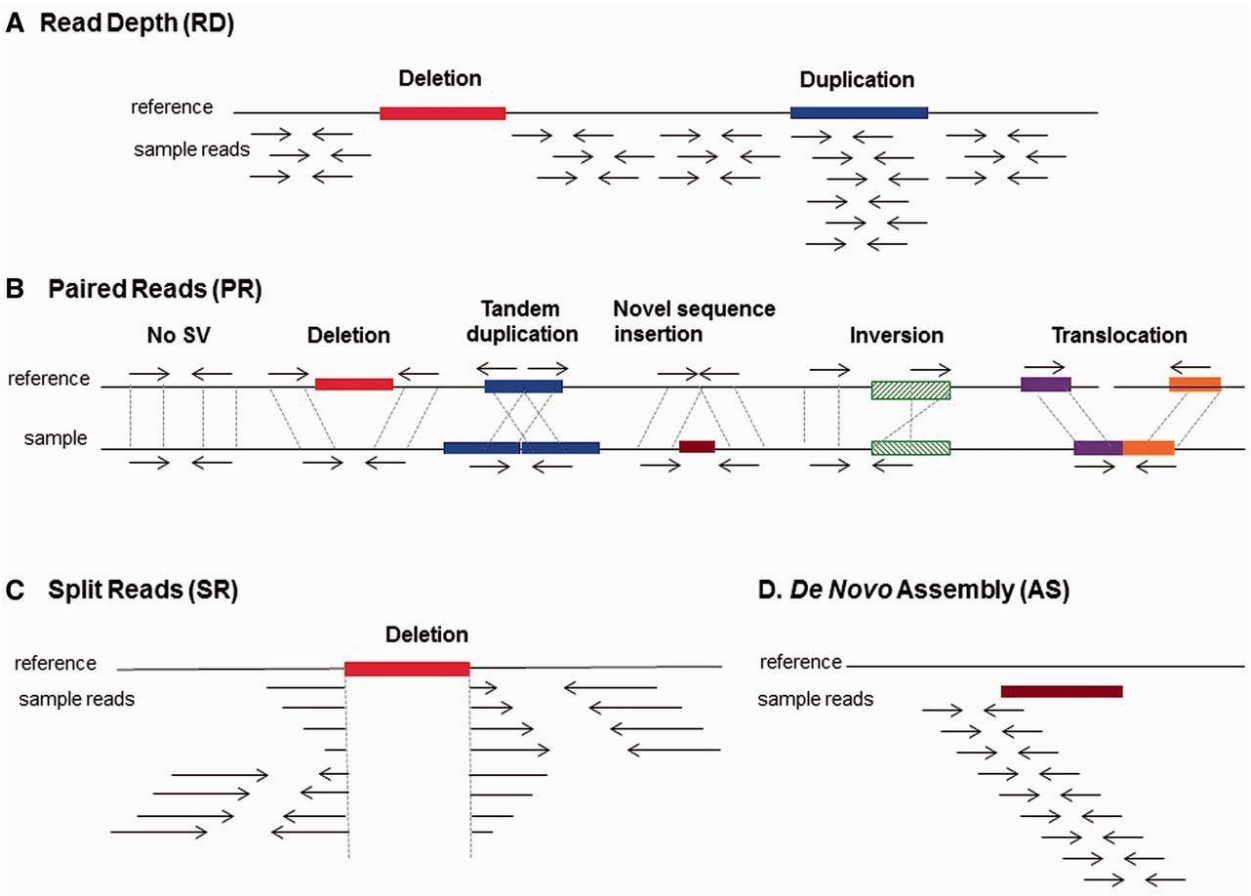


Figure 3 from (Escaramís et al., 2015). Strategies and signal for SVs detection.

The **read depth** indicates how many times a particular position of the genome has been sequenced. This feature is one of the most commonly used for identifying CNVs. For example, a deletion leads to a decrease in sequencing depth, while a duplication leads to an increase in sequencing depth. Although RD is indicative of SVs, variation in sequencing depth can also be due to artifacts or biases from PCR. Methods for detecting read depth rely on statistical approaches to account for heterogeneity in sequencing depth and to distinguish between artifacts and true variants. The choice of interval size, also known as sliding windows, to compare variations in sequencing depth within a given region is crucial, and this strategy detects large rather than small CNVs with greater accuracy.

The **paired-read** signal (PR) is based on the identification of clusters of divergent read pairs which indicate the presence of SV breaks between reads. One of the most common sequencing methods is paired-end sequencing, in which paired-end reads are generated by sequencing both ends of a fragment of a given size. Therefore, paired-end reads are expected to be mapped at a certain distance from each other, corresponding to the size of the insert, and in opposite orientations. However, the presence of SVs alters this type of expected signature and results in mismatched read-pair alignments, as shown in Figure 3B. These discordantly mapped paired reads may be a) further apart (or closer together) than expected based on the insertion size of the library, as in the case of a deletion of a DNA fragment, b) in reverse orientation, as in insertions/duplications, c) in the wrong order (pointing apart in the reference genome), as in an inversion, or d) mapped to different chromosomes, indicating a translocation. Breakpoint resolution in this approach depends on the mean and standard deviation of the library insertion size and coverage. The major drawback of this approach is that SVs are not the only signal that could perturb the discordant read mapping model. Artifacts, sequencing errors, and repetitive sequences can create discordant alignments that can fool the tools by mimicking a breakpoint and lead to the discovery of false positives.

The **split-read** signal is based on the identification of noncontiguous sequence reads. For example, a deletion in a sequenced sample causes the reads to be split to align to two non-

contiguous parts of the reference. The split-read method is a powerful method for detecting small and medium-sized variants and by definition, provides accurate resolution at the single nucleotide level, although the presence of small homology regions reduces accuracy to 1-10 nucleotides, while large homology regions result in even lower accuracy. In addition, the split-read method is limited for large variants or those in repetitive regions due to its local mapping approach

***De novo* sequence assembly** (AS), traditionally used to generate reference genomes, also allows detection of SV. AS of all reads of the genome is very expensive and requires significant sequencing depth compared to mapping-based methods. However, some variants, such as insertions, are difficult to detect with mapping. One possible strategy is to perform local assembly from a subset of reads. Moreover, AS can also be used to refine the breakpoints of complex SVs as well as all types of SVs.

Most available tools, especially those developed recently, rely on a combination of several approaches described above to increase the sensitivity and accuracy of detecting SV. The reason that so many tools (over 80) have been developed is that none of them has yet been widely used throughout the scientific community, and many efforts are still needed to improve call accuracy, increase sensitivity and usability, and shorten computation times (Liu et al., 2022). However, to avoid confusion among researchers who have a variety of tools to choose from, benchmark studies have been conducted in recent years to evaluate the relative advantages and disadvantages and to suggest best practices for selecting SV algorithms.

Published benchmark studies based on real data have found that no single SV caller algorithm can call all types of SVs with high precision and recall (Cameron et al., 2019) (Kosugi et al., 2019).

For example, very large changes of several Mb are often better detected by tools that rely on RD approaches. Using consensus calls generated by multiple SV callers may be a solution to achieve better precision and recall rate for a large number of SVs, but there is currently no clarity on how multiple tools should be used together. While creating a union of all SV calls made by multiple callers may increase sensitivity at the expense of precision, a more restrictive approach, such as requiring that a SV be named consensually by three or more SV callers, often results in low sensitivity but high precision (Becker et al., 2018). In addition, Kosugi et al. found that caller performance diverges widely between simulated data, which is commonly

used in the testing phase of a new tool, and real data from a truth set. Simulated data have low commutability and often underrepresent the complexity of SVs (as well as the genomic regions in which they occur). A balance between precision and sensitivity could be achieved by increasing the number and diversity of SV benchmark datasets that come from real data rather than simulated data, and by eliminating confounding factors that contribute to systematic miscalls. However, it is difficult to generate true high-confidence datasets, and they are often limited to a subset of SVs, particularly deletions. Due to the efforts of the Genome in a Bottle Consortium (GIAB) (Zook et al., 2020), a benchmark set of SV calls has been published for the Ashkenazi Jewish trio's son, HG002. The HG002 truth set contains only deletions and insertions identified by consensus calls from different platforms and multiple callers. In contrast, information on duplications, inversions, and translocations remains limited, considering that short-read data often report access to up to 4000 translocations, of which ~50-70% are repeat expansions (Sedlazeck et al., 2018). In addition, the HG002 truth set was obtained by aligning sequencing reads with an older version of the human reference genome GRCh37. However, a comprehensive assembly-based whole genome benchmark with the most used version of the reference genome, such as GRCh38, is not yet available and the available truth set has lower quality compared to SNV gold standard truth sets. Thus, it remains to be clarified to what extent the low performance in the context of real-world data is due to the low reliability of the callers and also to the shortcomings of the only high-quality truth set currently available. Another important issue is the use of genotypers after the SV calling step. Several tools have been developed for genotyping, such as SV2, SVTyper, GraphTyper and others (Antaki et al., 2018) (Eggertsson et al., 2017). Although genotyping performance can also be affected by size and type, recent work shows that these tools generally perform better than callers (Chander et al., 2019).

1.4 Toward non-coding genome exploration

The contribution of noncoding variation to common diseases and traits has long been studied, whereas for rare diseases most genetic analyses have sought the cause in protein coding regions of the genome. However, this approach has only been able to diagnose about 30-40% of rare genetic disorders. The reasons for this are many, but likely lie in unexplored SNVs or SVs in the non-coding genome (Krude et al., 2021). In recent years, it has been shown that a large part of the non-coding genome is functional and contains genetic variants that contribute to disease development. Great progress has been made in defining the non-coding elements in the genome; a schematic representation is shown in Figure 4. The 5' and 3' untranslated regions (UTRs) of mRNAs and introns account for up to ~35% of the human genome, and transposable elements and tandem repeats account for another 50%. Regulatory elements such as promoters, silencers, and enhancers tightly control gene protein expression by binding transcription factors (TFs). In addition, thousands of non-coding RNAs (ncRNAs), including short ncRNAs (b200 nucleotides) and long ncRNAs (lncRNAs) (N200 nucleotides), are transcribed from the non-coding genome. Most interactions between regulatory elements and exons are usually physically compartmentalized into topologically associating domains (TADs). TADs are key units of the three-dimensional (3D) nuclear architecture in the human genome. TADs are typically < 1 Mb in size and delineate the regions of chromosomes where sequences preferentially interact with each other rather than with elements in other regions of the genome. They may contain a portion of a gene or multiple genes more likely to be coregulated than genes not in the same TADs (French and Edwards, 2020).

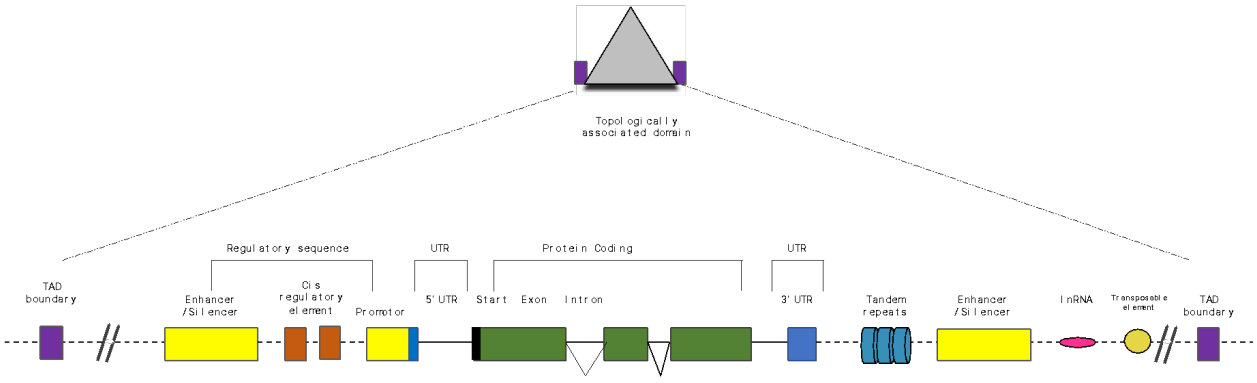


Figure 4. Schematic representation of the functional elements in noncoding DNA. Chromosomes are divided into topologically associated domains (TADs) corresponding to regions of highly interacting chromatin that appear as pyramidal structures. Within the boundaries of TADs, regulatory elements (yellow rectangles) such as promoters, enhancers, and silencer elements can form chromatin loops mediated by TF. Transcriptional elements are exons (green rectangles) and noncoding RNA genes (lncRNA; magenta ellipse). The UTRs flank the beginning and end of protein-coding regions and play a critical role in post-transcriptional gene regulation processes. Transposable elements (golden ovals) and tandem repeats (blue ovals) are widely used repetitive DNA sequences.

The most popular method for measuring interactions between chromosomal regions and providing a thorough view of 3D organization is high-throughput chromosome conformation capture (Hi-C) (Lieberman-Aiden et al; 2009). Each coding and noncoding variant within the TAD-boundary system has the potential to influence gene misexpression and, in some cases, disease through the repositioning of regulatory boundaries and/or the relocation of enhancer elements into different regulatory environments (Melo et al., 2020).

The role of noncoding variants in rare diseases is becoming increasingly clear. For example, a study using the Deciphering Developmental Disorders (DDD) dataset identified CNVs and SNVs in non-coding regions of MEF2C that cause severe DD due to a loss-of-function mechanism (Wright et al., 2021). Mutations in noncoding regions of GJB1 and ABCA4 are likely associated with X-linked Charcot-Marie-Tooth disease (Tomaselli et al., 2017). CNVs on enhancers at the IHH locus are associated with syndactyly and craniosynostosis (Will et al., 2017). However, because current means of interpreting noncoding mutations are severely limited, all pathogenic candidate variants in noncoding regions, as in these cited studies, must generally be proven causative by functional studies, band-shift assays, reporter gene assays, ChIP-seq experiments, and animal models. Current classification algorithms for predicting the impact of variants in the noncoding genome are not yet robust and often classify these variants as variants of uncertain significance (VUS). The fact that only 0.18% of pathogenic or possibly pathogenic variants reported in the ClinVar database are found in noncoding regions suggests that there are no general rules for interpreting variants that lie outside coding regions. In 2022, Ellingford et al published a general guideline to adapt the American College of Medical Genetics (ACMG) recommendation to noncoding variants and establish rules that should be evaluated to interpret them. Moreover, the prioritization of functional variants is based on functional annotations that are incomplete and uninformative. Therefore, the widespread use

of WGS can help both to generate data useful for training classification models and to develop computational models capable of predicting their effects. To improve knowledge of this class of variants, the research community today proposes to test and share the non-coding variants found by submitting them to ClinVar or DECIPHER (Landrum et al., 2018) (Bragin et al., 2014).

In the future, integration of large-scale projects such as ENCODE, FANTOM and individual studies should drive clinical interpretation and lead to codification of their role in human phenotype and rare diseases (Hon et al., 2017)(Davis et al., 2018). Since non-coding mutations were largely ignored for many years, it is likely that some functional elements have not yet been discovered.

1.5 Neurodevelopmental disorder and WGS application

Rare genetic diseases (RGD) are individually rare but collectively common, affecting the lives of approximately 25 million people in Europe alone. Neurodevelopmental disorders account for a large proportion of these rare genetic diseases and continue to place an enormous financial, logistical, and emotional burden on families, society, and the healthcare system. Neurodevelopmental disorders (NDD) are a group of heterogeneous conditions that affect the development of the central nervous system and are characterized by impairments in cognition, memory, language, and motor skills. NDDs include autism spectrum disorders (ASD), intellectual disability (ID), developmental disability (DD), attention deficit/hyperactivity disorder (ADHD), specific learning disorders (in reading, written expression, and/or mathematics), and motor disorders, and more broadly, disorders such as epilepsy and schizophrenia (Savatt and Myers, 2021). Given the considerable genetic heterogeneity, presence of comorbidities, and ascertainment bias, making a diagnosis is challenging. Discovering the molecular etiology of NDD can take years, and some people do not receive a diagnosis at all. This diagnostic odyssey is often associated with unnecessary tests and procedures, incorrect diagnoses, delays in effective management of disease progression, and emotional frustration and uncertainty in further reproductive decisions by parents (Schuermans et al., 2022). The vast majority of NDDs are monogenic (Mendelian) disorders resulting from a mutation in a gene inherited via an autosomal dominant, recessive, or X-linked inheritance, or the mutation may occur de novo. The causative variant may be found in the exome or in non-protein coding regions, including splice sites, exon-intron boundaries, introns, mitochondrial DNA, and regulatory regions. Current genetic tests for NDD include single gene testing, multi-gene panel testing, karyotype, microarray (CMA) and WES (Table 2). NGS technology has dramatically improved the diagnostic rate of NDDs and has also promoted the discovery of new disease genes over the past decade. In particular, WES has led to tremendous progress in deciphering monogenic forms of NDD with more than 900 genes having been reported, leading to it becoming a first-line diagnostic test (Parenti et al., 2020). At the same time, several public reference population databases have been developed to understand the role of genetic variation and to classify SNVs and SVs into common or potentially disease-causing ones, e.g., dbSNP (https://www.ncbi.nlm.nih.gov/snp/), gnomAD

(https://gnomad.broadinstitute.org/), gnomADSV, 1000 Genomes (https://www.internationalgenome.org/), Decipher (https://www.deciphergenomics.org/), and the 100,000 Genomes Project (https://www.genomicsengland.co.uk/initiatives/100000-genomes-project). Despite this great progress in understanding the disease, a genetic diagnosis is made in only about one-in-three patients (Srivastava et al., 2019). Methods to address this lack of heritability, particularly WGS, as well as transcriptomics and LR, are increasingly being incorporated into diagnostics, with a focus on the importance of regulatory non-coding sequences and SVs that are not detected by WES. WGS is the only test that can detect almost all types of genetic variants (Table 2), and its introduction into clinical practice could reduce the number of genetic tests that patients need performed. Still, the application of WGS in clinical practice means identifying a class of genomic variants whose functional implications are not well-understood and require additional experimentation for confirmation. Moreover, a solid infrastructure for data processing and interpretation is required, considering that WGS identifies approximately 3,4 million small variants instead of 50,000 variants in WES. In this context, the boundaries between diagnostics and research are becoming increasingly blurred, and close collaboration between clinicians, geneticists, molecular biologists and bioinformaticians is required.

Current Testing Options	SNVs/ InDels	CNVs	SVs	STRs	Mitochondrial	Number of loci/regions evaluated
Single Gene (Sanger)	yes	No	No	No	Yes	Average ~ 27,000
Gene Panel	yes	Limited	No	No	Yes	Related to genes number
Karyotype	No	No	Yes	No	No	~ 500
CMA	No	Yes	No	No	No	~ 0.05 - 2 million
WES	No	Limited	Limited	Limited	Yes	5 million
WGS	Yes	Yes	Yes	Yes	Yes	3 billion

Table 2. Current genetic test option for Rare Disease.

1.6 Computational challenge for WGS

Due to widespread sequencing data, computer science and genomics are becoming increasingly intertwined. The volume and complexity of high-throughput sequencing data require specialized computational methods for data storage, processing, and analysis. The clinical application of NGS, in particular, has been driven by technological innovations in both sequencing technology and genomic data processing. With the goal of bringing WGS into clinical practice, new computational challenges arise. In contrast to the ~50,000 variants found in a WES experiment, WGS can identify about 3 million variants, a difference of 70 times, and generates a volume of data 12 times larger than the volume of exome data. (Krude et al., 2021). In addition, genomics data will surpass the current major generators of "big data" such as YouTube (~1-2 exabytes/year) and Twitter (~1.36 petabytes/year) (Stephens et al., 2015). Therefore, new paradigms for long-term storage, computational resources, and processing of data are needed. The main challenges remain processing time, variant interpretation, and data management. In recent years, cloud computing and data storage have lowered the cost of maintaining servers and allowed for more efficient, scalable workflows. However, as the cost of cloud instances and long-term storage remains large, standalone servers are a more viable solution (Panda et al., 2021). Nevertheless, a significant infrastructure effort is required to facilitate the widespread use of WGS data in the clinical setting.

The processing of WGS data involves multiple steps, starting with alignment and variant calling, which are implemented primarily by open-source tools. The computing infrastructure required by these tools for WGS data includes a multi-core system for parallelizing computational operations and scalable, tightly controlled memory. Workflow management systems (WfMS) have recently obtained popularity in the field of genomics because of their ability to automate the execution of computational tasks and control the allocation of resources during the various steps of data processing (Ahmed et al., 2021). Several WfMSs have been developed in genomics, such as Nextflow, CWL, Snakemake, and WDL (Di Tommaso et al., 2017). These systems, therefore, grant users the ability to write reliable, scalable and reproducible pipelines in a context that manages dependencies, allocates and recovers memory, and keeps track of tasks and errors.

Recently, new approaches based on the computing power of the Graphics Processing Unit (GPU) architecture have been developed to reduce the computational time required to process WGS data. Systems such as Illumina DRAGEN and NVIDIA Parabricks allow for tremendous speedup in WGS data processing compared to the standard pipeline CPU. DRAGEN is a field-programmable gate array technology (FPGA) system that accelerates WGS data processing by 30 times compared to a standard CPU, while NVIDIA Parabricks results in 48 times end-to-end acceleration (Ham et al., 2020). Benchmarking studies are quite difficult because DRAGEN does not specify all the tools involved in the pipeline. However, preliminary studies seem to be moving in the direction that integrating GPUs into genomics reduces computation time and increases throughput (Carpi et al., 2022).

Aim

WGS is emerging as a promising diagnostic test for patients with NDD because it provides more uniform sequence coverage than WES in coding regions and also allows identification of regulatory and structural variants within non-coding regions. Although several studies have confirmed the usefulness of WGS in the clinical setting, standards for the practical implementation of clinical WGS analysis are lacking. The main unresolved issues are:

- ⇒ lack of a common pipeline for structural variant detection
- ⇒ clinical application requires a solid infrastructure to support data processing and handling
- ⇒ a lack of a prioritization approach that drives clinical downstream analysis of a large number of variants (e.g., SNVs and SVs).

There are several tools developed for SV detection on SR WGS data, but no single caller has been widely used in clinical diagnostics. In addition, the performance of each tool differs, depending on SV type and size range. Therefore, using calls generated by multiple SV tools may achieve better precision and recall rate. Building a pipeline for clinical evaluation of SVs requires:

1) an assessment of the types and sizes of SVs that can be most reliably identified from SR WGS data (to achieve satisfactory precision)

2) an assessment of how multiple tools should be used together (to achieve a satisfactory recall).

Here we evaluated which SV types can be most reliably called using several state-of-the-art tools on three different real data sets: HG002 sample (the son of the Ashkenazi trio from the GIAB consortium), NA12878 sample (from the 1KG project), and REACH00236 sample (an internal gold standard crossed from Pacific Bioscience and Oxford Nanopore Technologies). We also compared a consensus approach, using only the variants from at least two callers to a union approach, using all SVs calls from multiple callers.

Although WGS is the most appropriate tool to detect all relevant human genetic variants in a single test, its clinical application is still in its infancy. Sequencing thousands of genomes for clinical testing requires specialized infrastructure and expertise in the healthcare system. One of the most useful approaches is to develop scalable and reproducible pipelines using workflow management systems. Here, pipelines are developed for detection and prioritization of SNVs and SVs in a clinical cohort of 25 NDD trios using WfMS. In addition, the DRAGEN Illumina system pipeline was tested to compare the time required, flexibility of data, and future application of this system for clinical investigations. Human genetic variants are extremely diverse, ranging from small variants affecting single base pairs to large structural variants affecting thousands or millions of nucleotides. Widespread application of WGS in clinical cohorts will enable detection of more pathogenic coding and noncoding SNVs and SVs and allow for a better understanding of the combined effects of different variants. Currently, different algorithms and separate pipelines are being developed for each class of variant. Here, we propose a prioritization method for clinical evaluation of SNVs and SVs. In addition, the diagnostic rate and diagnostic outcome of using WGS for the NDD cohort is reported.

Methods

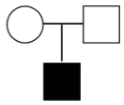
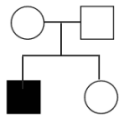
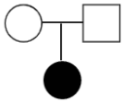
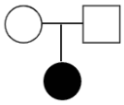
3.1 Patient recruitment and family collection

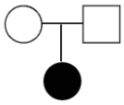
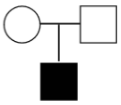
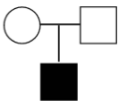
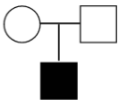
A total of 21 trios, three quartets and one multigenerational family in which one or multiple members were clinically diagnosed with a neurodevelopmental delay (NDD) suspected to be of genetic origin were enrolled in the Department of Medical Genetics of the Sant'Orsola-Malpighi University Hospital in Bologna. The inclusion criteria were clinical diagnosis of NDD and previous inconclusive genetic testing. Notably, these families had previously undergone several genetic testings, including sequencing/MLPA of clinically indicated genes, aCGH or aSNP analysis to detect CNVs, and in the vast majority of cases WES. This study was conducted as part of the project RF-2018-12366314 entitled "Whole Genome Sequencing into the diagnostic workflow of rare diseases: a cost-effectiveness evaluation in a heterogeneous population of patients with inconclusive Whole Exome Sequencing" and supported by the Italian Ministry of Health. All participants received information about the analysis and signed an informed consent form. In the case of minor patients or patients who could not express their will, the parents or legal representatives signed the consent. All family members were Italian (thus presumably belonging to the so-defined group of non-Finnish Europeans in gnomAD), and there were no self-reported consanguineous matings. Some participants were included in the project NIG (Network for Italian Genomes, 32%), a project that promotes the creation of a reference genome for the Italian population and the frequency of genomic variants in the Italian population. For all patients, phenotypic information was collected from medical records and captured by the standardized Human Phenotypic Ontology (HPO).

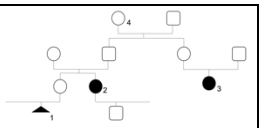
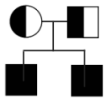
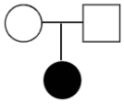
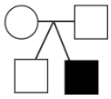
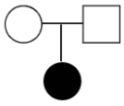
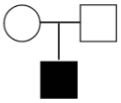
Neurodevelopmental profile. Intellectual Disability (ID) was present in all patients (54% male and 48% female), with varying severity and most of them had at least one additional

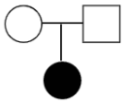
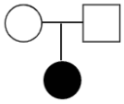
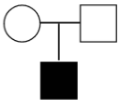
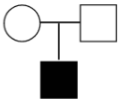
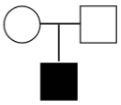
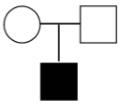
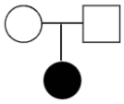
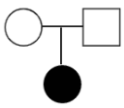
feature. The most common comorbid features were neurological conditions: epilepsy (64%), brain abnormalities (24%), dysmorphisms (20%), and hypotonia (12%).

Clinical history of the included families (FID) is briefly described in the table below (Table 3).

FID	Clinical features	Pedigree
FID.1	**NDD features**: Delayed psychomotor development, severe ID, absent speech and self-injurious behavior. **Other clinical features**: drooling, swallowing and feeding problems; microcephaly, brain abnormalities: opercular polymicrogyria. **Family history**: sporadic **Negative genetic tests**: karyotype, investigation of sub telomeric rearrangements, FISH to explore 22q11.2 deletion, array CGH (average resolution 100kb), NGS panel of 182 genes associated with brain malformations, clinical exome (performed in singleton mode, performed in 2015).	
FID.2	**NDD features**: ID, language impairment **Other clinical features**: small for gestational age and postnatal growth failure; hypertrophic cardiomyopathy; microcephaly, brain abnormalities: hypoplasia of the corpus callosum and cerebellar vermis hypoplasia, myelination delay. Muscle biopsy revealed numerous ragged-red fibres and mitochondrial abnormalities on ultrastructural examination. **Family history**: Agenesis of the corpus callosum was identified in his sister. **Negative genetic tests**: Array CGH (average resolution 100 kb), TS, and WES in trio mode.	
FID.3	**NDD features**: Mild ID and language impairment **Other clinical features**: fine and gross motor skills reduced, dilatation of renal calices and overweight. **Family history**: sporadic **Negative genetic tests**: WES in trio mode	
FID.4	**NDD features**: Global developmental delay with severe ID, absent speech **Other clinical features**: severe hypotonia from neonatal age, myoclonic seizures, gastrointestinal problem; plagiocephaly, broad thumbs in feet,	

	brain abnormality: corpus callosum hypoplasia; facial dysmorphisms; skeletal abnormalities **Family history**: sporadic **Negative genetic tests**: Array CGH, TS of genes involved in epileptic encephalopathy, WES in trio mode	
FID.5	**NDD features**: severe ID **Other clinical features**: non-cyanotic heart disease, bilateral coloboma, paresis of facial nerve (VII cranial nerve), hypoacusis, monoliteral hip dysplasia, imperforate hymen, square face, cleft palate; brain abnormality: hypoplastic cerebellar vermis, secondary amenorrhea, facial dysmorphisms **Family history**: brother with psychomotor delay **Negative genetic tests**: karyotype, Array CGH, analysis of CHD7, WES in trio mode	
FID.6	**NDD features**: Developmental delay with severe ID, absent speech **Other clinical features**: intrauterine growth restriction and oligohydramnios, hypotonia, failure to thrive, growth failure, growth hormone deficiency, facial dysmorphisms **Family history**: recessive due to uniparental disomy (UPD) of chromosome 7 **Negative genetic tests**: Array CGH, fragile X testing, WES in trio mode	
FID.7	**NDD features**: Delayed psychomotor development with severe ID **Other clinical features**: axial hypotonia, mild dysmorphic facial features, exotropia, brain abnormality: deficit of IV facial nerve, enlargement of subarachnoid spaces of bilateral fronto-temporal convexity **Family history**: sporadic **Negative genetic tests**: Array CGH, clinical WES (single mode)	
FID.8	**NDD features**: mild ID **Other clinical features**: laryngomalacia, recurrent respiratory infections, obesity, short limb dwarfism, strabismus, hypoacusia, hypotonia, facial dysmorphisms **Family history**: sporadic **Negative genetic tests**: Array CGH, Methylation analysis, clinical WES (single mode)	

FID.9	**NDD features**: Not Applicable (NA) **Other clinical features**: Male fetus (1) with absent tibia, holoprosencephaly, polydactyly, syndactyly, dysmorphism, ectopic kidney **Family history**: aunt (mother sister,2) with absent tibia and polydactyly; mother's cousin with holoprosencephaly (3). **Negative genetic tests**: custom micro-array and Sanger sequencing of ZRS	
FID.10	**NDD features**: Neurodevelopmental delay with severe ID **Other clinical features**: Seizure, microcephaly, brain abnormality: thin corpus callosum and abnormal myelination; astigmatism; alternative exotropia; ERG anomalies **Family history**: brother affected **Negative genetic tests**: karyotype, investigation of sub telomeric rearrangements, analysis of ATRX, WES (single mode)	
FID.11	**NDD features**: Mild ID **Other clinical features**: poor fine coordination, paroxysmal motor disorders of sleep, mild facial dysmorphisms **Family history**: sporadic **Negative genetic tests**: karyotype revealed a paracentric inversion on Chrom X	
FID.12	**NDD features**: ID mild, pervasive developmental disorder with atypical autistic spectrum disorder **Other clinical features**: facial dysmorphisms **Family history**: unaffected twin **Negative genetic tests**: WES in single mode	
FID.13	**NDD features**: not performed due to sudden infant death **Other clinical features**: two episodes of central apnea, one of which was fatal **Family history**: sporadic **Negative genetic tests**: WES in single mode	
FID.14	**NDD features**: Delayed psychomotor development with severe ID **Other clinical features**: Epileptic encephalopathy, mild dysmorphic facial features **Family history**: X-linked, the unaffected mother had two brothers, a stillborn twin and a brother who died at the age of 1 month. **Negative genetic tests**: Array CGH, WES in single mode	

FID.15	**NDD features**: Severe ID, however perinatal asphyxia was reported **Other clinical features**: Epileptic encephalopathy, hemiparesis **Family history**: sporadic **Negative genetic tests**: Array CGH, WES in single mode	
FID.16	**NDD features**: Delayed psychomotor development and mild ID **Other clinical features**: drug-resistant epilepsy, facial dysmorphism **Family history**: sporadic **Negative genetic tests**: Array CGH, WES in single mode	
FID.17	**NDD features**: severe ID with absent speech **Other clinical features**: Epileptic encephalopathy **Family history**: sporadic **Negative genetic tests**: Array CGH, WES in single mode	
FID.18	**NDD features**: Delayed psychomotor development and severe ID **Other clinical features**: Epileptic encephalopathy, microcephaly, polymicrogelia, cortical malformation **Family history**: family cases reported for cortical malformation and ID **Negative genetic tests**: Array CGH, WES in single mode	
FID.19	**NDD features**: Delayed psychomotor development and ID **Other clinical features**: Epileptic encephalopathy, hearing loss due to recessive variant in GJB2 (c.35del, p.Gly12Valfs*2), brain abnormality : Cerebral atrophy **Family history**: sporadic **Negative genetic tests**: Array CGH, WES in single mode	
FID.20	**NDD features**: Mild ID and behavioral disorders with attention deficit-hyperactivity disorder and oppositional defiant disorder **Other clinical features**: Epileptic encephalopathy, facial dysmorphism **Family history**: sporadic **Negative genetic tests**: karyotype, fragile X testing, Array CGH, WES in single mode	
FID.21	**NDD features**: ID not specified **Other clinical features**: Epileptic encephalopathy and behavioral disorders **Family history**: sister with seizures **Negative genetic tests**: Array CGH, WES in single mode	
FID.22	**NDD features**: Developmental delay **Other clinical features**: Epileptic encephalopathy, Microcephaly **Family history**: sporadic	

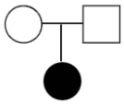
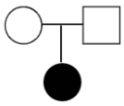
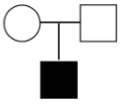
	Negative genetic tests: Array CGH, WES in single mode	
FID.23	**NDD features**: Severe ID **Other clinical features**: drug-resistant epilepsy, autistic features, recurrent urinary tract infection **Family history**: sporadic **Negative genetic tests**: WES in single mode	
FID.24	**NDD features**: Delayed psychomotor development, severe ID **Other clinical features**: seizures, deficit of the superior oblique muscle with exotropia, obesity, congenital hip dysplasia, brain abnormality: cyst of the third ventricle (surgically treated) **Family history**: brother with mild ID **Negative genetic tests**: WES in trio mode	
FID.25	**NDD features**: Delayed psychomotor development and ID **Other clinical features**: drug-resistant epilepsy, Hypotonia, mild dysmorphisms **Family history**: maternal cousin with schizophrenia **Negative genetic tests**: Array CGH, TS for gene involved in epileptic encephalopathy, TS for Pitt-Hopkins, WES in trio mode	

Table 3. Clinical features in NDD. The first column FID contains the ID of the family. The second column summarizes the clinical features obtained from the medical records. The third column indicates the family members who participated in the WGS study; additional unaffected family members (i.e siblings) are not reported in pedigree.

3.2 DNA sequencing

WGS was performed as described by the manufacturer (Illumina, San Diego, CA, USA) at IIT laboratories (Genoa, Italy). Briefly, 200 ng of genomic DNA (gDNA) from the subjects was processed using the Illumina TruSeq Nano DNA Library Prep Kit following an optimized gDNA shearing process. Multiple indexed patient libraries were sequenced simultaneously on a single S4 flow cell using the NovaSeqTM 6000 system (Illumina Inc.), which can sequence

up to 48 genomes/run. Patient DNA libraries were sequenced at an average of 30X coverage using 150x2 bp paired-end chemistry.

3.3 SNVs/InDels detection, genotyping and annotation

The clinical application of WGS involves developing bioinformatic pipelines that enable accurate and efficient germline variant calling. The Genome Analysis Toolkit (GATK), developed by the Broad Institute, is an open-source genome analysis package that contains all the tools for calling germline small variants (SNVs/InDels) (https://gatk.broadinstitute.org/hc/en-us). GATK is widely used in the genomics community for detecting small germline variants and applies a variety of state-of-the-art statistical methods (e.g., logistic regression, hidden Markov chain, and Naïve Bayes classification) to accurately identify differences between reads and the reference genome caused by either true genetic variants or errors. However, without any optimization, most GATK4 tools require more than 60 hours to complete a 30X WGS pipeline analysis. Following the GATK4 best practice workflow, a pipeline for small variants was implemented using the Snakemake workflow managers (*sn-pipeline*) (Koster and Rahmann, 2012). The pipeline consists of the following steps: quality control, read alignment, variant calling, and annotation (Figure 4).

Snakemake is a scalable workflow management system that uses a language similar to the standard Python syntax. It divides the entire workflow into rules, where each input is the output of the rule corresponding to the previous step. A system based on multiple steps makes the complex WGS processing flow easy to follow. Wildcards tags allow parameters to be passed to any rule in the pipeline by deriving the value of the parameter from the target filename. In addition, Snakemake recognizes which rules are independent of each other and can be executed in parallel to reduce idle time CPU and speed up the workflow. All required tools are automatically set up in isolated virtual environments using Conda package repositories.

Read alignment: the raw reads were aligned to the human reference genome GRCh38, which contains the primary assembly as well as ALT contigs, additional decoy contigs and HLA genes. Alignment is performed using BWA-MEM and the -M option, which turns the split reads into secondary alignments (Li and Durbin, 2009). The BWA output is piped and sorted

directly into an output BAM file using SAMtools (Li et al., 2009). In addition, SAMtools was preferred over Picard for the detection of duplicated reads because it speeds up the runtime process. For the GATK Base Quality Score Recalibration step, the genome was divided into 20 interval fractions in order to parallelize the step (one interval per thread). Sequence intervals were determined by running gatk ScatterIntervalsByNs with the GRCh38 reference genome, and the interval list is divided by the number of threads available on a virtual machine used to set the pipeline parameters. The recalibrated BAMs have a size of approximately 95 GB per sample.

Variant calling: Variant calling of SNVs and indels was performed with the GATK HaplotypeCaller using 20 intervals. Output is in GVCF mode and stored in a GenomicsDB file to improve scalability and accelerate the next step: joint-genotyping of multiple samples. A GenomicsDB import was performed for each of the 20 intervals to speed up the process. GenotypeGVCFs checks the available information for each locus for both variant and non-variant alleles across all samples and creates a VCF file containing only the sites found to be variant in at least one sample. The VCF files for each interval were merged into a compressed file using GatherVcfs.

Annotation. Before annotating the small variants, the GATK best practices workflow recommends filtering SNVs and InDels usingVQSR because the raw variants may contain many artifacts. The core algorithm in VQSR is a Gaussian mixture model that aims to classify variants based on the clustering of their annotation values given a training set of variants with high confidence. Then, the VQSR tools use this model to assign a new confidence value to each variant, called VQSLOD, a logarithmic ratio of the probabilities that the variant belongs to the positive and negative models. The resulting VCF file was split by chromosome and annotated with the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016). Each variant was annotated with genes, transcripts, and consequences, as well as other useful prediction values obtained from CADD (v.1.6), polyphen (v.2.2.2), SIFT (v.5.2.2), MPC (from dbNSFP4.3a_grch38), and allele frequency information from GnomAD genome allele frequency (v3.2.1), predictions of splicing and UTRs regions with SpliceAI and UTRannotator, clinical information such as ClinVar annotations (202205), and more.

Quality controls. Quality controls were performed on the raw data and on the sequencing file. FASTQ files were quality controlled using fastp, which performs trimming of adapters,

filtering based on quality information, and trimming of low-quality reads (Li and Durbin, 2009). In addition, BAM files were checked using SAMtools with the quickcheck option and Picard with the ValidateSamFile option. Mosdepth was used for calculating genome-wide sequencing coverage.

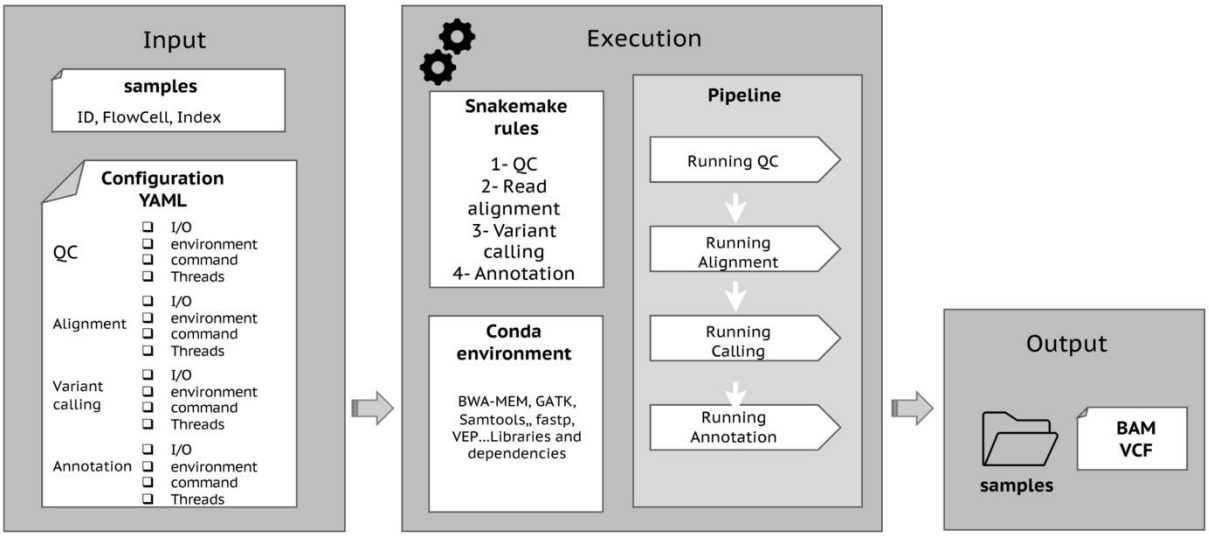


Figure 5. **Schematic representation of SNVs/InDels pipeline running using Snakemake**.

3.4 De novo SNVs identification.

De novo SNVs and indels were called in a custom pipeline using Platypus v0.8.1.1. Putative de novo variants were detected using the Platypus bayesiandenovofilter.py script, which takes as input Platypus VCF and pedigree file. The script was run with the following thresholds: genotype and mapping quality ≥ 30, minimum sequencing depth of 8 reads in the proband and both parents. Finally, only Mendelian inconsistencies on the autosomes (1-22), X chromosome (X), and mitochondrial chromosome (M) are considered and considered DNMs for

downstream analysis using the filter PASS. The output VCF file was annotated with VEP. The pipeline used is stored in GitHub repository https://github.com/Manuelaio/WGS_SNV_pipeline/tree/main/DNM.

3.5 DRAGEN Illumina pipeline.

The DRAGEN (v4.0.3.) germline pipeline was run using the Basespace Illumina web application. The germline pipeline includes highly optimized algorithms for mapping, alignment, sorting, duplicate marking, and calling of SNVs/Indel, CNVs, SVs, STRs, and ROHs. In addition, the pipeline provides quality control reports at three levels: for FASTQs, for alignment files (BAMs/CRAMs) and for VCFs. A schematic overview of the DRAGEN pipeline is shown below.

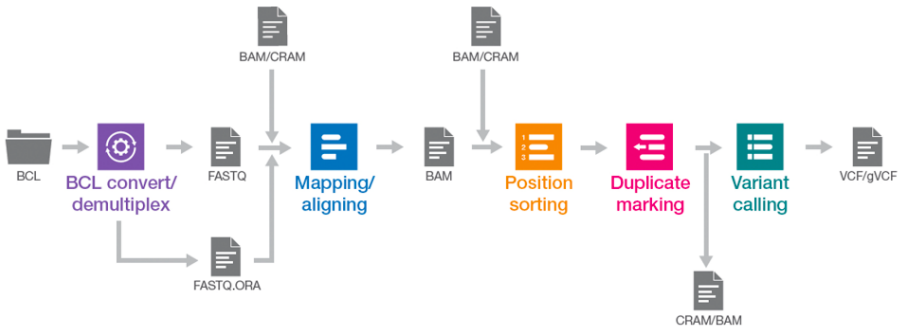


Figure 6. **DRAGEN Germline Pipeline** from (www.illumina.com). The figure shows a general schematic of the DRAGEN germline pipeline. The pipeline takes as input a FASTQ or CRAM file and performs alignment to reference genome, duplicate marking and calling of all types of variants.

3.6 SVs detection, genotyping and filtering

SVs calling tools. Of the more than 80 tools currently available for short-read WGS data, we have made a pre-selection of 8 tools based on different detection methods (split reads, read depth, paired reads, and de novo assembly) that do not require multiple samples (e.g. matched

sample as control and tumor samples) and that were previously assessed to be among the best performing approaches (Kosugi et al., 2019). The tools were then selected based on their compatibility with the following computation requirements: installation via the conda channel (https://docs.conda.io/en/latest/) and execution on the Linux system without .Net core requirment. The selected tools were run as SLURM jobs and with the recommended settings, but the specific commands and options used to test the tools were provided in the following GitHub repository (https://github.com/Manuelaio/Sv_pipeline/tree/main/benchmark). We adopted the VCFv4.1 as the universal format used in this study. Scripts to convert custom formats to VCFv4.1 are also required for some tools and are included in the repository.

Tool	Detection method	Reference
Manta (v.1.6.0)	SR, PR, AS	(Chen et al., 2016)
Delly (v.0.9.1)	SR, PR	(Rausch et al., 2012)
SVABA (v.1.1.0)	SR, PR, AS	(Wala et al., 2018)
ERDS (v.1.1)	RD	(Zhu et al., 2012)
Lumpy/Smoove (v.0.2.8)	SR, PR, RD	(Layer et al., 2014)
CNVpytor (v.1.2.1)	RD	(Suvakov et al., 2021)
Canvas (v.1.35.1)	RD	(Roller et al., 2016)
MELT (v.2.2.2)	PR	(Garder et al., 2017)
SURVIVOR (v.1.0.7)		(Jeffares et al., 2017)

Table 4. **SV-calling tools selected for benchmarking analysis.**

To compare a consensus approach with a union approach, VCFs obtained from different callers were combined with SURVIVOR as shown in Figure 7 (Jeffares et al., 2017). The individual SVs of the same type and strand were combined with a maximum allowable distance of 1Kb measured pairwise between breakpoints.

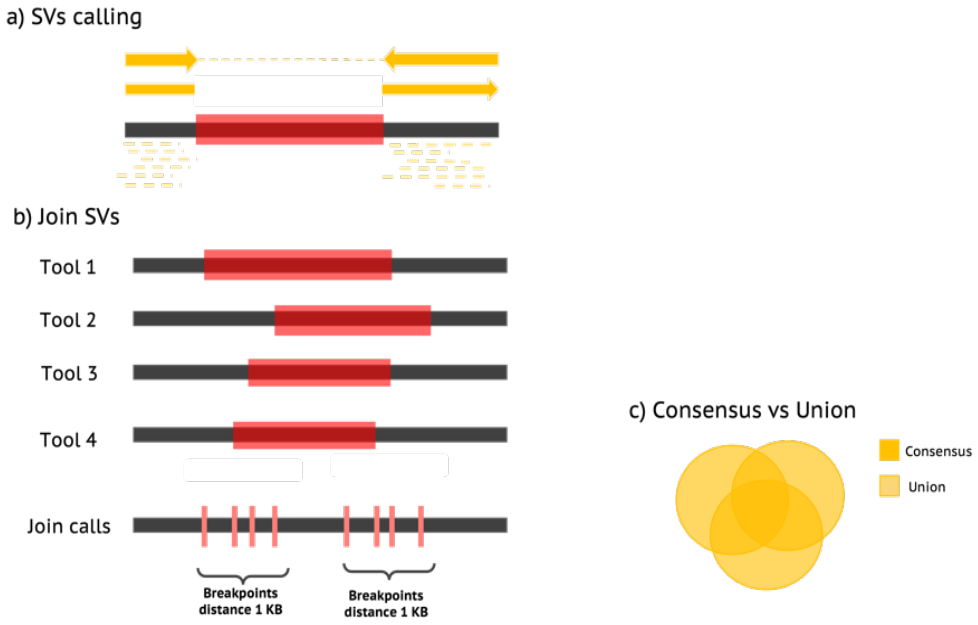


Figure 7. Flowchart of analysis steps. a) Variant detection: an example of a DEL detected by a combination of different signals in the following order: PR, SR, RD. b) The breakpoints of SVs may be inaccurate and different tools may call the same variant with different breakpoints (Tool 1, Tool 2, Tool 3, Tool 4). For this reason, when merging, we overlap variants that have the same orientation and a maximum breakpoint distance of 1KB as specified by SURVIVOR. c) Comparison between a union approach and a consensus approach. In the consensus approach, the SV must be called by multiple callers (dark yellow area), while in the union approach, all SVs found are evaluated (light yellow + dark yellow).

3.7 Reference SV dataset for real data.

The real datasets used for the benchmarking analysis include: the associated GIAB NIST Tier1 v.06 Benchmarking callset provided by the Genome in a Bottle (GIAB) consortium of the Ashkenazi son sample (HG002/NA24385), the REACH00236 sample, an in-house gold standard obtained from the intersection of calls from Pacific Bioscience and Oxford Nanopore Technologies and available at the University of California - San Diego, and the NA12878 sample from the 1KG project (https://www.genome.gov/27528684/1000-genomes-project) used in the previous study.

The GIAB high confidence SVs calls include curated deletion and insertion genotype calls determined by consensus calling from more than 30 callers with different sequencing technologies (Illumina Sequencing, long reads, 10x Genomics) from the GIAB community. The v0.6 SV benchmark set for HG002 is only available in hg19 (ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/) and several user limitations are reported, such as the sequence being inaccurate, the truth set resolving only simple SVs, and containing only about 50% of the SVs in the genome. A liftover to hg38 was performed using Picard tools (https://broadinstitute.github.io/picard/).

Sample REACH00236 contains all types of SVs and was obtained from the intersection between PB and ONT calls (Brandler et al., 2018).

The reference set for NA12878 has been described in detail in a previous publication (Kosugi et al., 2019). Briefly, this set was obtained by combining several datasets and also includes a set of experimentally verified inversions from long read studies and the InvFEST database (http://invfestdb.uab.cat). For NA12878, the reference dataset is only available in hg19, so a liftover was performed to obtain an hg38 version. Alignment files for the samples used in the benchmarking analysis were downloaded from the following repositories:http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/illumina_platinum_pedigree/data/CEU/NA12878/alignment/ and https://www.nist.gov/programs-projects/genome-bottle, with the exception of REACH00236, an in-house sample available at UCSD. All samples were aligned using BWA-MEM with the -M option for secondary alignment, and the recalibrated BAM files were used as input to the callers.

3.8 SV detection benchmarking

To measure the performance of each SV detection tool we compared them to the truth sets (TSet). Since the performance of each tool depends on the type and size of SV, precision and recall were calculated for each class and for different size ranges. Specifically, the following size ranges were considered: 50bp-300bp, 300-500bp, 500bp-1Kb, 1Kb-5Kb, 5Kb-10Kb, 10Kb-20Kb, 20Kb-30Kb, 30Kb-40Kb, 40Kb-50Kb, and greater than 50 Kb. Filters were applied (before benchmarking) on variants in canonical chromosomes (1-22, plus X, Y, and M) and were filtered based on the PASS filter flag. For benchmarking analysis, Bedtools was used with the reciprocal-overlap option. True positive (TP) calls were defined as SVs that had ≧ 50% reciprocal overlap with the reference TSet. In the case of insertions, because the insertions have no physical span over the reference, reciprocal overlap was based on the midpoint between the start position and the length of the INS +/-100 bp. False positive calls (FP) were defined as SVs that did not overlap with the TSet. False negative calls (FN) were defined as calls exclusive to the truth set. To compare the accuracy of the SV callers, the following metrics were used:

Precision	**Recall**	***F-Score***	**FDR**
TP/(TP + FP)	TP/(TP + FN)	2*Recall*Precision/(Recall+Precision)	FP/(FP + TP)

All benchmark analyses were performed in R (v.4.2.1) while VCF manipulation was performed in Python (v.3.8.1) using the pysam module (0.20.0).

3.9 Building a Nextflow pipeline.

The SV pipeline (*nx-pipeline*) is written in Nextflow. Nextflow is a widely used workflow manager for bioinformatics (Wratten et al., 2021). It is based on a domain-specific language (DSL) with a dataflow paradigm that makes it compatible with any scripting language. In Nextflow, users can run processes from Docker and/or Singularity containers (or with conda environments), which enables users to more easily create portable, reproducible pipelines. It also provides built-in support for high-performance computing environments such as SLURM and cloud computing services such as AWS, Azure Cloud, and Google Cloud. The pipeline was built with the Nextfow version (v.22.04.0) and tested on the Expanse compute cluster at San Diego Supercomputer Center (https://www.sdsc.edu/services/hpc/expanse/) and on the in-house infrastructure at IRCSS AOU BO hosted by Lepida in Ravenna, Emilia-Romagna, Italy. Reproducibility of the pipeline is ensured through the use of versioned Docker images containing SV callers with their dependencies. Due to security concerns regarding the use of Docker containers on clusters, during execution in the cluster environments, these Docker containers are converted into Singularity containers which run the same process. Singularity execution was also included to facilitate the use of the pipeline in high-performance data centers where the use of Docker is restricted by security regulations (Kurtzer et al., 2017).
The pipeline was built in independent modules, but sequentially. Users can design pipeline execution through a configuration file, which specifies the execution environment (e.g. AWS Batch, SLURM, Torque), the number of cores, memory, and time to allot for each type of process. Jobs within each module can be executed in parallel to accelerate the analysis of large numbers of samples.

Sample QC Module: This process was developed to gather quality control metrics from input files (BAMs/CRAMs)

Variant Calling Modules: The pipeline includes three tools for SV calling: Manta (v.1.6.0), Delly (v.0.9.1), Smoove (v.0.2.8); CNVpytor (v.1.0) for detection of CNVs and ExpansionHunter for detection of STRs for STRs detection (v2.5.5)

Merging, genotyping and filtration Module: The CNVpytor output, CNV-VCF, was merged into a multisample file, and telomeres, heterochromatin, contig, and scalf regions (the gaps in GRCh38 genome assembly) were filtered out. Instead, the output of Manta, Delly and Smoove

were merged with SURVIVOR in each sample, and SVs with a length of less than 50 bp and with a quality filter other than PASS were filtered out.

Post-processing analysis and *De novo* SVs identification: The Nextflow pipeline creates a VCF with unions of all SVs calls across multiple callers. SURVIVOR is also used to merge SVs across samples to create a cohort VCF (Figure 8). A joint-genotyped VCF then allows the allele frequency of variants in the cohort to be defined, enabling the identification of singleton and recurrent variants. After merging, each variant in each sample was annotated with ClassifyCNV and assigned to one of three quality scores based on the size of the SVs and the number of callers supporting the variant. Clinical interpretation of SVs is not yet standardized, however the ACMG has recently published guidelines for clinical classification of CNVs (Riggs et al., 2020). Because the ACMG guidelines consider a large number of rules and metrics, we use a tool called ClassifyCNV, which rapidly classifies variants according to the guidelines (Gurbich and Ilinsky, 2020). However, to guide the downstream analysis, each variant in each sample was assigned to one of three categories, "Weak," "Medium," and "Strong," based on the current benchmark results.

After creating a joint-call set for the cohort and annotating with AMCG consequence information and custom scores for each variant, a Nextflow pipeline for *de novo* SVs calling (dnSVs) was implemented using SV2 (Antaki et al., 2018). SV2 is a machine learning algorithm for genotyping deletions and duplications detected in SR WGS data. It uses the features of read depth, discordant read ratio, split read ratio, and heterozygous allele ratio, to estimate genotype likelihoods for each DUP and DEL variant in each sample. These genotype likelihoods were also used as quality metrics, and the variants with the DENOVO_FILTER tag are selected as dnSV candidates. The potential dnSVs detected by SV2 were further filtered by the following procedures. First, a manual visual inspection of the IGV view of each dnSV in the corresponding trio was performed and only singletons were considered. The VCF file was annotated with VEP (v.108) using some custom plugins such as the gnomadSV dataset for population-level allele frequency and overlap with segmental duplication regions (Collins et al., 2020).

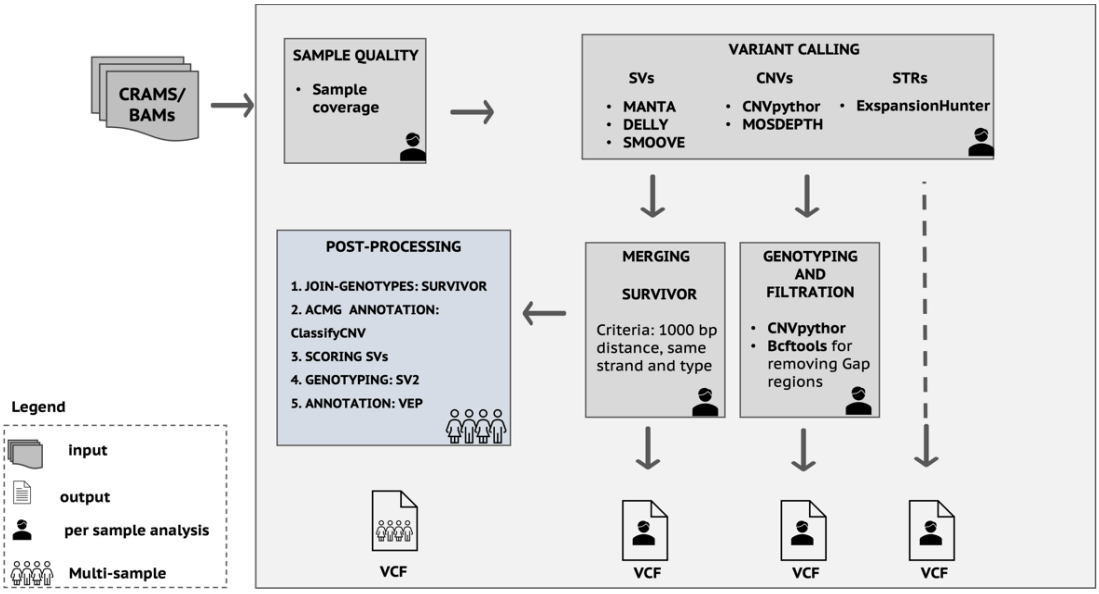


Figure 8. **Schematic representation of the *nx-pipeline***. The *nx-pipeline* takes as input the CRAMs/BAMs files and calculates sequencing coverage as a quality control. In the second module, *nx-pipeline* runs three different tools per sample to call all types of SVs, CNVpytor and mosdepth for CNVs and ExspansionHunter for STRs. For each sample, the output files from the SV tools are merged with SURVIVOR. In the post-processing module, SVs from multiple samples are joined, annotated with both clinical consequence (ACMG prediction) and reliability score, genotyped with SV2, and finally annotated with VEP.

3.10 Clinical prioritization

Interpretation of variants is the major bottleneck in clinical genome sequencing. This usually requires manual curation of a large number of variants,searching multiple databases, and computational tools. To streamline variant interpretation, we are developing Rabdomyzer, an in-house program that performs gene-based prioritization of variants identified in sequencing data.

Rabdomyzer is an open-source application written in Perl and implemented as a command line tool. It requires an annotated SNVs/Indels VCF file, either multi-sample and single sample, and a model file containing sample information, phenotypes, case-control status, and relationships between samples. The program consists of three modules: parsing the VCF file, filtering, and merging. An overview of the Rabdomyzer program workflow is summarized in Figure 9.

Rabdomyzer takes as input an annotated VCF file and during the parsing VCF module phase, samples and user-defined annotation fields are extracted (e.g. CADD, HGVSc, Clinvar). Rabdomyzer is suitable for processing VCFs containing all possible annotations from all possible transcripts that overlap with a given position in the genome. The transcript is selected with the most severe consequence defined by default using the consequence severity order provided by Ensembl. However, users can also specify a file with a consequence order according to their own assessment of severity.

In the filter module, a gene-centric key-system drives variant filtration among a family according to the inheritance model (model file).

Variants are filtered based on three inheritance models:

- Dominant model: in autosomal dominant disorders, the affected proband shares a single variant with the affected parent or with another affected family member. Rabdomyzer filters for those variants where the affected parent (or other family member) and the affected proband are heterozygous. In the absence of the affected parent, Rabdomyzer returns all heterozygous variants identified in the proband that are reference in the parents or other family members used as controls.
- Recessive model-homozygous state: in autosomal recessive diseases, the affected proband inherits a pathogenic mutation from each parent (the parents are carriers). Rabdomyzer

filters for those variants where the parents are heterozygous carriers of the variant, and the affected individual has a homozygous state. Since chromosome X is designated as homozygous in males, an X-linked inheritance can also be analyzed in this model.

- Recessive model-compound heterozygous state: In the case of such a recessive model, Rabdomyzer filters for at least two heterozygous variants in the proband, where one parent is heterozygous, and the other parent should have a reference genotype.

The merging module annotates each gene with a Rabdomyzer-database containing OMIM morbid genes, tissue-specific gene expression and regulation (GTEx), and the SORVA dataset contains calculations on the number of individuals who have a rare variant in a given gene. The final output is a separate worksheet for each inheritance model.

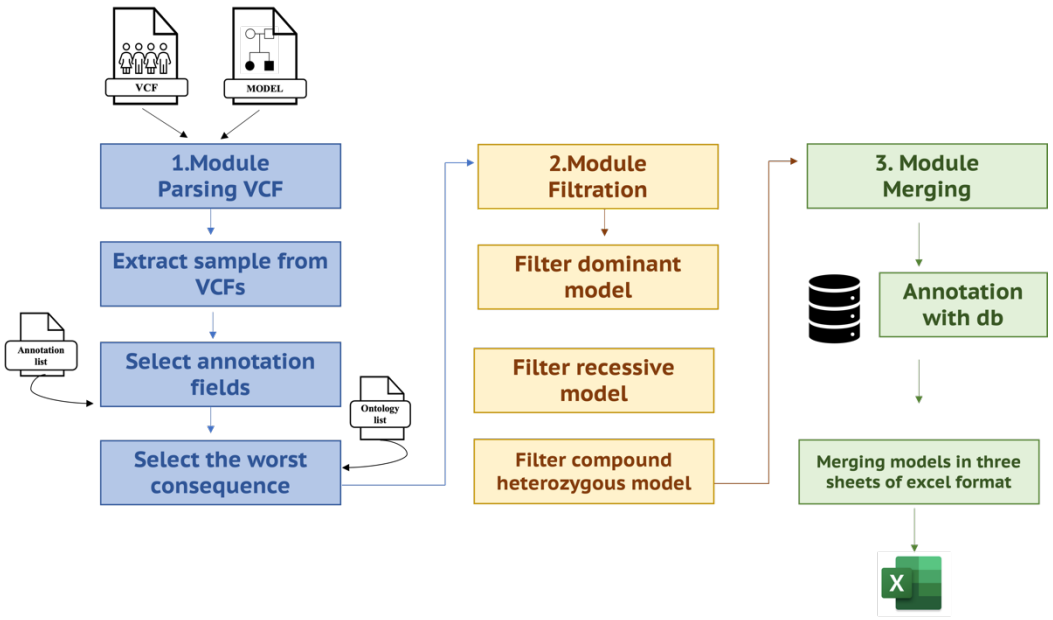


Figure 9**. An overview of the architecture of Rabdomyzer**. The required input files are VCF and model file. The model file is a tabulate file that contains in the first column the proband ID and eventually, separated by commas, the ID of the other affected family members. The second and third columns of the model file contain the ID of the parents or 0 if they are unavailable or affected. The three-step workflow for prioritizing SNVs/Indels is shown here: 1) parsing VCF module allows to extract the annotation fields from INFO field and the worst-consequence transcripts, 2) variant filtering based on the inheritance models, and 3) annotation with the Rabdomyzer database and preparing the output.

After running Rabdomyzer, exonic SNVs/Indels were filtered with restricted thresholds. Assessment of missense variants was restricted to variants with CADD score ≥ 20 or MPC ≥ 2 that were absent in the gnomAD database, in case of DNMs, or had a maximum allele frequency (MAF) ≤ 0.001% for recessive variants. Instead, evaluation of loss-of-function (LoF) variants were restricted to LoF-intolerant genes with a LoF intolerance (pLI) probability > of 0.97 or a LoF observed/expected upper bound fraction (LOEUF) ≥ 0.37. For splicing variants, we limited our assessment to those classified as splice-modifying by SpliceAI (DELTA > 0.5), while for 5'UTR variants, we focused on those that create or disrupt upstream open reading frames (uORF) according to UTRnator. Prioritization and interpretation of noncoding variants, especially intronic and intergenic variants, is not straightforward. We attempted to follow the recommendations recently published by Ellingford (Ellingford et al., 2022) for the clinical interpretation of noncoding variants. However, due to the high number of variants, the analysis of DNMs non-coding variants was restricted to variants not present in GnomAD or that segregate in affected families. In addition, we also used in-silico scores score for predicting the effect of non-coding variant (CADD > 17.4; ReMM > = 0.985; FATHMM_MKL > = 0.993) (Seaby et al., 2022) (Shihab et al., 2015) (Table 5).

A Rabdomyzer module for parsing SV-VCF is in progress. Prioritization of SVs was performed using a home-made Python script that follows the steps performed for SNVs. In addition, SV VCF was also annotated during the annotation step using synNDD, a systematic and curated catalog of published gene-disease associations implicated in NDD (Kochinke et al., 2016).

Variant type	Inheritance	Filtration	Prioritization
SNV/Indels	*de novo*	novel; ≥ 8 reads; QUAL >30	GnomAD ≤ 0.001% or novel
	Dominant	Gnomad novel, inherited from parents, ≥ 8 reads	CADD ≥ 20
	Compound heterozygous and Recessive	Gnomad ≤ 0.001, inherited one from and one from mother; ≥ 8 reads	SpliceAI: DELTA > 0.5 UTRannotator: annotation of uORF
	X-linked	Gnomad ≤ 0.001, inherited from mother,≥ 8 reads	Intronic: CADD > 17.4; ReMM > = 0.985; FATHMM_MKL > = 0.993
SV	*de novo*	novel; Score =STRONG, SR and PR ≥ 30 % of reads, novel in-house frequency	Gene = SYSNDD gene
STR	Various	ExpansionHunter catalogue	Repeat size > reference threshold

Table 5. Filtration and prioritization threshold used for different variant and different inheritance model

3.11 Diagnostic outcome definition.

The diagnostic outcome was evaluated based on the inheritance model at the individual family level. Three possible outcomes were defined:

1) Diagnosis (outcome 1): a pathogenic/likely pathogenic variant has been discovered in a disease gene associated with phenotype that is related to the patient's phenotype and explains the observed phenotype.

2) Potentially diagnosed (outcome 2): variant/s of unknown significance in an already established disease gene were identified that could explain the patient's phenotype. Or a pathogenic variant/s has been identified in a candidate disease gene that may be related to the patient's phenotype (or part thereof).

3) No conclusive result (outcome 3): no pathogenic variant/s were detected that could potentially explain the patient's phenotype.

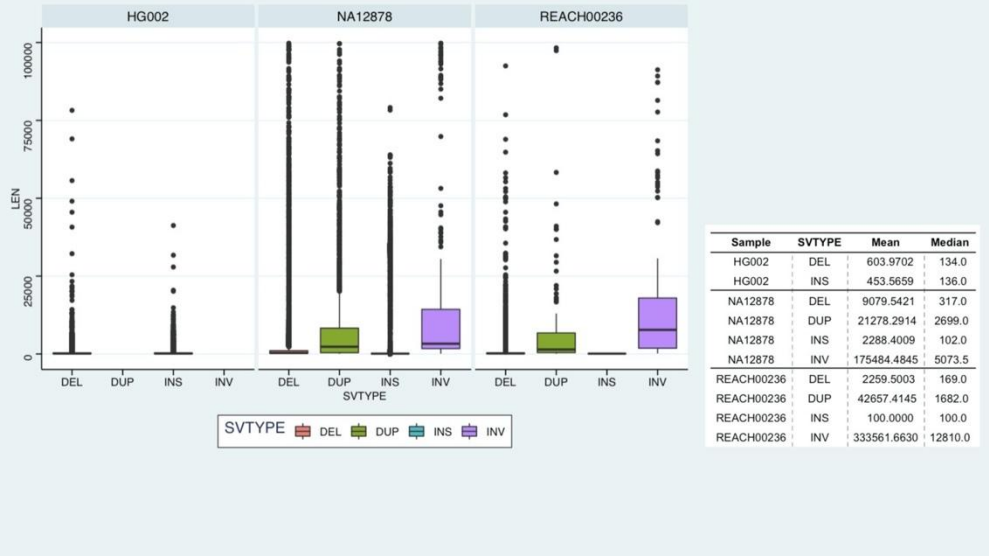
Moreover, clinical interpretation of STRs was limited to a variants catalog (provided by ExpansionHunter) with 30 well described disease associated STR loci (AR, ATN1, C9ORF72, DMPK, FMR1, FXN, HTT, ATXN1, ATXN2, ATXN3, ATXN7, ATXN10, ATXN8OS, AFF2, CACNA1A, CBL, CNBP, CSTB, DIP2B, GLS, JPH3, NIPA1, NOP56, NOTCH2NL, PABPN1, PHOX2B, PPP2R2B, RFC1, TBP, TCF4). Allele sizes for both replicates were extracted from ExpansionHunter data for all samples and compared to locus-specific expansion thresholds.

Results

4.1 SV benchmark analysis

We pre-selected 8 available SV tools for detecting SVs in SR WGS data. Two tools that did not satisfy our computational environment were excluded and finally the following were selected: Manta, Delly, Lumpy, ERDS, CNVpytor, SURVIVOR. The SVs called by each tool were compared to the real datasets to measure their performance on different types and sizes of SV and to evaluate how these tools can be used together (consensus vs union).

Three samples were used as real data sets (from different sources): HG002, NA12878, REACH00236. HG002 and NA12878 TSets were lifted over to hg38 (3.4% and 3.7% of the variants were not successfully lifted over). The size of SVs in the TSet ranged from 50 bp to 1 MB. In HG002 and REACH00236, almost half were smaller than 100 bp, as shown in Figure 10. In addition, the high- confidence SVs published for HG002 contain only deletions and insertions.



Sample	SVTYPE	Mean	Median
HG002	DEL	603.9702	134.0
HG002	INS	453.5659	136.0
NA12878	DEL	9079.5421	317.0
NA12878	DUP	21278.2914	2699.0
NA12878	INS	2288.4009	102.0
NA12878	INV	175484.4845	5073.5
REACH00236	DEL	2259.5003	169.0
REACH00236	DUP	42657.4145	1682.0
REACH00236	INS	100.0000	100.0
REACH00236	INV	333561.6630	12810.0

Figure 10. **SV length distribution in the truth set** . The figure shows the distribution of the size of SVs by class. The median value indicates that the real data sets are enriched with small deletions and small insertions.

4.2 Comparing the performance of SV callers on TSets.

The VCF file containing variants called from each tool were pre-filtered as described in the Methods section before comparison, and the number of SVs detected is shown in table 6.

Tot	DEL	DUP	INS	INV	TRA	Caller	Sample
14514	7215	n/a	7326	n/a	n/a	GIAB-TSet	HG002
10945	4994	600	2549	598	2204	Manta	HG002
2556	1155	823	48	530	0	Delly	HG002
3683	2215	1316	26	126	0	Svaba	HG002
2470	1526	944	0	0	0	CNVpytor	HG002
5250	3524	647	0	65	1014	Smoove/Lumpy	HG002
25337	9042	2512	13525	258	n/a	1KG-TSet	NA12878
9942	4395	691	2288	644	1924	Manta	NA12878
4524	3133	851	1	539	0	Delly	NA12878
4150	1827	2129	26	168	0	Svaba	NA12878
3103	2885	218	0	0	0	CNVpytor	NA12878
6250	3694	882	0	88	1674	Smoove/Lumpy	NA12878
17321	7645	152	9343	181	0	UCSD-TSet	REACH000236
10340	4615	732	2129	642	2222	Manta	REACH000236
5266	3532	1055	48	631	0	Delly	REACH000236
3579	2094	1287	52	146	0	Svaba	REACH000236
6806	1049	5757	0	0	0	CNVpytor	REACH000236
6481	4073	856	0	65	1487	Smoove/Lumpy	REACH000236

Table 6. SV calls summary called by tools and TSet.

This visual comparison shows that the number of CNVs called by each tool is lower than the number of CNVs in the corresponding TSet, while the number of balanced SVs as INV and TRA is bigger. Benchmarking metrics were computed for each caller and for each SV type across the size bins. Since some of the DUPs in the TSet could be reported as INS as described

by Zook et al. (Zook et al., 2020), the performance of INS/DUPs was evaluated jointly, and the results are shown in Figure 11.

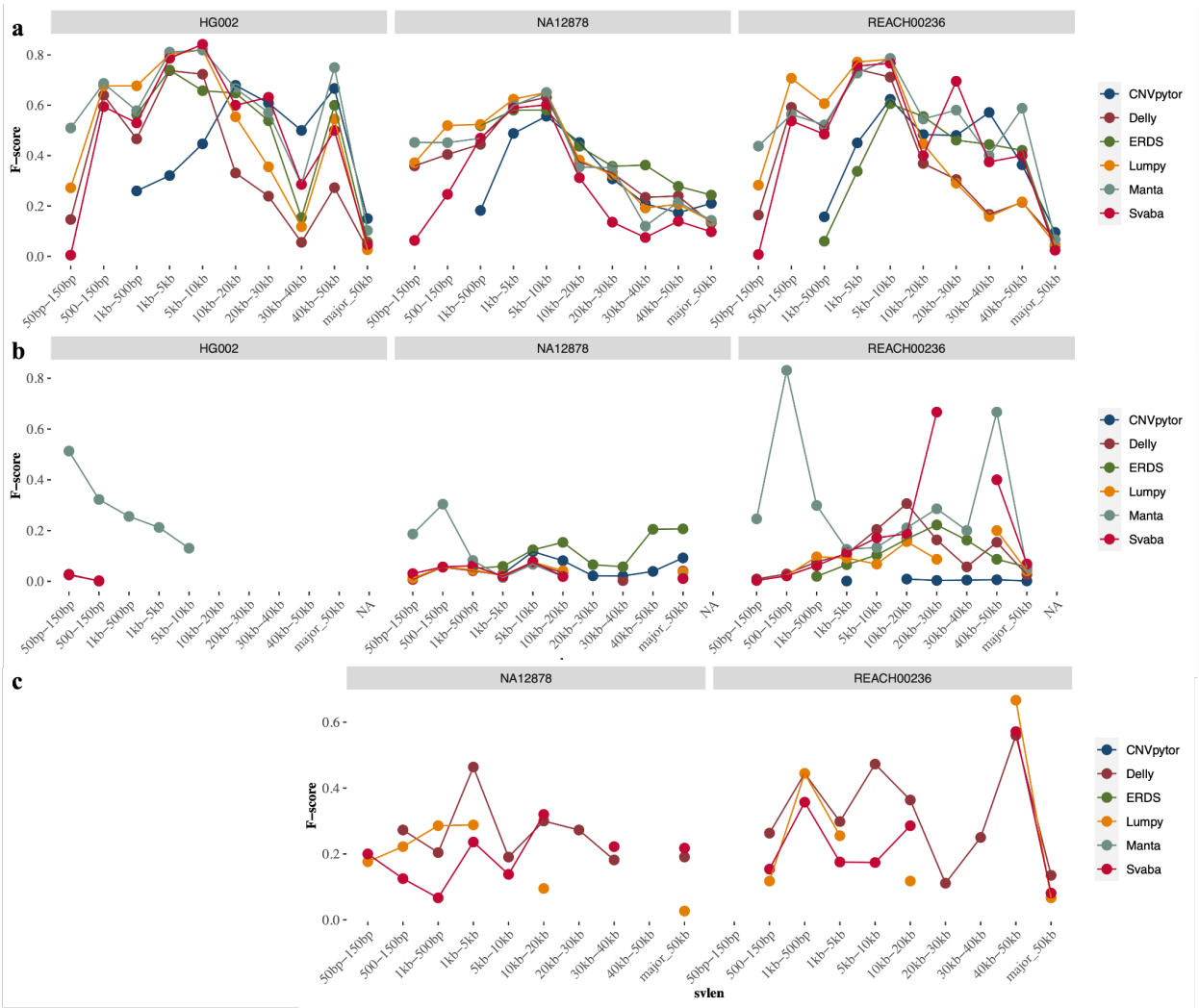


Figure 11. **Accuracy of SV algorithms.** F-score, from DELs (a), INS /DUP (b), and INV (c) were determined using 3 TSets: HG002, NA12878, and REACH00236. F-score, a combined statistic for precision and recall, is shown for each caller (colours of dots and line) and size range (x-axis)**.**

In general the F-score trend in NA12878 is slightly lower for the vast majority of sizes and tools, which is likely due to the quality of the benchmark set, which is older than HG002 and contains several redundant SV calls and inaccuracies (Liu et al., 2022). The precision and recall for calling deletions varied greatly by algorithm, and the size of SV as shown in Figure 11.a. Algorithms based on PR and SR, such as Manta, Delly, Svaba, and Lumpy, have high levels of accuracy and recall in a range of 150bp to 30Kb, while algorithms based on RD, such as ERDS and CNVpythor, show poorer performance in this range but perform better starting

from 5 Kb calls. In general, the performance of the tools decreases with increasing size of DEL in all samples, and the F-score for DEL > 50 Kb reaches low values (< 0.2). In general, it's not clear whether this behavior is due to the absence of these variants in real datasets or to errors during the variant calling (i.e .5 DELs greater than 50 KB are listed in HG002 TSet). In contrast, most of the INS /DUPs called do not match TSet, as found in previous studies (Kosugi et al., 2019) (Vialle et al., 2022) (Cameron et al., 2019) (Figure 11.b).

Regarding INS /DUPs, Manta performed best in each of the three samples exhibiting a 0.8 F-score value for INS/DUP ranging from 50 bp to 500bp. In general, the low F-score values observed for other tools, especially in the HG002 sample, are partly due to incomplete reference in TSet and partly due to overrepresentation of small INS in the TSet (52% of INS in the HG002 TSet are smaller than 150 bp). From this point of view, the REACH00236 TSet is the most comprehensive truth set as it comes from LR sequencing and confirms that Manta achieves a good F-score level compared with other tools that achieved an F-score of less than 0.2. Benchmark analysis of inversions (INV) was limited to REACH00236 and NA12878 (Figure 11.c).

Callers perform poorly on INV compared to DELs and generally achieve a low F-score (Figure 11.c). However, tools such as Delly and Lumpy achieve a satisfactory F-score, especially for precision, which reaches around 50% and 80% for INVs with ranging sizes of 1-5 KB in both NA12878 and REACH00236.

4.3 Union and consensus approach evaluation.

Benchmark results showed that the performance of the tools varies substantially depending on the type and size range of SVs. The DEL SVs are identified with the greatest precision and accuracy. Tools based on PR and SR are able to identify DELs ranging from 50 bp to 30 KB in all three samples with adequate/satisfactory performance. However, Figure 11.a shows that no tool is better than the rest for all types/size bins. Tools relying on RD show poor performance for the better ranges of the PR-SP tools, while they are better at identifying size variants of 30-40KB.

Calling INSs/DUPs is challenging, and the poor performance is likely also due to the fact that DUPs in particular are underrepresented in the TSets. In this case, Manta performs best among the tools tested. Delly and Lumpy, on the other hand, allow the detection of INVs with adeguate performance. Based on these considerations, combining the SV calls from different tools seems to be the best solution. The calls from each tool were combined using SURVIVOR, as described in the methods. At this stage, we excluded SVABA from subsequent analyses. SVABA returns a VCF in which the SV type is encoded as breck-end (BND), unlike the other tools that output the SV type (i.e. DEL, DUP, INV, INS, TRA). Although this problem has been reported several times in the GitHub repository, no consistent method for converting the SV types is provided by the owners. In addition, the ALT alleles are encoded differently from the other tools, leading to problems in the SURVIVOR merging step. Based on these considerations, we decide to exclude SVABA. ERDS was also excluded from subsequent analyses. ERDS requires the BAM format as input, while the other tools accept both the BAM and CRAM formats as input, the latter becoming more commonly used. Manta, Delly, Lumpy/Smoove (MDL) and CNVpythor were used for testing the union vs consensus approach.

As shown in the Venn diagram, many SVs are caller-specific (Figure 12.a), with the percentage of unique calls ranging from 21% of Delly's unique calls to 93% of CNVpytor's unique calls, while the second largest category of SVs were consensus calls from two of the callers, such as Delly, Manta, and Lumpy/Smoove (Figure 12.b). CNVpytor relies on RD as its detection signal, and as the previous benchmark analyses show, its performance increases with the size of the SVs, unlike the tools based on PR and SR. The difference in performance explains the large number of unique calls made by CNVpytor (Figure 12.b). Next, we investigated whether the merge approach is better than a consensus approach, especially for DELs where no tool outperformed the rest among all sizes, as in the case of INSs/DUPs. As Figure 12-a shows, the consensus approach would miss about 62.5% of the true positives identified by Manta as unique calls. In contrast, Delly and Lumpy's unique calls do not provide the same advantage.

a


DELLY (4799)
LUMPY (5149)
MANTA (5429)
CNVpytor (6805)
980
1570
982
371
2
2013
33
2480
6349
393
155
11
1
16

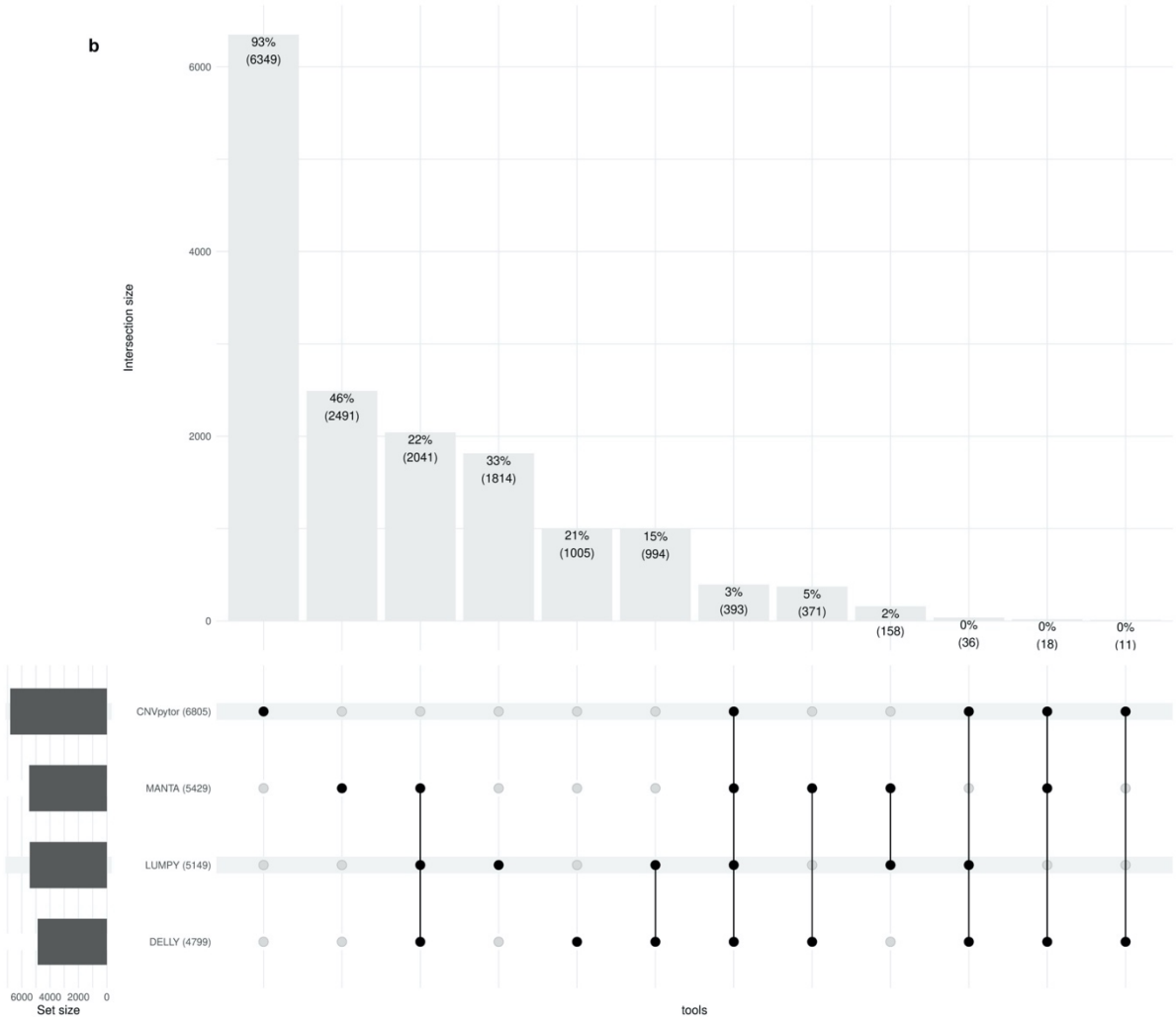
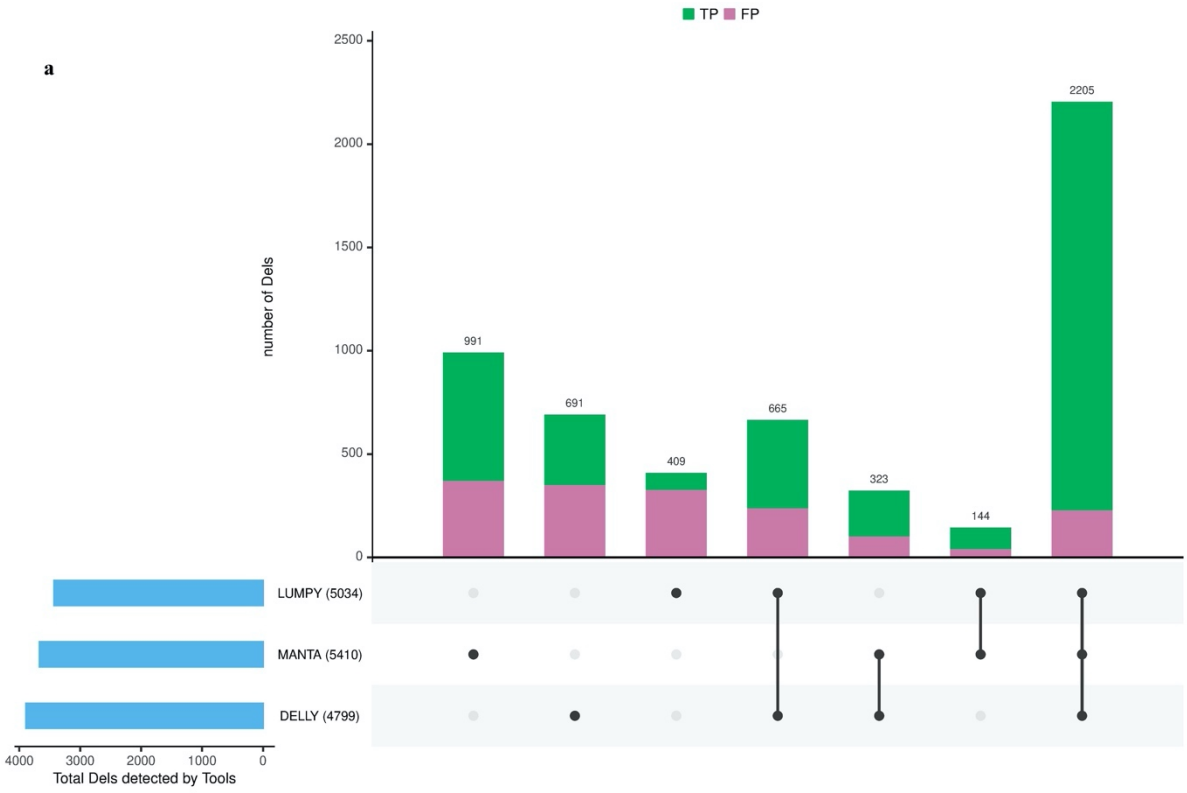


Figure 12. **(a) Venn diagrams** include the germline SVs detected in sample REACH00236 by joining Manta, DELLY, LUMPY, and CNVpytor calls. Most CNVpytor SVs are caller-specific, while SVs detected by MDL callers show a higher degree of overlap. **(b) Upset plot**. The upset plot shows the number and percentage of SVs called by a single tool and by multiple tools (consensus). The horizontal bars represent the total number of SVs identified by each tool, while the vertical bars show the occurrence and co-occurrence of SVs identified by different combinations of tools, indicated by dots and lines in the box below. The number of occurrences and percentage are shown within each bar.

Furthermore, we compute the precision, recall, and F1 score for the consensus and union approaches. We used two merging sets, one with the 4 selected callers and one without CNVpytor, to test the effects of unique calls (93%) on the F-score (Figure 13.b). According to

the performance metrics, a conservative approach (Union) achieved a higher F-score than the consensus approach. The many unique calls of CNVpytor affected the F-score, so we excluded it from the Union calls. To limit the assessment of FPs during downstream analysis, information on a set of callers supporting variants was modeled for scoring the reliability of each variant in each sample as described in the Methods. In addition, several published papers agree that FPs occur in low-complexity genomic regions, such as repeats and GC-rich regions, which can lead to ambiguous read mapping that results in false read alignments in all samples that are then readily detectable in a cohort VCF (Cameron et al., 2019) (Gong et al., 2021)(Ho et al., 2020).



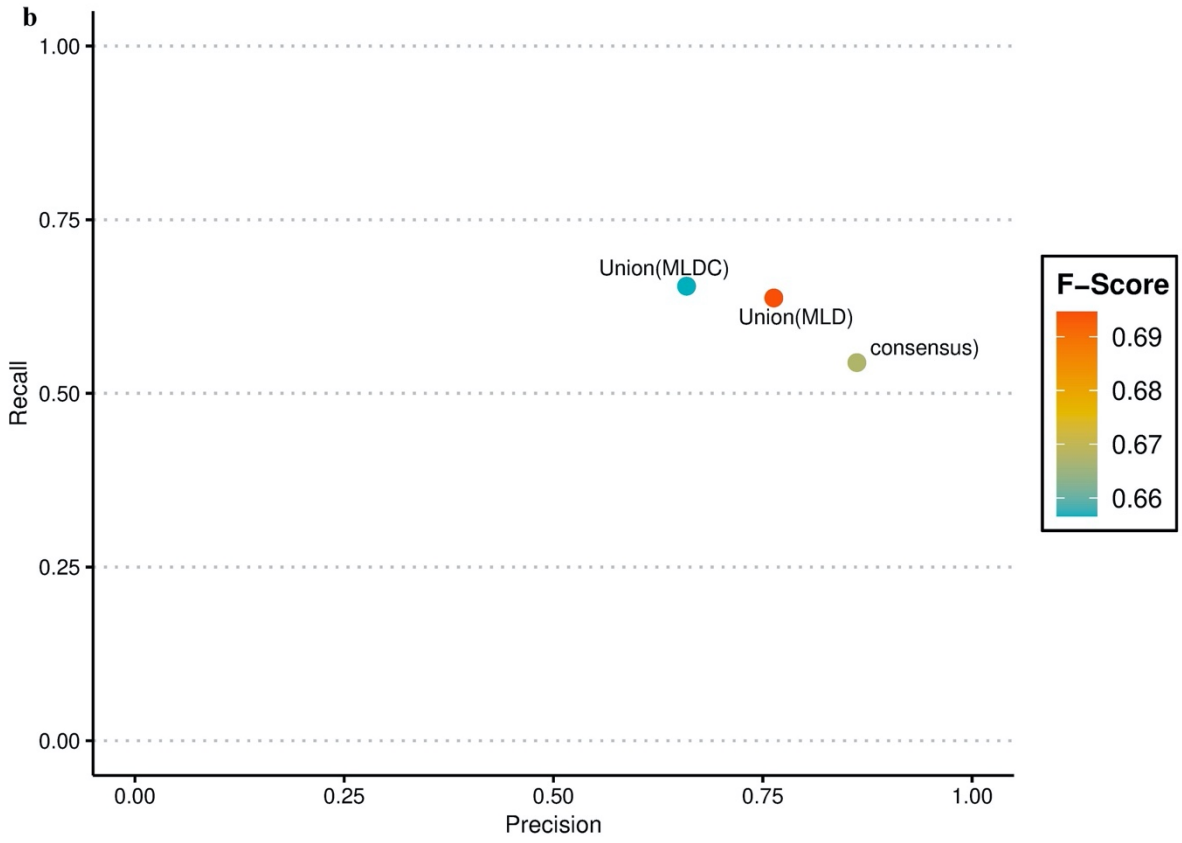


Figure 13. (a) DELs calls detected in sample REACH00236 using a union of Manta, DELLY, LUMPY, CNVpytor. The horizontal bars show the total number of Dels identified by each tool. The vertical bars show the occurrence and the co-occurence of Dels identified by different combinations of tools, and the number at the top shows the cardinality of the set. The black dots and lines indicate which tool or combination is represented in the corresponding vertical bars. (**b**) Precision and recall in consensus and union calls. There are two types of union sets, one set contains all callers (MLDC), and the other one does not include CNVpytor (MLD). The F-Score values are expressed by a color gradient, and the legend on the right side of the panel defines the value in terms of color.

4.4 Nextflow pipeline.

To enable automated parallel execution of tools on a high-performance computing (HPC) system, we created a pipeline (*nx-pipeline*) for clinical detection of SVs using Nextflow as the WfMS. We tested the pipeline with each WGS dataset (i.e., HG002, NA12878, and REACH00236) on Slurm-based HPC systems, and Table 4 shows the computational resources used.

Caller	Sample	Threads	time	peak mem
Manta	HG002	12	25m 54s	25.8 GB
	NA12878	12	31m 29s	21.3 GB
	REACH00236	12	55m 29s	25.4 GB
Delly	HG002	2	2h 11m 58s	8 GB
	NA12878	2	1h 33m 27s	4.6 GB
	REACH00236	2	2h 41m 46s	8.2 GB
Lumpy/ Smoove	HG002	4	2h 11m 34s	8 GB
	NA12878	4	1h 33m 16s	4.6 GB
	REACH00236	4	2h 41m 24s	8.2 GB

Table 4. **Compute resources used by each SV callers in *nx-pipeline*.**

The resultant variants of each caller were first filtered, considering only the PASS variants with a length of at least 50 bp, and then merged with SURVIVOR. The output of *nx-pipeline* shows higher precision/recall metrics for DEL in all samples. The benchmark result of INS /DUP shows that Manta remains the best tool in terms of precision for all samples. However, considering that TSet is incomplete for both HG002 and NA12878, the results in sample REACH00236 suggest that our pipeline achieves recall values close to Manta.

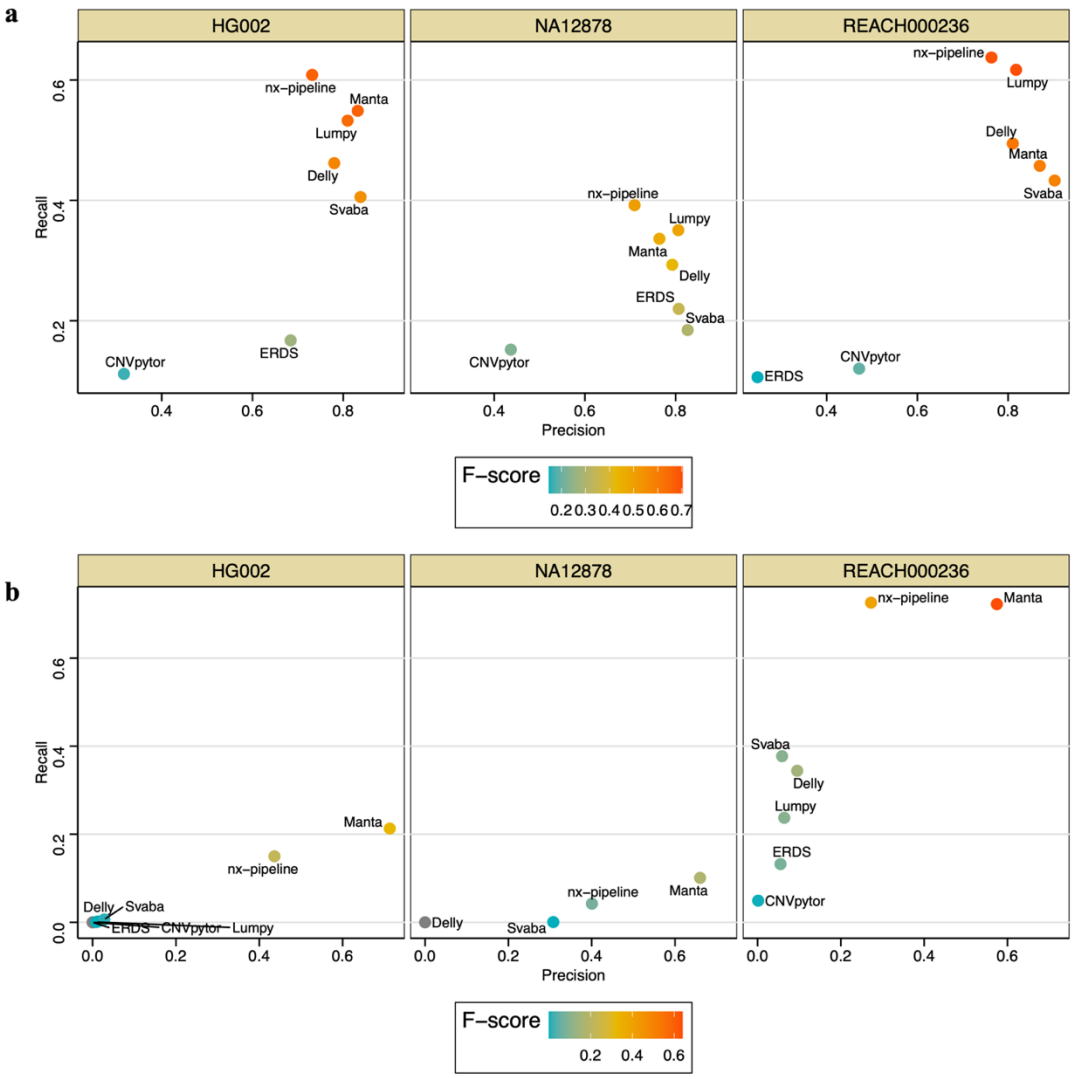


Figure 14. Overall precision, recall and F1 score of individual SV callers and *nx-pipeline* in DEL (a) and in INS/DUP(b). The F-Score values are expressed by a color gradient, and the legend on the right side of the panel defines the value in terms of color.

Finally, to support downstream analyses, we converted the results of the benchmark analysis into a reliability score. The score was defined as follows:

- WEAK: supported only by Smoove or only by Delly o supported only by Manta and with size > 30KB
- MEDIUM: supported by Manta only and with size ≤ 30KB supported by 2 or more callers and with size > 30KB

- STRONG: supported by 3 callers and with a size ≤ 30KB or INS supported by Manta and with a size ≤ 500 bp.

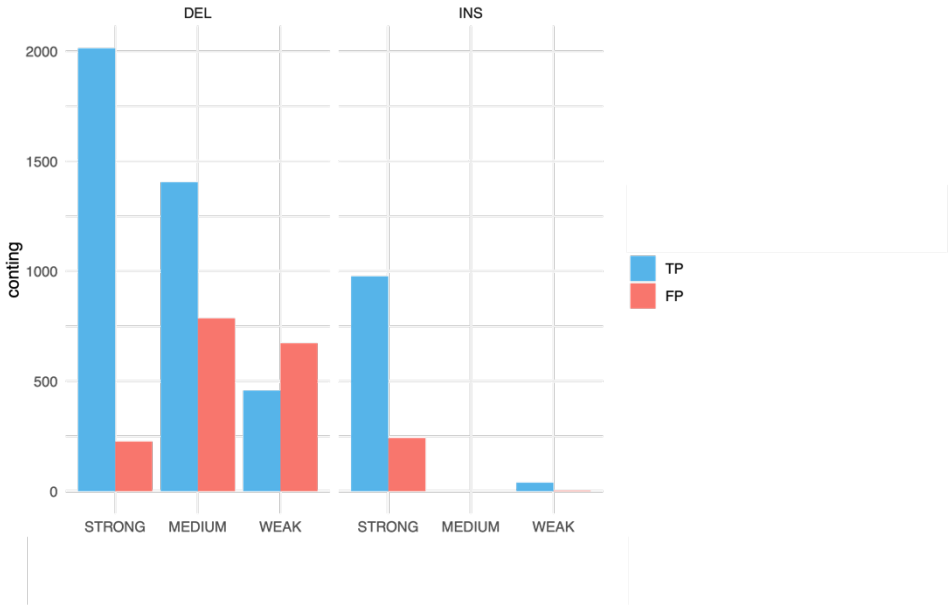


Figure 15: Number of TPs and FPs in REACH00236 based on score. Absolute number of TS both for DELs and INS based on different scores.

Results show that 90% of STRONG DELs are TPs (2014 at 2240 calls) and 80% of STRONG INSs are TPs (997 at 241).

4.5 Time Comparison: CPU vs FPGA Pipeline.

We performed WGS on 77 DNA samples from a cohort of 25 families. To process 6.7T FASTQs, workflows were run with Slurm on the Lepida server, a 1-node HPC cluster consisting of 85 Intel(R) Xeon(R) Gold 5218R CPUs available at Sant'Orsola Malpighi Hospital. More than 100 GB of temporary files (BAMs, GVCFs) were created for each sample, and an average of 30 GB were stored on CRAM. Overall, samples were processed at different speeds due to varying system load or availability at the time of use. The submitted job was successfully completed in approximately 77.9 hours for the *sn-pipeline* and 3 hours for the *nx-pipeline*, not including the alignment step. For the *sn-pipeline*, FASTQ file quality checking, alignment, and duplicate marking consumed more than 75% of the compute time required for the whole pipeline. We compare the compute time required to process the data from our CPU system to the DRAGEN (v4.0.3.) germline pipeline based on an FPGA system. The DRAGEN pipeline was the fastest, taking 1 hour and 30 minutes to process the WGS trio data and generating an average of 24 GB of data for the sample (Table 5).

	Lepida	**DRAGEN (v4.0.3)**
Time for upload file	0	~ 3 h
Processing time	~ 80 h	1 h 30 minutes
Memory	80GB	< 100GB
CPU	20	FPGA system
Output folder	~ 30GB	~ 24GB

Table 7. **Comparison of computation efficiency between Lepida server and DRAGEN on WGS trio.** The runtime comparison is shown in the first line. DRAGEN is significantly faster than the pipeline running on the hospital server. The memory required during data processing is comparable and the final output file size is also roughly comparable.

The computational efficiency of the DRAGEN-based pipeline is dramatically higher than the pipeline created for the CPU system, suggesting that a GPU or FPGA system can significantly reduce the processing time required for WGS data. However, gold standard tools such as GATK are not yet included in DRAGEN, and the generated VCF files require additional steps

such as joint-genotyping and VQRS to be suitable for downstream analysis (i.e., annotation and prioritization).

4.6 QC and Variant calling

After quality control all samples remained for further analysis. On average, 93.7% of genomic regions were covered at least 20X. The average sequencing depth was 47.6 X (Figure 16.a), and the average insert length was 344. (Figure 16.b).

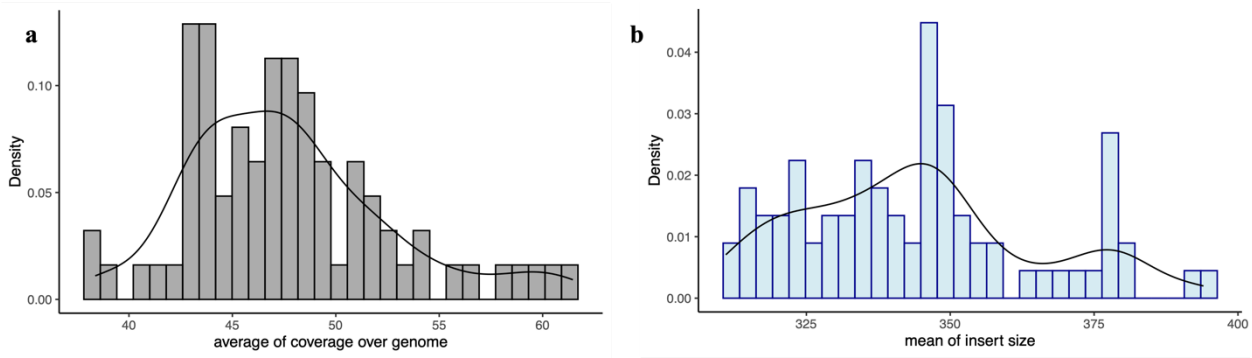
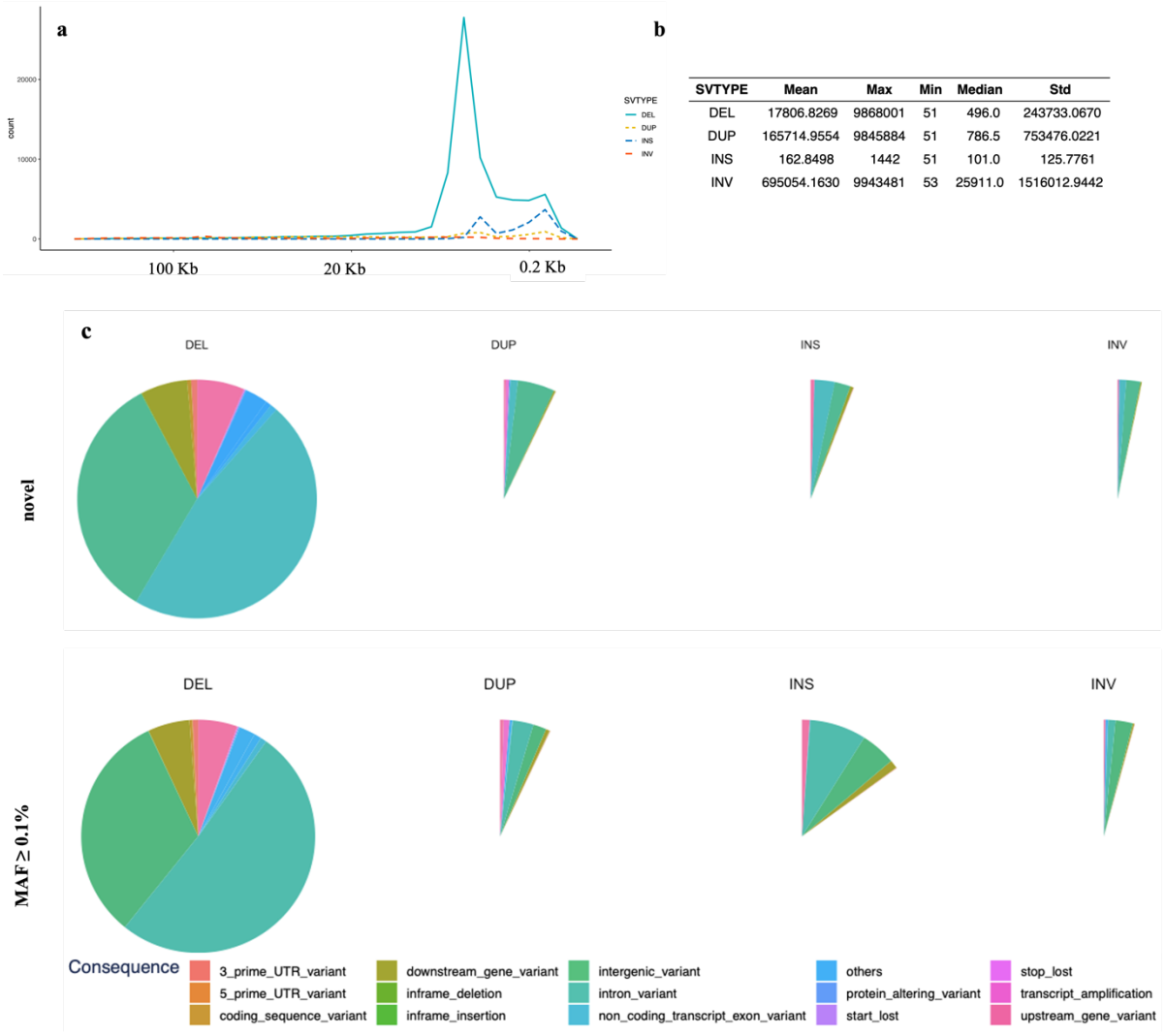


Figure 16. **Distribution of insert size and coverage in sequencing data of 77 samples**. a) Median and the mean of coverage was 47.1 and 47.6. b respectively) Median insert size was 344.

A total of 97,457 SVs were identified using *nx-pipeline* in 78 samples, which were used for all subsequent downstream analyses. Most SVs were small and consisted mainly of CNVs, with the frequency decreasing with decreasing variant size (Figure 17.a). To assess the frequency of variants, the VCF file was annotated with gnomAD- SV using VEP. We found that about 20% of SVs are novel and 54% of these SVs were discovered as singletons (allele count in cohort = 1) and about 87% of the singletons had a length of less than 1Kb. The vast majority of SVs, regardless of class and MAF, are located in intronic or intergenic genomic regions (Figure 17.c).



SVTYPE	Mean	Max	Min	Median	Std
DEL	17806.8269	9868001	51	496.0	243733.0670
DUP	165714.9554	9845884	51	786.5	753476.0221
INS	162.8498	1442	51	101.0	125.7761
INV	695054.1630	9943481	53	25911.0	1516012.9442

Figure 17. SV size distribution in log10 scale per SV type. Within the figure, a table summarizing the statistical values of the length of each class of variants.

Finally, after genotyping by machine learning (SV2), prioritization and manual inspection with IGV, 4 potential dnSVs were identified that were absent in the general population (gnomad- SV).

Using *sn-pipeline*, we identified 21,451,053 small variants, including 15,472,622 SNPs and 5,978,431 indels in 77 samples. Among those, 2,408 high-confidence DN SNVs and Indels

were prioritized. The number of DNMs in each proband ranged from 68 to 127 in the genome (zero to four in the exome) (Figure 18).

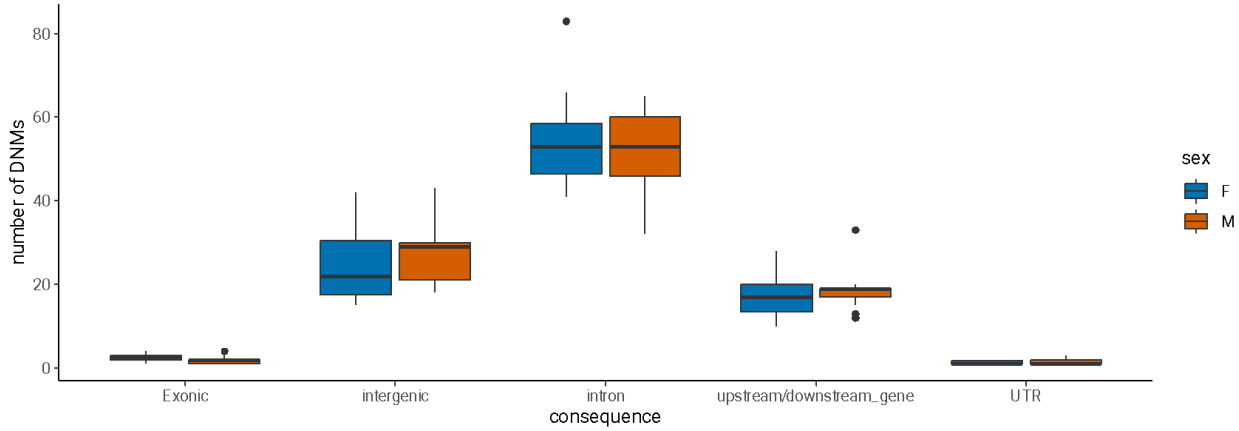


Figure 18**. Number of DMNs in female and male proband**. Exonic variants include missense, stop, frameshift, inframe deletion, and synonymous VEP consequence. A large proportion of DMNS arise in introns or intragenic regions in both males and females.

4.7 Clinical interpretation of WGS data.

Considering that millions of variants are identified, the first step of the downstream analysis was to narrow the search space to variants with traits most likely to cause genetic disease, based on the criteria mentioned in the method section. Our downstream analysis relies on Rabdomyzer, a genomic variant filtering and gene-driven prioritization software that captures all variants occurring in each gene based on different modes of inheritance. Rabdomyzer was performed for each family, and after accurate interpretation of the variants based on the criteria mentioned in the Methods section, we obtained a genetic diagnosis in 8 samples (32%) and a genetic candidate diagnosis in 5 samples (20%). In contrast, the genetic cause remained unclear in 12 patients (48%). Clinical findings of conclusive genetic diagnosis and possible diagnosis are shown in table 10

Case	Sex	Gene	Transcript	Variant	Inheritance	Outcome
FID_1	M	CREBBP	NM_004380.3	p.Arg1868Trp	AD (de novo)	1
FID_2	M	MT-ATP8	ENST00000361851.1	p.Lys57*	Mitochondrial	1
FID_3	F	RIF1	NM_018151.5	2-151426560-6821_del	AD (denovo)	1
FID_4	F	MED12	NM_005120.3	c.4477_4527+56dup	AD (denovo)	1
FID_5	F	CHD7	NM_017780.4	c.5210+1235A > G	AD (denovo	2
FID_7	M	HECW2	NM_001348768.2	p.Arg1330Trp	AD (denovo)	1
FID_8	M	AFF4	NM_014423.4	p.Arg258Trp	AD (denovo)	1
FID_9	F	LMBR1	NR_146959.1	c.424-5999T>G	AD	2
FID_10	M	TRIT1	NR_132405.1	p.Trp228Arg, p.Arg150Ter	AR	1
FID_11	F	-	-	inv(X)(p22.13q28)	AD	2
FID_19	M	NLGN4X	ENST00000381095.8	X-6197436+308_ins	XRL	2
FID_22	F	CYFIP2	NM_001037333.3	p.Asp877GlufsTer57	AD (denovo)	1
FID_24	F	RFT1	NM_052859.4	p.Lys152Glu, C.*2407C > T	AR	2

Table 10. Clinical findings and summary of pathogenic variants in the NDD. Outcome 1: diagnosis, Outcome 2: possible diagnosis. The table shows that the most common mode of inheritance identified in our cohort was AD which was found in 63% of our patients. The variant column includes the aminoacidic change, where possible, or cDNA position or the position of SVs.

4.8 Conclusive diagnosis

FID_1 D*e novo* p.Arg1868Trp (NM_004380.3) in the exon 31 of CREBBP are identified in proband of FID_1. CRREBP is a ubiquitously expressed gene that encodes CREB binding, a histone acetyl transferase, and is a transcriptional activator that interacts with several transcription factors and proteins. Heterozygous loss of function is known to cause Rubinstein-Taybi syndrome type 1 (RTS), the main features of which are facial dysmorphia with a characteristic grimace smile and broad thumbs, and halluces. In 2016, Menke et al. reported 10 different de novo missense variants affecting specific regions of exons 30 and 31 of these two genes, but without the specific phenotype of RTS (Menke et al., 2018). In particular, the phenotypes included feeding problems, autistic behavior, recurrent upper respiratory tract

infections, and hearing impairment. Since then, several research groups have described additional individuals who share the same genotype characteristics. In 2019, it was proposed that this syndrome be called Menke-Hennekam syndrome (MHS) after the authors who first described it. The same missense variant found in the FDI_1 proband (p.Arg1868Trp) was previously described in patients with MHS who had feeding problems, absent speech, microcephaly, and psychomotor delays (Menke et al., 2016)(Banka et al., 2019). In addition, the variant has been reported as pathogenic in ClinVar and is absent in Gnomad.

FID_2. A mitochondrial variant m.8535A > G (p.Lys57*) in MT -ATP8 was identified in proband of FID_2.

Mitochondrial ATP synthase (complex V) is a macromolecule consisting of 18 protein subunits, 16 of which are encoded by nuclear DNA and 2 by mitochondrial DNA (mtDNA) (MT -ATP8 and MT -ATP6). Pathogenic mutations in the ATP6 gene have been associated with Leigh syndrome or syndrome of neuropathy, ataxia, and retinitis pigmentosa (NARP), whereas pathogenic mutations in the ATP8 gene are less known in the literature and have a variable phenotype ranging from cardiomyopathies to epilepsy (Dautant et al., 2018). Because m.8535A > G had not yet been described, a functional study performed in our Medical Genetic Unit confirmed the pathogenicity of the variant (functional reduction of complex V). The percentage of heteroplasmy in the proband was 96% in blood, 96% in muscle, and 86% in urine. Moreover, the same pathogenic variant was absent in the tissues of the mother and sister. The clinicians diagnosed mitochondrial complex V deficiency.

FID_4. A small dSV was discovered in proband of FID_4. The dSV is a heterozygous small insertion of 57bp (chrX:71132902-71132959) affecting the last part of exon 32 of MED12 (NM_005120.3) and a small part of intron 32 (Figure 18). MED12 is one of 31 subunits (MED1-MED31) of the large multiprotein complex Mediator, which regulates gene expression and interacts with RNA polymerase II. Hemizygous missense variants in MED12 cause three distinct X-linked neurodevelopmental disorders (NDD) in males. A recent study reported 18 females in whom de novo variants in MED12 result in NDDs that partially overlap with the previously known phenotypic spectrum of affected males (Polla et al., 2021). The

clinical features reported for truncating variants in MED12 are consistent with the syndromic presentations of the patients. The INS was annotated as STRONG with our reliability scoring and also detected by GATK in the small variants pipeline.

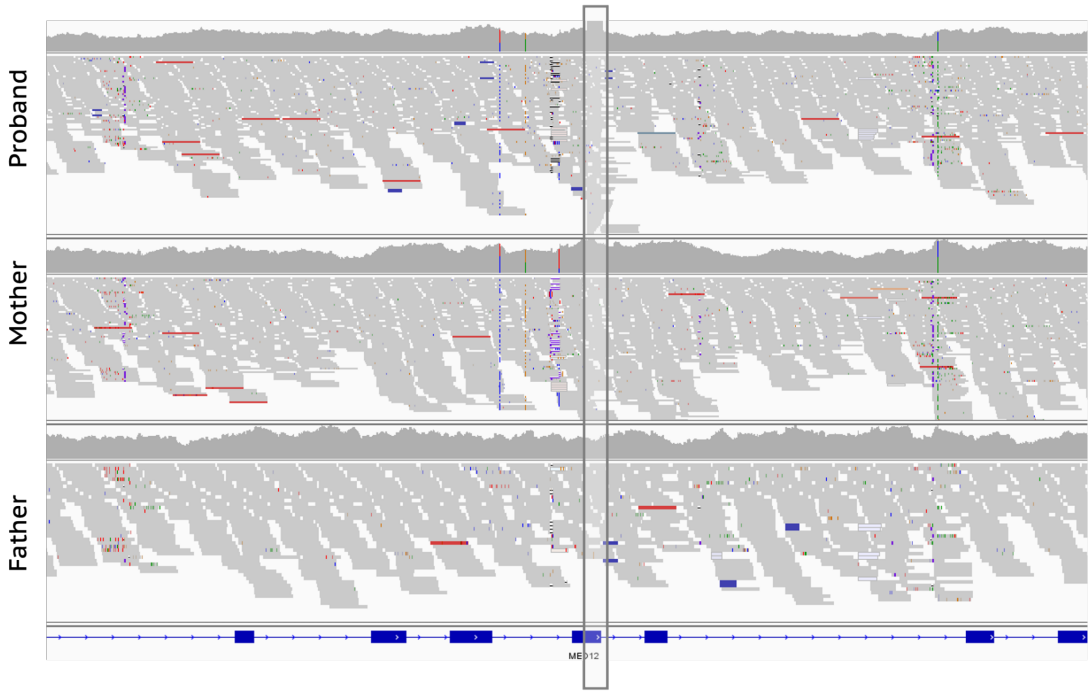


Figure 18. Visualization of genomic alignment of WGS data in Integrative Genomics Viewer (IGV) software. A 56 bp de novo INS in MED12 on Xq13.1-Xq13.1 (chrX:71132902-71132959) encompassing exon 32 of the proband.

FID_3. A dSV DEL (chr2:151426560-151433381) in RIF1 gene (NM_018151.5) was found in the proband of FID_3. The deletion of 6,821 Kb includes the entire exon 9, which was completely skipped, and introns 8 and 9 were partially deleted, resulting in a frameshift like alteration (Figure 19). RIF1 (Rap1-interacting-factor-1) appears to be highly intolerant of loss-of-function variants, as the predicted LoF mutations are extremely rare (LOEUF: 0 and pLI: 1). It has many functions, from controlling telomere length to regulating nuclear architecture, maintaining epigenetic state, and controlling cell cycle progression, but studies are still needed to understand its contribution during development (Richards et al., 2022). A recent study identified two frameshift variants in patients with global developmental disorders and ID (Seaby et al., 2022). The deletion was annotated as STRONG using our reliability scoring and

is detected in two different workflows, *nx-pipeline* and in Dragen, as well as using multiple lines of evidence, including RD (CNVpython with +/- 380bp from *nx-pipeline* output and Canvas BS-Illumina-Dragen with +/-360), SR and PR.

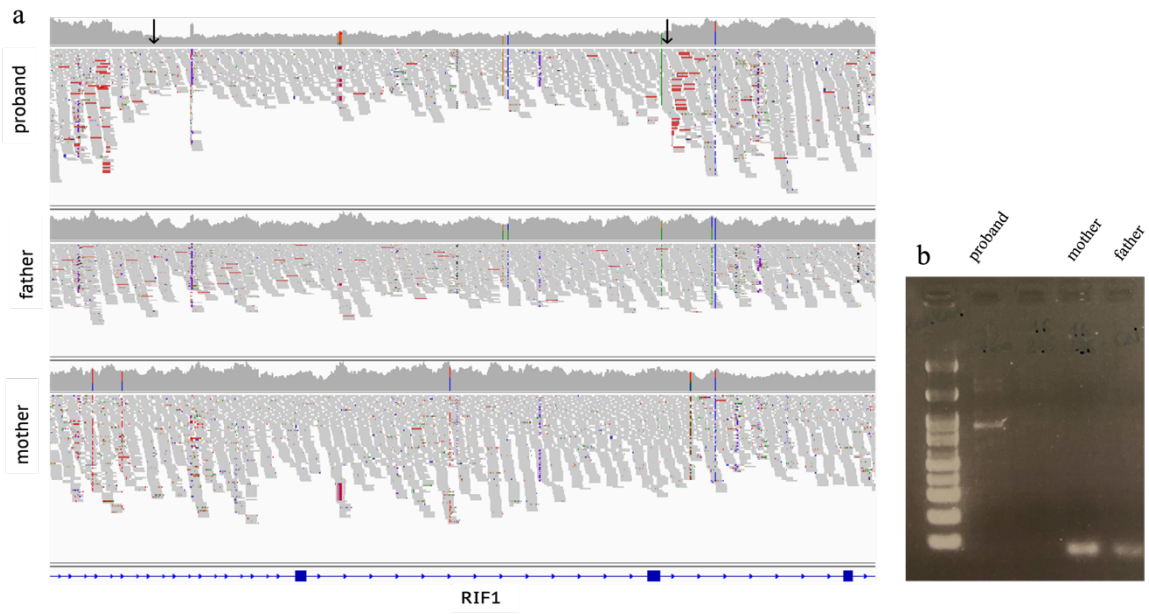


Figure 19. a. Visualization of genomic alignment of WGS data in Integrative Genomics Viewer (IGV) software. A 6 kb de novo DEL in RIF1 on 2q23(151426560-151433380) encompassing exon 8 of the proband. b. PCR results confirm that the deleted region amplified only in proband. The primers were designed using Primer3 (forward: GGAATCATGGTAGGTTCATTTCCCACAAG, reverse: GTTTGTGTGCCTCTGATTCAAAG).

FID_7. D*e novo* p.Arg1330Trp in the exon 23 of HECW2 (NM_001348768.2) are identified in proband of FID_7. HECW2 encodes an E3 ubiquitin ligase that plays an important role in neural crest cell proliferation, migration, and differentiation. It is mainly expressed in brain, lung, and heart tissues. Recently, variants in HECW2 have been reported to cause a neurodevelopmental disorder with hypotonia, seizures, and impaired language. In particular, the missense variant p.Arg1330Trp has been reported as a recurrent variant and accounts for 20% of reported cases. The variant is reported as pathogenic in ClinVar and is not present in the general population.

FID_8. A heterozygous missense variant in exon 20 of AFF4 (p.Arg258Trp) was identified in the proband from FID_8. This missense variant occurs as DNMs, was reported as pathogenic in ClinVar, and is absent in Gnomad. The p.Arg258Trp (NM_014423.4) is the most common recurrent mutation in CHOPS syndrome. CHOPS is an abbreviation for a list of features of the disorder, including cognitive impairment, coarse facial features, heart defects, obesity, lung involvement, short stature, and skeletal abnormalities. Previous studies have shown that all DNMs associated with CHOPS syndrome are located in the evolutionarily highly conserved ALF homology domain of AFF4 (Raible et al., 2019). Although the p.Arg258Trp is located in the exome region, the clinical WES had a negative result because the gene is not included in the target region. Based on those evidence, the clinicians make a diagnosis of CHOPS in the proband.

FID_10. Compound heterozygous in the TRIT1 gene was found in the proband and the affected brother in FID_10: a missense variant (p.Trp228Arg, NR_132405.1) inherited from the mother and a stop variant (p.Arg150Ter, NR_132405.1) inherited from the father. The TRIT1 gene encodes a tRNA isopentenyl transferase (ITPase) that is involved in the post-transcriptional modification of tRNAs and is essential for folding, stability, and maintenance of the correct reading frame during protein translation. TRIT1 defect is a rare autosomal recessive transcription disorder associated with developmental delay, myoclonic seizures, delayed myelination, and abnormalities of the corpus callosum. To date, only 13 patients are known (Muylle et al., 2022). The p.Trp228Arg missense variant identified in the samples has an allele frequency of 0.00000657 in GnomAD, and individuals homozygous for the alternate variant are not present in the database. The variant is classified as pathogenic according to ACMG guidelines and is not listed in ClinVar. The variant p.Arg150Ter is not present in the GnomAD database and is classified as likely pathogenic according to ACMG guidelines and is not reported in ClinVar.

FID_22. Small *de novo* indel (c.2631_2632del(p.Asp877GlufsTer88), NM_001037333.3) in exon24 of the gene CYFIP2 was found in the proband of FID_22. CYFIP2 encodes a component of the regulatory complex of the WASP -family of verprolin-homologous proteins (WAVE). Proteins of the WAVE -family play a central role in actin remodeling, axon

elongation, dendritic spine morphogenesis, and synaptic plasticity. De novo missense and LoF variants in CYFIP2 have recently been associated with early-onset epileptic encephalopathy, intellectual disability (ID), seizures, and muscular hypotonia (Nakashima et al., 2018). The frameshift variant of CYFIP2 has not yet been described but is predicted to be intolerant (pLI = 1) to both missense and LoF variants. The frameshift variant p.Asp877GlufsTer88 does not occur in the general population, is not described in ClinVar, and is predicted to be LP based on ACMG recommendation.

4.9 Candidate diagnosis

FID_5. In the proband FID_5, an intronic DNM (c.5210+1235A > G, NM_017780.4) was found in the CHD7 genes. The variant was prioritized because it is not present in the general population and the HPO terms suggested involvement of this gene. Since childhood, CHARGE syndrome was the main clinical suspect for the FID_5 proband. The CHARGE syndrome is a rare congenital disorder characterized by multiple abnormalities, including coloboma of the eye, cardiac defects, atresia of the choans, growth and/or developmental delay, genitourinary defects, and ear abnormalities with or without deafness (Qin et al., 2020). Nowadays, clinical diagnostic criteria for CHARGE syndrome include other clinical features such as ID, cranial nerve abnormalities, square face, and others. More than 1,000 variants of CHD7 have been identified, and 90-95% of patients carrying a CHD7 variant meet the diagnostic criteria for CHARGE syndrome. The CHD7 variant found in the proband FID_5 is absent in GnomAD, is not listed in ClinVar, and is located in a deep intronic region. Because of the high clinical interest in the gene, we used the intron tool RegSNPs to predict the disease-causing probability of human intronic SNVs and obtained a probability of 75% (Lin et al., 2019). To evaluate the potential pathogenicity of the intronic mutation, a functional study is currently being performed in our laboratory. Beyond this intronic variant, no other candidate variants were identified in this sample with our pipelines.

FID_9. A heterozygous variant was found in intron 5 of LMBR1 (c.424-5999T > G, NR_146959.1) in three affected samples of FID_9, one male fetus and two females. This gene

encodes a member of the LMBR1-like membrane protein family, and intron 5 of this gene contains a highly conserved, cis-acting regulatory module for the sonic hedgehog gene (SHH). Expression of SHH is regulated by an enhancer called zone of polarizing activity regulatory sequence (ZRS) and located within intron 5 of the LMBR1 gene 1 Mb from the target gene SHH. Previous study showed that mis regulation of SHH and mutations in intron 5 of LMBR1 cause several disorders, including Laurin-Sandrow syndrome, which is highly suspicious for this family (Lettice et al., 2003) (Xu et al., 2020). The intron variant found in this family is outside the ZRS region and seems to be located in an enhancer (Figure 20). The variant is not present in Gnomad, not reported in ClinVar, and given the strong relevance of this genomic region for these disease phenotypes, a functional study is planned to test the effects of the variant on enhancer activity. Segregation study showed that the same variant is also present in 4,I (unaffected). The high variable expressivity and incomplete penetrance have been described in ZRS-associated syndromes (Vanlerberghe et al., 2014) (Girisha et al., 2014) (Norbnop et al., 2014). It is reported that the complex interaction between SHH and its regulatory elements in limb development could be an explanation for the observed variability. However, further studies are needed to define the association between the intronic variant in LMBR1 and the family phenotype.

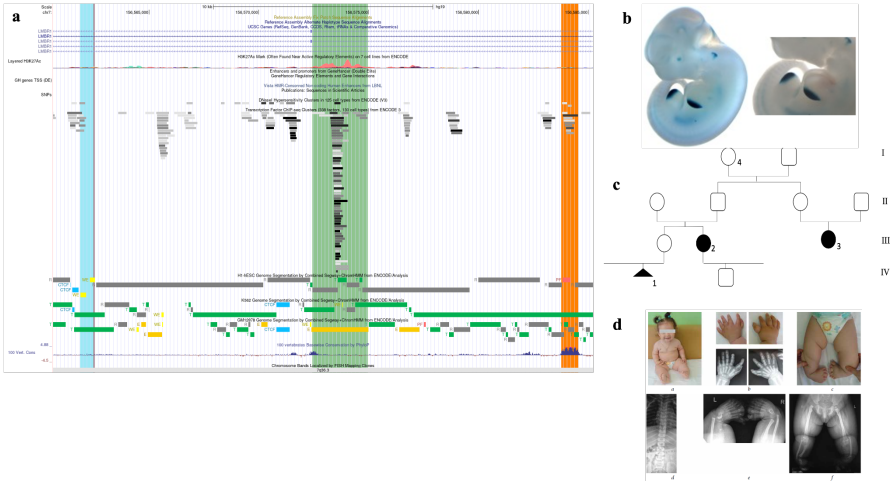


Figure 20. a) Non-coding mutation in LMBR1 (red line) located in a region predicted to contain a weak enhancer. In intron 5, the ZRS limb enhancer is located at about 21 kb (orange box). b) The activity of the human ZRS enhancer has been tested in mouse embryos, and the activity of the enhancer is critical for normal limb development in mice (Kvon et al Cell 2016). c) Pedigree of FID_9: In affected proband IV,1 DNA was

extracted from chorionic villi and the clinical features are tibial agenesis, polydactyly (8 fingers), syndactyly, ectopic kidney, dysmorphism. Affected sample III,2 had tibial agenesis and polydactyly. Affected sample III,3 had holoprosencephaly. The segregation study identified the same variant in sample IV,4 which has no clinical features. d) Clinical manifestation of ZRS-associated syndromes

FID_11. A large *de novo* pericentric inversion was identified by *nx-pipeline* and prioritized. This INV was also previously detected by conventional karyotypes. The large inversion, approximately 130 Mb in size, involves most of chromosome X (p21.3q27) in the proband of FID_11. The junctions were supported by the signal PR with a STRONG score and confirmed by orthogonal experimental approaches (PCR and Sanger, Figure 21). The breakpoints identified in the *nx-pipeline* are located 4 bases upstream of the true junctions identified by Sanger. The INV is located in the intergenic region and does not directly affect genes. *Denovo* INVs occur at a frequency of 1%-2% in the general population, and further studies, such as X chromosome inactivation, are needed to define molecular significance (Pettersson et al., 2020). In addition, FID_11 carries a missense DNM in GCH1 (p.Thr94Met) that is not present in the general population and is associated with Dystonia 5 in the ClinVar database, which does not explain ID in this patient (Yu et al., 2013).

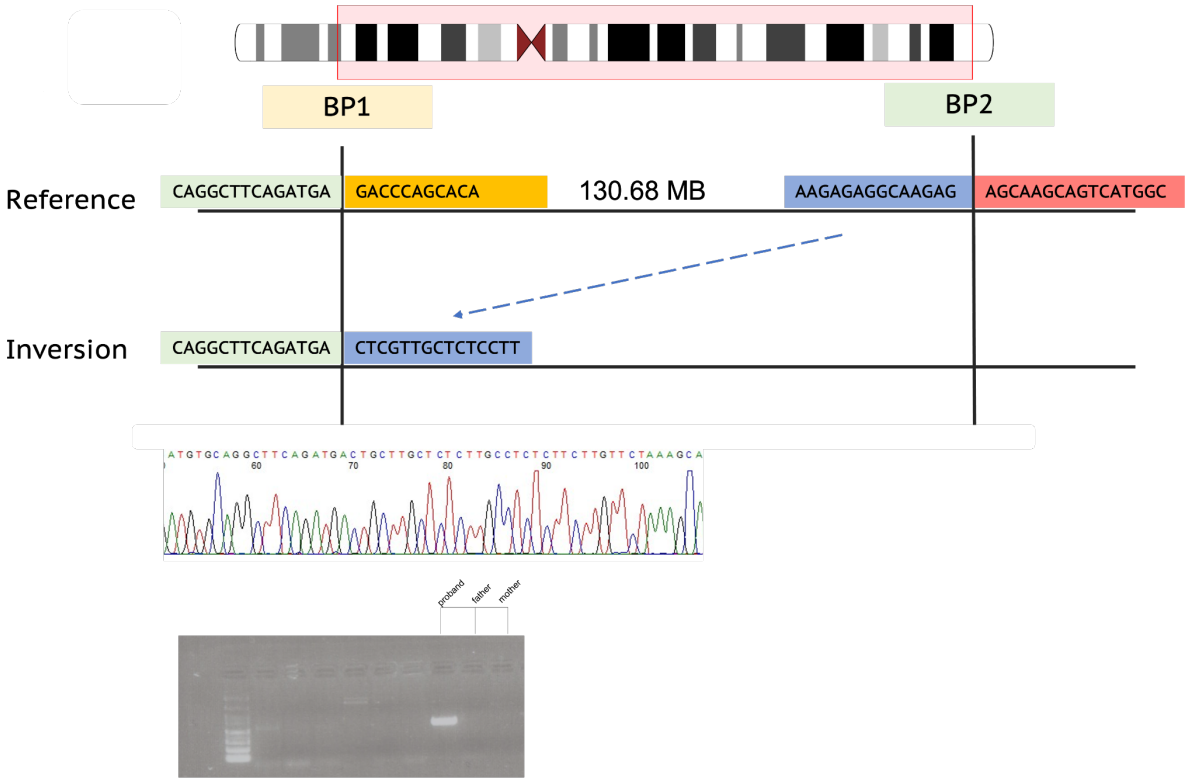


Figure 21. **Large *de novo* paracentric inversion on chromosome X**. Ideogram with chromosome X highlighting the inverted representation by the red block. Horizontal lines represent the reference (Reference) and inverted (Inversion) rearrangements. Vertical arrows mark the two inversion breakpoints (BP1 and BP2). Sequences involved in the rearrangement are shown in colored boxes. Sanger chromatograms show the sequence of the first breakpoints in the Inv chromosome. PCR results confirm that the inverted regions are amplified only in the proband and not in the parents. The primers were designed to amplify only in the presence of an inversion and along the boundary of the first breakpoint. The primers were designed using Primer3 (forward: AGCAGTGATACCCAGACAGT, reverse: GCTCTTACTGCTTCAGGGTTC) (primer3.ut.e).

FID_19. A dnSV in NLGN4X was found in proband of FID_19. The INS is located in the intron of NLGN4X, a gene associated with ASD/ID. NLGN4X belongs to neuroligin (NLGN), a family of postsynaptic cell adhesion molecules that regulate neuronal development and synaptic transmission. NLGN have different isoforms and distinct localizations, with human NLGN4X is localized in excitatory synapses. Among clinical syndromes reported by OMIM, mutations in NLGN4X cause seizures in approximately 30% of cases, the most important clinical feature reported in this patient. The INS was confirmed to be de novo by PCR. However, the functional impact of this mutation and its association with phenotype are not yet clear, and functional studies are needed to confirm pathogenicity.

FID_24. Compound heterozygous in the RFT1 gene in the proband of FID_24 : a missense variant (p.Lys152Glu, NM_052859.4) inherited from the father and a noncoding variant inherited from the mother (c.*2407C > T, NM_052859.4). RFT1 RFT1-congenital disorder of glycosylation (CDG) syndrome is a recessive N-glycosylation disorder associated with developmental delay and nonspecific epilepsy. In general, impaired glycosylation is associated with a wide spectrum of clinical manifestations, ranging from paucisymptomatic individuals to extremely severe phenotypes with multiorgan involvement. The wide variability of phenotypes makes CDG challenge to diagnose. Confirmation of CDG relies on enzymatic assays in leukocytes or fibroblasts and lipid-linked oligosaccharide (LLO) analysis in fibroblasts, as well as serum sialotransferrin dosing (Barba et al., 2016). The c.454A>G (p.Lys152Glu) was previously described in patients with CDG, occurs at a frequency of 0.0001% in the general population, and was included in ClinVar as a pathogenic variant, whereas the 3'UTR variant was absent in the general population and was not included in any other database (Vleugels et al., 2009). This heterozygous compound may explain part of the phenotype but not precocious puberty and secondary amenorrhea. In contrast, this phenotype could be explained by a missense variant in BRWD1 (p.Arg219His) inherited from the father and associated with premature ovarian failure. Biochemical analyses are underway to confirm the disruption of glycosylation.

Discussion

In this work, we aimed to evaluate the clinical application of WGS in a diagnostic cohort of patients with NDD in terms of both the computational framework required to process WGS data and the diagnostic yield. Our data support the hypothesis that the development of a workflow system with robust bioinformatics pipelines for genomic data processing, calling and prioritization support the clinical application of WGS in diagnostic practice. The overall diagnostic yield achieved with WGS is 32% with an additional 20% of potential diagnosis. Genetic tests such as WES and clinical WES have been previously performed in all patients of our NDD cohort, which gave negative results. However, some of the variants, especially located in the coding regions, could also be identified through a reanalysis of WES data (Table 11).

Gene	Transcript	Variant	Inheritance	Outcome	Detectable with other NGS technique
CREBBP	NM_004380.3	p.Arg1868Trp	AD (de novo)	1	Potentially
TRIT1	NR_132405.1	p.Trp228Arg, p.Arg150Ter	AR	1	Potentially
HECW2	NM_001348768.2	p.Arg1330Trp	AD (denovo)	1	Potentially
AFF4	NM_014423.4	p.Arg258Trp	AD (denovo)	1	Potentially
CYFIP2	NM_001037333.3	p.Asp877GlufsTer57	AD (denovo)	1	Potentially
MT-ATP8	ENST00000361851.1	p.Lys57*	Mitochondrial	1	Unlikely
RIF1	NM_018151.5	2-151426560-6821_del	AD (denovo)	1	Unlikely
MED12	NM_005120.3	c.4477_4527+56dup	AD (denovo)	1	Unlikely
CHD7	NM_017780.4	c.5210+1235A > G	AD (denovo	2	Unlikely
LMBR1	NR_146959.1	c.424-5999T>G	AD	2	Unlikely
-	-	inv(X)(p22.13q28)	AD	2	Unlikely
NLGN4X	ENST00000381095.8	X-6197436+308_ins	XRL	2	Unlikely
RFT1	NM_052859.4	p.Lys152Glu, C.*2407C > T	AR	2	Unlikely

Table 11. **Clinical findings in the NDD**. The table shows that 5 of the 7 variants that led to a certain diagnosis are in the coding region and are therefore potentially also identifiable via WES. In contrast, none of the variants that were flagged as "possible diagnoses" could be identified via WES.

The de novo pathogenic variant (p.Arg1868Trp) affecting exon 31 of the CREBBP was previously described in patients with Menke-Hennekam syndrome. The first description of the syndrome dates from 2016, whereas the WES of the proband of FID_1 was performed in 2015, a year before the syndrome was published, when variants in the CREBBP gene were causative of Rubinstein-Taybi syndrome, which was not consistent with the patient's phenotype.

The de novo pathogenic variants (p.Arg1330Trp) in the HECW2 gene and (Arg258Trp) in the AFF4 gene identified by WGS were also previously described in patients with neurodevelopmental disorder and CHOPS syndrome, respectively. These protein-coding variants had not been previously identified by clinical WES because the genes were not captured by the capture kit. In the proband FID _22, a small de novo indel (c.2631_2632del(p.Asp877GlufsTer88) was found in the CYFIP2 gene exon 24. Missense and loss-of-function variants in CYFIP2 were found in patients with a developmental and epileptic encephalopathy phenotype (Begemann A et al. al; 2020). Although the WES enrichment kit included the CYFIP2 gene, the variant was not included due to the low quality of variant the WES data. Analysis of mitochondrial DNA allowed identification of the m.8535A > G (Lys57*) variant of the MT -ATP8 gene with a high percentage of heteroplasmy in the different tissues analyzed (muscle, blood, and urine) in the proband from FID _2. This variant m.8535A > G in the MT -ATP8 gene has never been described in the literature, is not present in the gnomAD population database, functional analyzes confirmed its pathogenicity. WES had not allowed the identification of the studied variant because of insufficient coverage of mitochondrial DNA. The 6,821 kb de novo deletion (2-151426560-6821_del) in the RIF1 gene has never been described in the literature and is not present in gnomadSV. Frameshift variants in RIF1 were observed in two patients with neurodevelopmental delayed, delayed gross motor development and intellectual disability (Seaby et al.; 2022). WGS is the best tool for identifying structural variants, and since the deletion breakpoints are in introns 8 and 9, WES is unable to detect this type of variant. The de novo insertion in the MED12 gene was also

detected in the FID _4 proband using WGS. Analysis of the WES yielded a negative result and although the first breakpoint is within the exon, the variant showed a poor quality, and the profile of the insertion was difficult to distinguish (Figure 22).

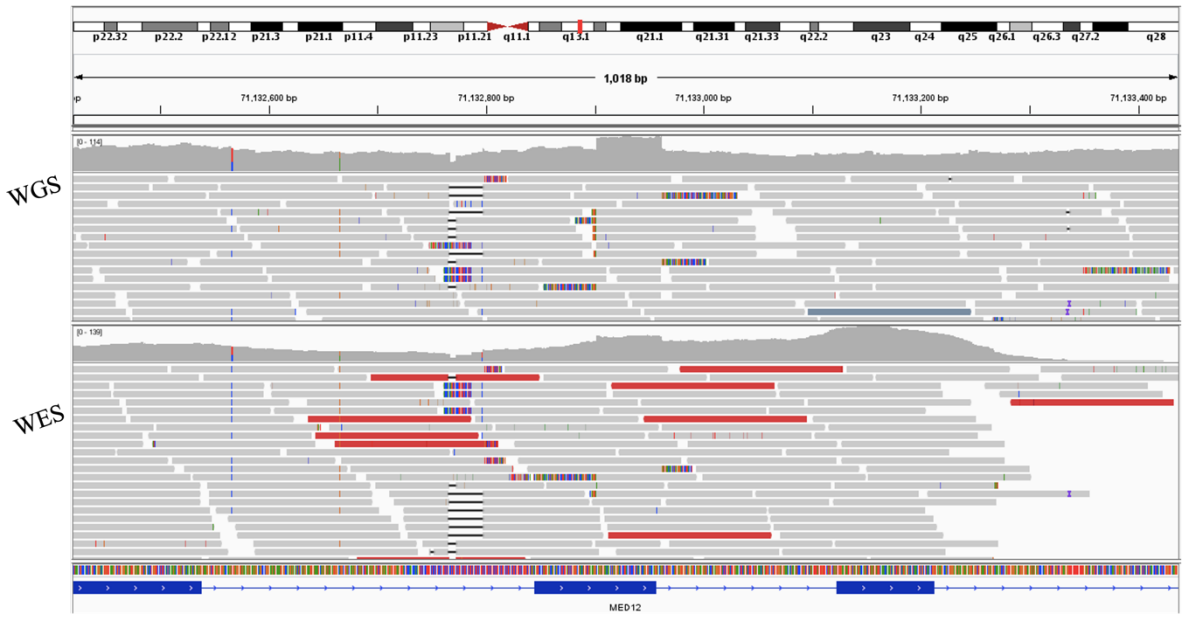


Figure 22. **Difference between WES and WGS in c.4477_4527+56dup identification.** IGV visualization of MED12 insertion in WGS and WES data. In the WGS data, the insertion is well visible, and the breakpoints are distinguishable. In WES the insertion is not detectable.

All variants labeled as possible diagnosis (outcome = 2) are detectable only by using the WGS. WGS is able to identify all types of relevant genomic variants in a single experimental dataset, facilitating downstream analysis. To evaluate the clinical application of WGS in NDD, the following steps were performed:

1) development of a robust bioinformatics workflow for genomic variant processing, calling and prioritization

2) integration of SV calling, genotyping and prioritization within this framework

3) evaluation of diagnostic yield.

It is well-known that the most important barrier to the adoption of WGS in clinical practice is data processing and management (Messner, 2017). The lack of standardized protocols for processing and prioritizing WGS variants prompted us to develop a standardized workflow for identifying all disease-relevant variants (i.e. SNVs/Indels, SVs, STRs) in NDD disorders: *sn-*

pipeline and *nx-pipeline*. *Sn-pipeline* is a robust framework for analyses of SNVs/Indels. The configuration was designed to include all necessary steps for processing and calling small variants. Required bioinformatics tools are automatically installed in isolated virtual environments that allow rapid pipeline setup on new systems and also reproducibility of analysis on different systems. In addition, this framework tracks tasks and errors, allowing immediate assessment of the quality of generated inputs and rapid identification of failed jobs. While small variant identification has been standardized through the use of "gold standard" tools commonly used for sequencing and variant calling, best practices for identification of SVs have yet to be defined (Liu et al., 2022). Therefore, as a second step, we proceeded to perform benchmark analyses to establish a diagnostic pipeline for the identification of SVs from genomic data. More than 80 tools are available for identifying SVs from WGS data. However, previous studies have shown that only some of them have achieved satisfactory precision and recall values on simulated and real data (Kosugi et al., 2019). Moreover, the performance of each tool varies depending on the type and size of SV. First, we evaluated the most reliable sizes based on the different types of SVs from SR WGS data in terms of precision and recall using the available tools. Our data suggest that DELs are the type of SVs identified with higher precision and recall by the vast majority of the selected tools. The most reliable size for DELs is between 150bp and 30KB. Above and below this threshold, the performance of the tools is poor (with an F-score < 0.2). The INS and DUP tools were evaluated together. Our data suggest that this type of SV is particularly difficult to identify from WGS data. However, Manta, which is also based on DA, demonstrates adequate performance values for INSs of small size (50-500bp). For INVs, the performance varies greatly depending on the size; Delly achieves high performances for large INVs (30KB and above). Our results are consistent with previous published works (Kosugi et al., 2019). Since no tool outperforms the other, the second goal was to determine how to combine them. There is no agreement regarding how to combine these tools (van Belzen et al., 2021). Consensus calls generated by multiple SV callers can achieve higher precision and recall. The union of all SVs calls across multiple tools results in a higher false positive rate, but often results in higher sensitivity estimates. Our data suggest that an approach based on the union of multiple callers gives better performance than the consensus approach (0.69 vs 0.66), especially when tools based on the same detection mechanisms are combined (Manta, Delly, Lumpy). Also, using a consensus

approach, the TP INS /DUP calls called by Manta and TP INVs calls called preferentially by Delly would be lost. To handle on caller-derived FP, we decided to guide the downstream analyses by prioritizing the most reliable SVs by modeling a score based on the benchmark analysis results. The SVs with the most reliable size called by multiple tools were flagged as highly reliable in the downstream analyses. Because false positives occur in all samples in regions of low genome complexity, the allelic frequency of the variant in the cohort could guide filtering (Sedlazeck et al., 2018). Due to the high-powered features of the underlying Nextflow engine, the execution of the pipeline is flexible and scalable for multiple samples and execution environments. Despite our best efforts, we are aware that a certain percentage of false positive calls remain in our final output data. Accurate and reproducible detection of SV supports reliable clinical implementation and reduces the possibility of misdiagnosis. However, the tools currently available do not achieve the accuracy achieved with small variants. Research continues, and new methods based on machine learning may lead to improvements in SV detection.

A potential limitation of our pipeline approach is computational time, which depends on the system CPU and remains a major bottleneck for clinical application of WGS. Processing WGS data required approximately 25 hours per genome (20 threads), resulting in a total of approximately 668 hours to process data from the entire cohort (77 samples). We then compared the computational time spent on the DRAGEN instance and found that it decreased dramatically. In this scenario, DRAGEN can replace the most expensive computational task in a CPU system, such as alignment and quality control. In contrast, variant calling could be performed for both SNVs/Indels and SVs with workflow managers.

A final aim of this work is evaluating the diagnostic yield of WGS in NDD. In the last decade, advances in genomic technologies and analyses have enormously increased our capability to identify variants and genes involved in NDD. In recent years, WGS has emerged as the most appropriate tool for the discovery of pathogenic variants in rare diseases, allowing comprehensive variant calling of all types of variations without the technical biases and limitations seen in other sequencing technologies (WES, TS) (Lionel et al., 2018). The overall diagnostic yield in our cohort was 32 %, a value consistent with the diagnostic yield in patients with NDD (Srivastava et al., 2019) (van der Sanden et al., 2023). However, we expected a higher diagnostic yield with WGS of approximately 40-50%, a diagnostic yield that is likely

to be achieved when accounting for SVs and non-coding SNVs that required further validation (52%) (Gilissen et al., 2014) (Álvarez-Mora et al., 2022). A possible explanation for the achieved diagnostic yield could be related to the variant interpretation, which is still focused on the exonic regions. Interpretation of regions beyond the exons remains severely limited. In particular, clinical interpretation of structural rearrangements, such as inversions, deletions, or duplications, and non-coding SNV regions is not straightforward and requires functional testing, often performed in research contexts where times are longer than in routine diagnostics. Approximately 50% of patients in our cohort did not receive a molecular diagnosis or a candidate diagnosis. There are some limitations in our work that may explain this result. First, identification and interpretation of SVs remain challenging. In particular, callers fail in detection of whole range size of SVs and some FN could be related to explain the phenotype. In addition, we focused our analysis on evaluating SNVs/Indels and SVs separately. We are aware that the co-occurrence of SVs and SNVs on the same locus is a relevant genetic mechanism. A future plan of this work is to implement a Rabdomyzer module to integrate the simultaneous analysis of co-occurrence of SNVs/SVs in compound heterozygous mode, with the aim of improving the diagnosis rate. Furthermore, in this study, we limited the analysis of STRs to 30 genes, and finally, our knowledge of noncoding variants is still limited, and we plan to include additional annotations for noncoding variants in our analysis. In particular, we plan to add a comprehensive set of annotations for regulatory regions using GREEN-DB (Giacopuzzi et al., 2022).

In conclusion, WGS is going to be increasingly adopted as a primary diagnostic test for patients with NDD. WGS identified clinically relevant variants and candidate variants, suggesting that its diagnostic potential is likely underestimated. Further efforts are needed, particularly to combine different types (i.e., SNVs and SVs) of variant identified by WGS and to improve the interpretation of variants that lie beyond the boundaries of the coding regions. In the coming years, WGS will generate a large amount of data that can be used to train machine learning algorithms and provide useful tools for interpreting and classifying non-coding variants and improving the identification of SVs

Reference

100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report, 2021. . N. Engl. J. Med. 385, 1868–1880. https://doi.org/10.1056/NEJMoa2035790

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 Biol. 17, 241. https://doi.org/10.1186/s13059-016-1110-1

Ahmed, A.E., Allen, J.M., Bhat, T., Burra, P., Fliege, C.E., Hart, S.N., Heldenbrand, J.R., Hudson, M.E., Istanto, D.D., Kalmbach, M.T., Kapraun, G.D., Kendig, K.I., Kendzior, M.C., Klee, E.W., Mattson, N., Ross, C.A., Sharif, S.M., Venkatakrishnan, R., Fadlelmola, F.M., Mainzer, L.S., 2021. Design considerations for workflow management systems use in production genomics research and the clinic. Sci. Rep. 11, 21680. https://doi.org/10.1038/s41598-021-99288-8

Alfares, A., Alsubaie, L., Aloraini, T., Alaskar, A., Althagafi, A., Alahmad, A., Rashid, M., Alswaid, A., Alothaim, A., Eyaid, W., Ababneh, F., Albalwi, M., Alotaibi, R., Almutairi, M., Altharawi, N., Alsamer, A., Abdelhakim, M., Kafkas, S., Mineta, K., Cheung, N., Abdallah, A.M., Büchmann-Møller, S., Fukasawa, Y., Zhao, X., Rajan, I., Hoehndorf, R., Al Mutairi, F., Gojobori, T., Alfadhel, M., 2020. What is the right sequencing approach? Solo VS extended family analysis in consanguineous populations. BMC Med. Genomics 13, 103. https://doi.org/10.1186/s12920-020-00743-8

Antaki, D., Brandler, W.M., Sebat, J., 2018. SV2: accurate structural variation genotyping and de novo mutation detection from whole genomes. Bioinformatics 34, 1774–1777. https://doi.org/10.1093/bioinformatics/btx813

Antaki, D., Guevara, J., Maihofer, A.X., Klein, M., Gujral, M., Grove, J., Carey, C.E., Hong, O., Arranz, M.J., Hervas, A., Corsello, C., Vaux, K.K., Muotri, A.R., Iakoucheva, L.M., Courchesne, E., Pierce, K., Gleeson, J.G., Robinson, E.B., Nievergelt, C.M., Sebat, J., 2022. A phenotypic spectrum of autism is attributable to the combined effects of rare variants, polygenic risk and sex. Nat. Genet. 54, 1284–1292. https://doi.org/10.1038/s41588-022-01064-5

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., Gibbs, R.A., Green, E.D., Hurles, M.E., Knoppers, B.M., Korbel, J.O., Lander, E.S., Lee, C., Lehrach, H., Mardis, E.R., Marth, G.T., McVean, G.A., Nickerson, D.A., Schmidt, J.P., Sherry, S.T., Wang, J., Wilson, R.K., Gibbs, R.A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E.S., Altshuler, D.M., Gabriel,

S.B., Gupta, N., Gharani, N., Toji, L.H., Gerry, N.P., Resch, A.M., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S.E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W.M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R.E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D.R., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M.W., Amstislavskiy, V.S., Borodina, T.A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Mardis, E.R., Wilson, R.K., Fulton, L., Fulton, R., Sherry, S.T., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., McVean, G.A., Durbin, R.M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T.M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Schmidt, J.P., Davies, C.J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C.L., Kong, Y., Marcketta, A., Gibbs, R.A., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., Wang, J., Coin, L.J.M., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Y., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G.T., Garrison, E.P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A.N., Wu, J., Zhang, M., Daly, M.J., DePristo, M.A., Handsaker, R.E., Altshuler, D.M., Banks, E., Bhatia, G., del Angel, G., Gabriel, S.B., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E.S., McCarroll, S.A., Nemesh, J.C., Poplin, R.E., Yoon, S.C., Lihm, J., Makarov, V., Clark, A.G., Gottipati, S., Keinan, A., Rodriguez-Flores, J.L., Korbel, J.O., Rausch, T., Fritz, M.H., Stütz, A.M., Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W.M., Ritchie, G.R.S., Smith, R.E., Zerbino, D., Zheng-Bradley, X., Sabeti, P.C., Shlyakhter, I., Schaffner, S.F., Vitti, J., Cooper, D.N., Ball, E.V., Stenson, P.D., Bentley, D.R., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E.E., Batzer, M.A., Konkel, M.K., Walker, J.A., MacArthur, D.G., Lek, M., Sudbrak, R., Amstislavskiy, V.S., Herwig, R., Mardis, E.R., Ding, L., Koboldt, D.C., Larson, D., Ye, K., Gravel, S., The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, European Molecular Biology Laboratory, E.B.I., Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University, National Eye Institute, N., 2015. A global reference for human genetic variation. Nature 526, 68–74. https://doi.org/10.1038/nature15393

Banka, S., Sayer, R., Breen, C., Barton, S., Pavaine, J., Sheppard, S.E., Bedoukian, E., Skraban, C., Cuddapah, V.A., Clayton-Smith, J., 2019. Genotype–phenotype specificity in Menke–Hennekam

syndrome caused by missense variants in exon 30 or 31 of CREBBP. Am. J. Med. Genet. A. 179, 1058–1062. https://doi.org/10.1002/ajmg.a.61131

Barba, C., Darra, F., Cusmai, R., Procopio, E., Dionisi Vici, C., Keldermans, L., Vuillaumier-Barrot, S., Lefeber, D.J., Guerrini, R., Group, C.D.G., 2016. Congenital disorders of glycosylation presenting as epileptic encephalopathy with migrating partial seizures in infancy. Dev. Med. Child Neurol. 58, 1085–1091. https://doi.org/10.1111/dmcn.13141

Becker, T., Lee, W.-P., Leone, J., Zhu, Q., Zhang, C., Liu, S., Sargent, J., Shanker, K., Mil-homens, A., Cerveira, E., Ryan, M., Cha, J., Navarro, F.C.P., Galeev, T., Gerstein, M., Mills, R.E., Shin, D.-G., Lee, C., Malhotra, A., 2018. FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. Genome Biol. 19, 38. https://doi.org/10.1186/s13059-018-1404-6

Begemann A, Sticht H, Begtrup A, Vitobello A, Faivre L, Banka S, Alhaddad B, Asadollahi R, Becker J, Bierhals T, Brown KE, Bruel AL, Brunet T, Carneiro M, Cremer K, Day R, Denommé-Pichon AS, Dyment DA, Engels H, Fisher R, Goh ES, Hajianpour MJ, Haertel LRM, Hauer N, Hempel M, Herget T, Johannsen J, Kraus C, Le Guyader G, Lesca G, Mau-Them FT, McDermott JH, McWalter K, Meyer P, Õunap K, Popp B, Reimand T, Riedhammer KM, Russo M, Sadleir LG, Saenz M, Schiff M, Schuler E, Syrbe S, Van der Ven AT, Verloes A, Willems M, Zweier C, Steindl K, Zweier M, Rauch A. New insights into the clinical and molecular spectrum of the novel CYFIP2-related neurodevelopmental disorder and impairment of the WRC-mediated actin dynamics. Genet Med. 2021 Mar;23(3):543-554. doi: 10.1038/s41436-020-01011-x. Epub 2020 Nov 5. PMID: 33149277; PMCID: PMC7935717.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M.J., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M.D., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara E. Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.-D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W.,

Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, Andrew C., Pike, Alger C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, John, Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., vandeVondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, Jane, Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R., Smith, A.J., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59. https://doi.org/10.1038/nature07517

Bragin, E., Chatzimichali, E.A., Wright, C.F., Hurles, M.E., Firth, H.V., Bevan, A.P., Swaminathan, G.J., 2014. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res. 42, D993. https://doi.org/10.1093/nar/gkt937

Brandler, W.M., Antaki, D., Gujral, M., Kleiber, M.L., Whitney, J., Maile, M.S., Hong, O., Chapman, T.R., Tan, S., Tandon, P., Pang, T., Tang, S.C., Vaux, K.K., Yang, Y., Harrington, E., Juul, S., Turner, D.J., Thiruvahindrapuram, B., Kaur, G., Wang, Z., Kingsmore, S.F., Gleeson, J.G., Bisson, D., Kakaradov, B., Telenti, A., Venter, J.C., Corominas, R., Toma, C., Cormand, B., Rueda, I., Guijarro, S., Messer, K.S., Nievergelt, C.M., Arranz, M.J., Courchesne, E., Pierce, K., Muotri, A.R., Iakoucheva, L.M., Hervas, A., Scherer, S.W., Corsello, C., Sebat, J., 2018. Paternally inherited cis-regulatory structural variants are associated with autism. Science 360, 327–331. https://doi.org/10.1126/science.aan2261

Cameron, D.L., Di Stefano, L., Papenfuss, A.T., 2019. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat. Commun. 10, 3240. https://doi.org/10.1038/s41467-019-11146-4

Cameron, D.L., Schröder, J., Penington, J.S., Do, H., Molania, R., Dobrovic, A., Speed, T.P., Papenfuss, A.T., 2017. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 27, 2050–2060. https://doi.org/10.1101/gr.222109.117

Carpi, G., Gorenstein, L., Harkins, T.T., Samadi, M., Vats, P., 2022. A GPU-accelerated compute framework for pathogen genomic variant identification to aid genomic epidemiology of infectious disease: a malaria case study. Brief. Bioinform. 23, bbac314. https://doi.org/10.1093/bib/bbac314

Caspar, S. m., Dubacher, N., Kopps, A. m., Meienberg, J., Henggeler, C., Matyas, G., 2018. Clinical sequencing: From raw data to diagnosis with lifetime value. Clin. Genet. 93, 508–519. https://doi.org/10.1111/cge.13190

Chander, V., Gibbs, R.A., Sedlazeck, F.J., 2019. Evaluation of computational genotyping of structural variation for clinical diagnoses. GigaScience 8, giz110. https://doi.org/10.1093/gigascience/giz110

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., Saunders, C.T., 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32, 1220–1222. https://doi.org/10.1093/bioinformatics/btv710

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., Sharpe, T., Stone, M.R., Valkanas, E., Fu, J., Tiao, G., Laricchia, K.M., Ruano-Rubio, V., Stevens, C., Gupta, N., Cusick, C., Margolin, L., Taylor, K.D., Lin, H.J., Rich, S.S., Post, W.S., Chen, Y.-D.I., Rotter, J.I., Nusbaum, C., Philippakis, A., Lander, E., Gabriel, S., Neale, B.M., Kathiresan, S., Daly, M.J., Banks, E., MacArthur, D.G., Talkowski, M.E., 2020. A structural variation reference for medical and population genetics. Nature 581, 444–451. https://doi.org/10.1038/s41586-020-2287-8

Dautant, A., Meier, T., Hahn, A., Tribouillard-Tanvier, D., di Rago, J.-P., Kucharczyk, R., 2018. ATP Synthase Diseases of Mitochondrial Genetic Origin. Front. Physiol. 9.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., Onate, K.C., Graham, K., Miyasato, S.R., Dreszer, T.R., Strattan, J.S., Jolanki, O., Tanaka, F.Y., Cherry, J.M., 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 46, D794–D801. https://doi.org/10.1093/nar/gkx1081

Delaney, S.K., Hultner, M.L., Jacob, H.J., Ledbetter, D.H., McCarthy, J.J., Ball, M., Beckman, K.B., Belmont, J.W., Bloss, C.S., Christman, M.F., Cosgrove, A., Damiani, S.A., Danis, T., Delledonne, M., Dougherty, M.J., Dudley, J.T., Faucett, W.A., Friedman, J.R., Haase, D.H., Hays, T.S., Heilsberg, S., Huber, J., Kaminsky, L., Ledbetter, N., Lee, W.H., Levin, E., Libiger, O., Linderman, M., Love, R.L., Magnus, D.C., Martland, A., McClure, S.L., Megill, S.E., Messier, H., Nussbaum, R.L., Palaniappan, L., Patay, B.A., Popovich, B.W., Quackenbush, J., Savant, M.J., Su, M.M., Terry, S.F., Tucker, S., Wong, W.T., Green, R.C., 2016. Toward clinical genomics in everyday medicine: perspectives and recommendations. Expert Rev. Mol. Diagn. 16, 521–532. https://doi.org/10.1586/14737159.2016.1146593

D'haene, E., Vergult, S., 2021. Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. Genet. Med. 23, 34–46. https://doi.org/10.1038/s41436-020-00974-1

Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., Notredame, C., 2017. Nextflow enables reproducible computational workflows. Nat. Biotechnol. 35, 316–319. https://doi.org/10.1038/nbt.3820

Dillon, O.J., Lunke, S., Stark, Z., Yeung, A., Thorne, N., Gaff, C., White, S.M., Tan, T.Y., 2018. Exome sequencing has higher diagnostic yield compared to simulated disease-specific panels in children with suspected monogenic disorders. Eur. J. Hum. Genet. 26, 644–651. https://doi.org/10.1038/s41431-018-0099-1

Dubois, F., Sidiropoulos, N., Weischenfeldt, J., Beroukhim, R., 2022. Structural variations in cancer and the 3D genome. Nat. Rev. Cancer 22, 533–546. https://doi.org/10.1038/s41568-022-00488-9

Eggertsson, H.P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K.E., Jonasdottir, Aslaug, Jonasdottir, Adalbjorg, Jonsdottir, I., Gudbjartsson, D.F., Melsted, P., Stefansson, K., Halldorsson, B.V., 2017. Graphtyper enables population-scale genotyping using pangenome graphs. Nat. Genet. 49, 1654–1660. https://doi.org/10.1038/ng.3964

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., Richardson, E., Singer-Berk, M., Taylor, J.C., Williams, M., Wood, J.C., Wright, C.F., Harrison, S.M., Whiffin, N., 2022. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. Genome Med. 14, 73. https://doi.org/10.1186/s13073-022-01073-3

Escaramís, G., Docampo, E., Rabionet, R., 2015. A decade of structural variants: description, history and methods to detect structural variation. Brief. Funct. Genomics 14, 305–314. https://doi.org/10.1093/bfgp/elv014

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

Giacopuzzi, E., Popitsch, N., Taylor, J.C., 2022. GREEN-DB: a framework for the annotation and prioritization of non-coding regulatory variants from whole-genome sequencing data. Nucleic Acids Res. 50, 2522–2535. https://doi.org/10.1093/nar/gkac130

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., Vissers, L.E.L.M., Veltman, J.A., 2014. Genome sequencing identifies major causes of severe intellectual disability. Nature 511, 344–347. https://doi.org/10.1038/nature13394

Girisha KM, Bidchol AM, Kamath PS, Shah KH, Mortier GR, Mundlos S, Shah H. A novel mutation (g.106737G>T) in zone of polarizing activity regulatory sequence (ZRS) causes variable limb phenotypes in Werner mesomelia. Am J Med Genet A. 2014 Apr;164A(4):898-906. doi: 10.1002/ajmg.a.36367. Epub 2014 Jan 29. PMID: 24478176.

Gong, T., Hayes, V.M., Chan, E.K.F., 2021. Detection of somatic structural variants from short-read next-generation sequencing data. Brief. Bioinform. 22, bbaa056. https://doi.org/10.1093/bib/bbaa056

Gurbich, T.A., Ilinsky, V.V., 2020. ClassifyCNV: a tool for clinical annotation of copy-number variants. Sci. Rep. 10, 20375. https://doi.org/10.1038/s41598-020-76425-3

Ham, T.J., Bruns-Smith, D., Sweeney, B., Lee, Y., Seo, S.H., Song, U.G., Oh, Y.H., Asanovic, K., Lee, J.W., Wills, L.W., 2020. Genesis: A Hardware Acceleration Framework for Genomic Data Analysis, in: 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA). Presented at the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), pp. 254–267. https://doi.org/10.1109/ISCA45697.2020.00031

Ho, S.S., Urban, A.E., Mills, R.E., 2020. Structural variation in the sequencing era. Nat. Rev. Genet. 21, 171–189. https://doi.org/10.1038/s41576-019-0180-9

Hon, C.-C., Ramilowski, J.A., Harshbarger, J., Bertin, N., Rackham, O.J.L., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T.M., Severin, J., Lizio, M., Kawaji, H., Kasukawa, T., Itoh, M., Burroughs, A.M., Noma, S., Djebali, S., Alam, T., Medvedeva, Y.A., Testa, A.C., Lipovich, L., Yip, C.-W., Abugessaisa, I., Mendez, M., Hasegawa, A., Tang, D., Lassmann, T., Heutink, P., Babina, M., Wells, C.A., Kojima, S., Nakamura, Y., Suzuki, H., Daub, C.O., de Hoon, M.J.L., Arner, E., Hayashizaki, Y., Carninci, P., Forrest, A.R.R., 2017. An atlas of human long non-coding RNAs with accurate 5′ ends. Nature 543, 199–204. https://doi.org/10.1038/nature21374

Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., Sedlazeck, F.J., 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat. Commun. 8, 14061. https://doi.org/10.1038/ncomms14061

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., Reijnders, M.R.F., Yeung, A., Yntema, H.G., Vissers, L.E.L.M., Juusola, J., Wright, C.F., Brunner, H.G., Firth, H.V., FitzPatrick, D.R., Barrett, J.C., Hurles, M.E., Gilissen, C., Retterer, K., 2020. Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature 586, 757–762. https://doi.org/10.1038/s41586-020-2832-5

Koboldt, D.C., 2020. Best practices for variant calling in clinical sequencing. Genome Med. 12, 91. https://doi.org/10.1186/s13073-020-00791-w

Kochinke, K., Zweier, C., Nijhof, B., Fenckova, M., Cizek, P., Honti, F., Keerthikumar, S., Oortveld, M.A.W., Kleefstra, T., Kramer, J.M., Webber, C., Huynen, M.A., Schenck, A., 2016. Systematic Phenomics Analysis Deconvolutes Genes Mutated in Intellectual Disability into Biologically Coherent Modules. Am. J. Hum. Genet. 98, 149–164. https://doi.org/10.1016/j.ajhg.2015.11.024

Kohailan, M., Aamer, W., Syed, N., Padmajeya, S., Hussein, S., Sayed, A., Janardhanan, J., Palaniswamy, S., El hajj, N., Al-Shabeeb Akil, A., Fakhro, K.A., 2022. Patterns and distribution of de novo mutations in multiplex Middle Eastern families. J. Hum. Genet. 67, 579–588. https://doi.org/10.1038/s10038-022-01054-9

Koster, J., Rahmann, S., 2012. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics 28, 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

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

Krude, H., Mundlos, S., Øien, N.C., Opitz, R., Schuelke, M., 2021. What can go wrong in the non-coding genome and how to interpret whole genome sequencing data. Med. Genet. 33, 121–131. https://doi.org/10.1515/medgen-2021-2071

Kurtzer, G.M., Sochat, V., Bauer, M.W., 2017. Singularity: Scientific containers for mobility of compute. PLOS ONE 12, e0177459. https://doi.org/10.1371/journal.pone.0177459

Lalonde, E., Rentas, S., Lin, F., Dulik, M.C., Skraban, C.M., Spinner, N.B., 2020. Genomic Diagnosis for Pediatric Disorders: Revolution and Evolution. Front. Pediatr. 8.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J.B., Kattman, B.L., Maglott, D.R., 2018. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 46, D1062–D1067. https://doi.org/10.1093/nar/gkx1153

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

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. Hum. Mutat. 36, 815–822. https://doi.org/10.1002/humu.22813

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., de Graaff, E., 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. 12, 1725–1735. https://doi.org/10.1093/hmg/ddg180

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 25, 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., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lieberman-Aiden E., van Berkum N.L., Williams L., Imakaev M., Ragoczy T., Telling A., Amit I., Lajoie B.R., Sabo P.J., Dorschner M.O. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–293

Lin, B., Hui, J., Mao, H., 2021. Nanopore Technology and Its Applications in Gene Sequencing. Biosensors 11, 214. https://doi.org/10.3390/bios11070214

Lin, H., Hargreaves, K.A., Li, R., Reiter, J.L., Wang, Y., Mort, M., Cooper, D.N., Zhou, Y., Zhang, C., Eadon, M.T., Dolan, M.E., Ipe, J., Skaar, T.C., Liu, Y., 2019. RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants. Genome Biol. 20, 254. https://doi.org/10.1186/s13059-019-1847-4

Lin, M., Whitmire, S., Chen, J., Farrel, A., Shi, X., Guo, J., 2017. Effects of short indels on protein structure and function in human genomes. Sci. Rep. 7, 9313. https://doi.org/10.1038/s41598-017-09287-x

Lin, X., Yang, Y., Melton, P.E., Singh, V., Simpson-Yap, S., Burdon, K.P., Taylor, B.V., Zhou, Y., 2022. Integrating Genetic Structural Variations and Whole-Genome Sequencing Into Clinical Neurology. Neurol. Genet. 8. https://doi.org/10.1212/NXG.0000000000200005

Lionel, A.C., Costain, G., Monfared, N., Walker, S., Reuter, M.S., Hosseini, S.M., Thiruvahindrapuram, B., Merico, D., Jobling, R., Nalpathamkalam, T., Pellecchia, G., Sung, W.W.L., Wang, Z., Bikangaga, P., Boelman, C., Carter, M.T., Cordeiro, D., Cytrynbaum, C., Dell, S.D., Dhir, P., Dowling, J.J., Heon, E., Hewson, S., Hiraki, L., Inbar-Feigenberg, M., Klatt, R., Kronick, J., Laxer, R.M., Licht, C., MacDonald, H., Mercimek-Andrews, S., Mendoza-Londono, R., Piscione, T., Schneider, R., Schulze, A., Silverman, E., Siriwardena, K., Snead, O.C., Sondheimer, N., Sutherland, J., Vincent, A., Wasserman, J.D., Weksberg, R., Shuman, C., Carew, C., Szego, M.J., Hayeems, R.Z., Basran, R., Stavropoulos, D.J., Ray, P.N., Bowdin, S., Meyn, M.S., Cohn, R.D., Scherer, S.W., Marshall, C.R., 2018. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet. Med. 20, 435–443. https://doi.org/10.1038/gim.2017.119

Liu, Z., Roberts, R., Mercer, T.R., Xu, J., Sedlazeck, F.J., Tong, W., 2022. Towards accurate and reliable resolution of structural variants for clinical diagnosis. Genome Biol. 23, 68. https://doi.org/10.1186/s13059-022-02636-8

Logsdon, G.A., Vollger, M.R., Eichler, E.E., 2020. Long-read human genome sequencing and its applications. Nat. Rev. Genet. 21, 597–614. https://doi.org/10.1038/s41576-020-0236-x

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N.J., Nicolae, D.L., Gamazon, E.R., Im, H.K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E.T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J.M., Wilder, E.L., Derr, L.K., Green, E.D., Struewing, J.P., Temple, G., Volpi, S., Boyer, J.T., Thomson, E.J., Guyer, M.S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T.R., Koester, S.E., Little, A.R., Bender, P.K., Lehner, T., Yao, Y., Compton, C.C., Vaught, J.B., Sawyer, S., Lockhart, N.C., Demchok, J., Moore, H.F., 2013. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45, 580–585. https://doi.org/10.1038/ng.2653

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. Genet. Med. 18, 1282–1289. https://doi.org/10.1038/gim.2016.58

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., Cunningham, F., 2016. The Ensembl Variant Effect Predictor. Genome Biol. 17, 122. https://doi.org/10.1186/s13059-016-0974-4

Melo US, Schöpflin R, Acuna-Hidalgo R, Mensah MA, Fischer-Zirnsak B, Holtgrewe M, Klever MK, Türkmen S, Heinrich V, Pluym ID, Matoso E, Bernardo de Sousa S, Louro P, Hülsemann W, Cohen M, Dufke A, Latos-Bieleńska A, Vingron M, Kalscheuer V, Quintero-Rivera F, Spielmann M, Mundlos S. Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases. Am J Hum Genet. 2020 Jun 4;106(6):872-884. doi: 10.1016/j.ajhg.2020.04.016

Menke, L.A., Study, T.D., Gardeitchik, T., Hammond, P., Heimdal, K.R., Houge, G., Hufnagel, S.B., Ji, J., Johansson, S., Kant, S.G., Kinning, E., Leon, E.L., Newbury-Ecob, R., Paolacci, S., Pfundt, R., Ragge, N.K., Rinne, T., Ruivenkamp, C., Saitta, S.C., Sun, Y., Tartaglia, M., Terhal, P.A., van Essen, A.J., Vigeland, M.D., Xiao, B., Hennekam, R.C., 2018. Further delineation of an entity caused by CREBBP and EP300 mutations but not resembling Rubinstein–Taybi syndrome. Am. J. Med. Genet. A. 176, 862–876. https://doi.org/10.1002/ajmg.a.38626

Menke, L.A., van Belzen, M.J., Alders, M., Cristofoli, F., Study, T.D., Ehmke, N., Fergelot, P., Foster, A., Gerkes, E.H., Hoffer, M.J.V., Horn, D., Kant, S.G., Lacombe, D., Leon, E., Maas, S.M., Melis, D., Muto, V., Park, S.-M., Peeters, H., Peters, D.J.M., Pfundt, R., van Ravenswaaij-Arts, C.M.A., Tartaglia, M., Hennekam, R.C.M., 2016. CREBBP mutations in individuals without Rubinstein–Taybi syndrome phenotype. Am. J. Med. Genet. A. 170, 2681–2693. https://doi.org/10.1002/ajmg.a.37800

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. Genet. Med. 20, 159–163. https://doi.org/10.1038/gim.2017.86

Muylle, E., Jiang, H., Johnsen, C., Byeon, S.K., Ranatunga, W., Garapati, K., Zenka, R.M., Preston, G., Pandey, A., Kozicz, T., Fang, F., Morava, E., 2022. TRIT1 defect leads to a recognizable phenotype of myoclonic epilepsy, speech delay, strabismus, progressive spasticity, and normal lactate levels. J. Inherit. Metab. Dis. 45, 1039–1047. https://doi.org/10.1002/jimd.12550

Nakashima, M., Kato, M., Aoto, K., Shiina, M., Belal, H., Mukaida, S., Kumada, S., Sato, A., Zerem, A., Lerman-Sagie, T., Lev, D., Leong, H.Y., Tsurusaki, Y., Mizuguchi, T., Miyatake, S., Miyake, N., Ogata, K., Saitsu, H., Matsumoto, N., 2018. De novo hotspot variants in CYFIP2 cause early-onset epileptic encephalopathy. Ann. Neurol. 83, 794–806. https://doi.org/10.1002/ana.25208

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S.J., Diekhans, M., Logsdon, G.A., Alonge, M., Antonarakis, S.E., Borchers, M., Bouffard, G.G., Brooks, S.Y., Caldas, G.V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L.G., Dishuck, P.C., Durbin, R., Dvorkina, T., Fiddes, I.T., Formenti, G., Fulton, R.S., Fungtammasan, A., Garrison, E., Grady, P.G.S., Graves-Lindsay, T.A., Hall, I.M., Hansen, N.F., Hartley, G.A., Haukness, M., Howe, K., Hunkapiller, M.W., Jain, C., Jain, M., Jarvis, E.D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V.V., Marschall, T., McCartney, A.M., McDaniel, J., Miller, D.E., Mullikin, J.C., Myers, E.W., Olson, N.D., Paten, B., Peluso, P., Pevzner, P.A., Porubsky, D., Potapova, T., Rogaev, E.I., Rosenfeld, J.A., Salzberg, S.L., Schneider, V.A., Sedlazeck, F.J., Shafin, K., Shew, C.J., Shumate, A., Sims, Y., Smit, A.F.A., Soto, D.C., Sović, I., Storer, J.M., Streets, A., Sullivan, B.A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B.P., Wenger, A., Wood, J.M.D., Xiao, C., Yan, S.M., Young, A.C., Zarate, S., Surti, U., McCoy, R.C., Dennis, M.Y., Alexandrov, I.A., Gerton, J.L., O'Neill, R.J., Timp, W., Zook, J.M., Schatz, M.C., Eichler, E.E., Miga, K.H., Phillippy, A.M., 2022. The complete sequence of a human genome. Science 376, 44–53. https://doi.org/10.1126/science.abj6987

Panda, A., Subramanian, K., Kahali, B., 2021. Implementation of human whole genome sequencing data analysis: A containerized framework for sustained and enhanced throughput. Inform. Med. Unlocked 25, 100684. https://doi.org/10.1016/j.imu.2021.100684

Parenti, I., Rabaneda, L.G., Schoen, H., Novarino, G., 2020. Neurodevelopmental Disorders: From Genetics to Functional Pathways. Trends Neurosci. 43, 608–621. https://doi.org/10.1016/j.tins.2020.05.004

Perenthaler, E., Yousefi, S., Niggl, E., Barakat, T.S., 2019. Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development. Front. Cell. Neurosci. 13.

Pettersson, M., Grochowski, C.M., Wincent, J., Eisfeldt, J., Breman, A.M., Cheung, S.W., Krepischi, A.C.V., Rosenberg, C., Lupski, J.R., Ottosson, J., Lovmar, L., Gacic, J., Lundberg, E.S., Nilsson, D., Carvalho, C.M.B., Lindstrand, A., 2020. Cytogenetically visible inversions are formed by multiple molecular mechanisms. Hum. Mutat. 41, 1979–1998. https://doi.org/10.1002/humu.24106

Polla, D.L., Bhoj, E.J., Verheij, J.B.G.M., Wassink-Ruiter, J.S.K., Reis, A., Deshpande, C., Gregor, A., Hill-Karfe, K., Silfhout, A.T.V., Pfundt, R., Bongers, E.M.H.F., Hakonarson, H., Berland, S., Gradek, G., Banka, S., Chandler, K., Gompertz, L., Huffels, S.C., Stumpel, C.T.R.M., Wennekes, R., Stegmann, A.P.A., Reardon, W., Leenders, E.K.S.M., Vries, B.B.A. de, Li, D., Zackai, E., Ragge, N., Lynch, S.A., Cuddapah, S., Bokhoven, H. van, Zweier, C., Brouwer, A.P.M. de, 2021. De novo variants in MED12 cause X-linked syndromic neurodevelopmental disorders in 18 females. Genet. Med. 23, 645–652. https://doi.org/10.1038/s41436-020-01040-6

Qin, Z., Su, J., Li, M., Yang, Q., Yi, Shang, Zheng, H., Zhang, Q., Chen, F., Yi, Sheng, Lu, W., Li, W., Huang, L., Xu, J., Shen, Y., Luo, J., 2020. Clinical and Genetic Analysis of CHD7 Expands the Genotype and Phenotype of CHARGE Syndrome. Front. Genet. 11.

Raible, S.E., Mehta, D., Bettale, C., Fiordaliso, S., Kaur, M., Medne, L., Rio, M., Haan, E., White, S.M., Cusmano-Ozog, K., Nishi, E., Guo, Y., Wu, H., Shi, X., Zhao, Q., Zhang, X., Lei, Q., Lu, A., He, X., Okamoto, N., Miyake, N., Piccione, J., Allen, J., Matsumoto, N., Pipan, M., Krantz, I.D., Izumi, K., 2019. Clinical and molecular spectrum of CHOPS syndrome. Am. J. Med. Genet. A. 179, 1126–1138. https://doi.org/10.1002/ajmg.a.61174

Rajan-Babu, I.-S., Peng, J.J., Chiu, R., Birch, P., Couse, M., Guimond, C., Lehman, A., Mwenifumbo, J., van Karnebeek, C., Friedman, J., Adam, S., Souich, C.D., Elliott, A., Lehman, A., Mwenifumbo, J., Nelson, T., van Karnebeek, C., Friedman, J., Li, C., Mohajeri, A., Dolzhenko, E., Eberle, M.A., Birol, I., Friedman, J.M., IMAGINE Study, CAUSES Study, 2021. Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions. Genome Med. 13, 126. https://doi.org/10.1186/s13073-021-00932-9

Rao, A.R., Nelson, S.F., 2018. Calculating the statistical significance of rare variants causal for Mendelian and complex disorders. BMC Med. Genomics 11, 53. https://doi.org/10.1186/s12920-018-0371-9

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

Richards, L., Das, S., Nordman, J.T., 2022. Rif1-Dependent Control of Replication Timing. Genes 13, 550. https://doi.org/10.3390/genes13030550

Riggs, E.R., Andersen, E.F., Cherry, A.M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D.I., South, S.T., Thorland, E.C., Pineda-Alvarez, D., Aradhya, S., Martin, C.L., 2020. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genet. Med. 22, 245–257. https://doi.org/10.1038/s41436-019-0686-8

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

Sahajpal, N.S., Jill Lai, C.-Y., Hastie, A., Mondal, A.K., Dehkordi, S.R., van der Made, C.I., Fedrigo, O., Al-Ajli, F., Jalnapurkar, S., Byrska-Bishop, M., Kanagal-Shamanna, R., Levy, B., Schieck, M., Illig, T., Bacanu, S.-A., Chou, J.S., Randolph, A.G., Rojiani, A.M., Zody, M.C., Brownstein, C.A., Beggs, A.H., Bafna, V., Jarvis, E.D., Hoischen, A., Chaubey, A., Kolhe, R., 2022. Optical genome mapping identifies rare structural variations as predisposition factors associated with severe COVID-19. iScience 25, 103760. https://doi.org/10.1016/j.isci.2022.103760

Savatt, J.M., Myers, S.M., 2021. Genetic Testing in Neurodevelopmental Disorders. Front. Pediatr. 9.

Schuermans, N., Hemelsoet, D., Terryn, W., Steyaert, S., Van Coster, R., Coucke, P.J., Steyaert, W., Callewaert, B., Bogaert, E., Verloo, P., Vanlander, A.V., Debackere, E., Ghijsels, J., LeBlanc, P., Verdin, H., Naesens, L., Haerynck, F., Callens, S., Dermaut, B., Poppe, B., De Bleecker, J., Santens, P., Boon, P., Laureys, G., Kerre, T., for UD-PrOZA, 2022. Shortcutting the diagnostic odyssey: the multidisciplinary Program for Undiagnosed Rare Diseases in adults (UD-PrOZA). Orphanet J. Rare Dis. 17, 210. https://doi.org/10.1186/s13023-022-02365-y

Seaby, E.G., Smedley, D., Taylor Tavares, A.L., Brittain, H., van Jaarsveld, R.H., Baralle, D., Rehm, H.L., O'Donnell-Luria, A., Ennis, S., 2022. A gene-to-patient approach uplifts novel disease gene discovery and identifies 18 putative novel disease genes. Genet. Med. 24, 1697–1707. https://doi.org/10.1016/j.gim.2022.04.019

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., Schatz, M.C., 2018. Accurate detection of complex structural variations using single molecule sequencing. Nat. Methods 15, 461–468. https://doi.org/10.1038/s41592-018-0001-7

Shihab, H.A., Rogers, M.F., Gough, J., Mort, M., Cooper, D.N., Day, I.N.M., Gaunt, T.R., Campbell, C., 2015. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 31, 1536–1543. https://doi.org/10.1093/bioinformatics/btv009

Spielmann, M., Lupiáñez, D.G., Mundlos, S., 2018. Structural variation in the 3D genome. Nat. Rev. Genet. 19, 453–467. https://doi.org/10.1038/s41576-018-0007-0

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. Genet. Med. 21, 2413–2421. https://doi.org/10.1038/s41436-019-0554-6

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., Robinson, G.E., 2015. Big Data: Astronomical or Genomical? PLOS Biol. 13, e1002195. https://doi.org/10.1371/journal.pbio.1002195

Stranneheim, H., Lagerstedt-Robinson, K., Magnusson, M., Kvarnung, M., Nilsson, D., Lesko, N., Engvall, M., Anderlid, B.-M., Arnell, H., Johansson, C.B., Barbaro, M., Björck, E., Bruhn, H., Eisfeldt, J., Freyer, C., Grigelioniene, G., Gustavsson, P., Hammarsjö, A., Hellström-Pigg, M., Iwarsson, E., Jemt, A., Laaksonen, M., Enoksson, S.L., Malmgren, H., Naess, K., Nordenskjöld, M., Oscarson, M., Pettersson, M., Rasi, C., Rosenbaum, A., Sahlin, E., Sardh, E., Stödberg, T., Tesi, B., Tham, E., Thonberg, H., Töhönen, V., von Döbeln, U., Vassiliou, D., Vonlanthen, S., Wikström, A.-C., Wincent, J., Winqvist, O., Wredenberg, A., Ygberg, S., Zetterström, R.H., Marits, P., Soller, M.J., Nordgren, A., Wirta, V., Lindstrand, A., Wedell, A., 2021. Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. Genome Med. 13, 40. https://doi.org/10.1186/s13073-021-00855-5

Suvakov, M., Panda, A., Diesh, C., Holmes, I., Abyzov, A., 2021. CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing. GigaScience 10, giab074. https://doi.org/10.1093/gigascience/giab074

Tan, T.Y., Lunke, S., Chong, B., Phelan, D., Fanjul-Fernandez, M., Marum, J.E., Kumar, V.S., Stark, Z., Yeung, A., Brown, N.J., Stutterd, C., Delatycki, M.B., Sadedin, S., Martyn, M., Goranitis, I., Thorne, N., Gaff, C.L., White, S.M., 2019. A head-to-head evaluation of the diagnostic efficacy and costs of trio versus singleton exome sequencing analysis. Eur. J. Hum. Genet. 27, 1791–1799. https://doi.org/10.1038/s41431-019-0471-9

Tankard, R.M., Bennett, M.F., Degorski, P., Delatycki, M.B., Lockhart, P.J., Bahlo, M., 2018. Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. Am. J. Hum. Genet. 103, 858–873. https://doi.org/10.1016/j.ajhg.2018.10.015

Tomaselli, P.J., Rossor, A.M., Horga, A., Jaunmuktane, Z., Carr, A., Saveri, P., Piscosquito, G., Pareyson, D., Laura, M., Blake, J.C., Poh, R., Polke, J., Houlden, H., Reilly, M.M., 2017. Mutations in noncoding regions of GJB1 are a major cause of X-linked CMT. Neurology 88, 1445–1453. https://doi.org/10.1212/WNL.0000000000003819

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., Stevens, S.J., van den Wijngaard, A., Brunner, H.G., Yntema, H.G., Gilissen, C., Nelen, M.R., Vissers, L.E.L.M., 2023. The performance of genome sequencing as a first-tier test for neurodevelopmental disorders. Eur. J. Hum. Genet. 31, 81–88. https://doi.org/10.1038/s41431-022-01185-9

Vanlerberghe C, Faivre L, Petit F, Fruchart O, Jourdain AS, Clavier F, Gay S, Manouvrier-Hanu S, Escande F. Intrafamilial variability of ZRS-associated syndrome: characterization of a mosaic ZRS mutation by pyrosequencing. Clin Genet. 2015 Nov;88(5):479-83. doi: 10.1111/cge.12534. Epub 2015 Jan 6. PMID: 25382487.

Veltman, J.A., Brunner, H.G., 2012. De novo mutations in human genetic disease. Nat. Rev. Genet. 13, 565–575. https://doi.org/10.1038/nrg3241

Vleugels, W., Haeuptle, M.A., Ng, B.G., Michalski, J.-C., Battini, R., Dionisi-Vici, C., Ludman, M.D., Jaeken, J., Foulquier, F., Freeze, H.H., Matthijs, G., Hennet, T., 2009. RFT1 deficiency in three novel CDG patients. Hum. Mutat. 30, 1428–1434. https://doi.org/10.1002/humu.21085

Wala, J.A., Bandopadhayay, P., Greenwald, N.F., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S., Li, Y., Weischenfeldt, J., Yao, X., Nusbaum, C., Campbell, P., Getz, G., Meyerson, M., Zhang, C.-Z., Imielinski, M., Beroukhim, R., 2018. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 28, 581–591. https://doi.org/10.1101/gr.221028.117

Will, A.J., Cova, G., Osterwalder, M., Chan, W.-L., Wittler, L., Brieske, N., Heinrich, V., de Villartay, J.-P., Vingron, M., Klopocki, E., Visel, A., Lupiáñez, D.G., Mundlos, S., 2017. Composition and dosage

of a multipartite enhancer cluster control developmental expression of Ihh (Indian hedgehog). Nat. Genet. 49, 1539–1545. https://doi.org/10.1038/ng.3939

Wratten, L., Wilm, A., Göke, J., 2021. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat. Methods 18, 1161–1168. https://doi.org/10.1038/s41592-021-01254-9

Wright, C.F., Quaife, N.M., Ramos-Hernández, L., Danecek, P., Ferla, M.P., Samocha, K.E., Kaplanis, J., Gardner, E.J., Eberhardt, R.Y., Chao, K.R., Karczewski, K.J., Morales, J., Gallone, G., Balasubramanian, M., Banka, S., Gompertz, L., Kerr, B., Kirby, A., Lynch, S.A., Morton, J.E.V., Pinz, H., Sansbury, F.H., Stewart, H., Zuccarelli, B.D., Cook, S.A., Taylor, J.C., Juusola, J., Retterer, K., Firth, H.V., Hurles, M.E., Lara-Pezzi, E., Barton, P.J.R., Whiffin, N., 2021. Non-coding region variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms. Am. J. Hum. Genet. 108, 1083–1094. https://doi.org/10.1016/j.ajhg.2021.04.025

Xu, C., Yang, X., Zhou, H., Li, Y., Xing, C., Zhou, T., Zhong, D., Lian, C., Yan, M., Chen, T., Liao, Z., Gao, B., Su, D., Wang, T., Sharma, S., Mohan, C., Ahituv, N., Malik, S., Li, Q.-Z., Su, P., 2020. A novel ZRS variant causes preaxial polydactyly type I by increased sonic hedgehog expression in the developing limb bud. Genet. Med. 22, 189–198. https://doi.org/10.1038/s41436-019-0626-7

Yoo, A.B., Jette, M.A., Grondona, M., 2003. SLURM: Simple Linux Utility for Resource Management, in: Feitelson, D., Rudolph, L., Schwiegelshohn, U. (Eds.), Job Scheduling Strategies for Parallel Processing, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 44–60. https://doi.org/10.1007/10968987_3

Yu, L., Zhou, H., Hu, F., Xu, Y., 2013. Two novel mutations of the GTP cyclohydrolase 1 gene and genotype–phenotype correlation in Chinese Dopa-responsive dystonia patients. Eur. J. Hum. Genet. 21, 731–735. https://doi.org/10.1038/ejhg.2012.239

Zhao, S., Agafonov, O., Azab, A., Stokowy, T., Hovig, E., 2020. Accuracy and efficiency of germline variant calling pipelines for human genome data. Sci. Rep. 10, 20222. https://doi.org/10.1038/s41598-020-77218-4

Zhu, M., Need, A.C., Han, Y., Ge, D., Maia, J.M., Zhu, Q., Heinzen, E.L., Cirulli, E.T., Pelak, K., He, M., Ruzzo, E.K., Gumbs, C., Singh, A., Feng, S., Shianna, K.V., Goldstein, D.B., 2012. Using ERDS to Infer Copy-Number Variants in High-Coverage Genomes. Am. J. Hum. Genet. 91, 408–421. https://doi.org/10.1016/j.ajhg.2012.07.004

Zook, J.M., Hansen, N.F., Olson, N.D., Chapman, L.M., Mullikin, J.C., Xiao, C., Sherry, S., Koren, S., Phillippy, A.M., Boutros, P.C., Sahraeian, S.M.E., Huang, V., Rouette, A., Alexander, N., Mason, C.E., Hajirasouliha, I., Ricketts, C., Lee, J., Tearle, R., Fiddes, I.T., Barrio, A.M., Wala, J., Carroll, A., Ghaffari, N., Rodriguez, O.L., Bashir, A., Jackman, S., Farrell, J.J., Wenger, A.M., Alkan, C., Soylev, A., Schatz, M.C., Garg, S., Church, G., Marschall, T., Chen, K., Fan, X., English, A.C., Rosenfeld, J.A., Zhou, W., Mills, R.E., Sage, J.M., Davis, J.R., Kaiser, M.D., Oliver, J.S., Catalano, A.P., Chaisson,

M.J., Spies, N., Sedlazeck, F.J., Salit, M., 2020. A robust benchmark for detection of germline large deletions and insertions. Nat. Biotechnol. 38, 1347–1355. https://doi.org/10.1038/s41587-020-0538-8

Sitography

OMIM: https://www.omim.org/

GENCODE: https://www.gencodegenes.org/

1000 Genomes Project Data Portal: https://www.internationalgenome.org/data-portal/sample

GATK: https://gatk.broadinstitute.org/hc/en-us

GIAB: https://www.nist.gov/programs-projects/genome-bottle

Sysndd: https://sysndd.dbmr.unibe.ch/

GitHub https://github.com/Manuelaio/WGS_SNV_pipeline/

Acknowledgments

Several people have been of essential support in the making of this thesis. First, I am grateful to Professor Jonathan Sebat and James Guevara, University of California San Diego, for their collaboration and contribution to the material of benchmark analysis and de novo SVs interpretation. I would like to thank Pamela Magini and Sonia Bonora, U.O. Genetica Medica, IRCC Azienda Ospedialiero-Universitaria of Bologna, helped in performing molecular validation of SNVs and SVs of this study.