UNIVERSITY OF VERONA
DEPARTMENT OF BIOTECHNOLOGY

DOCTORAL PROGRAM IN BIOTECHNOLOGY
XXXVIII CYCLE

Exome CNV Calling and Constellation Mapped Read for Clinical Structural Variant Analysis

S.S.D. BIO/18

Coordinator: Prof.ssa Flavia Guzzo

Tutor: Prof. Massimo Delledonne

Doctoral Student: Matteo Orlandi

La borsa di dottorato è stata cofinanziata con le risorse del PNRR:

• per il DM 351 nell'ambito della Missione 4 ("Istruzione e ricerca") – Componente 1 ("Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università"), Investimento 3.4. ("Didattica e competenze universitarie avanzate") e Investimento 4.1 ("Estensione del numero di dottorati di ricerca e dottorati innovativi per la pubblica amministrazione e il patrimonio culturale") - progetto M4C1 –Inv. 3.4 e progetto M4C1 – Inv. 4.1

• per il DM 352, nell'ambito della Missione 4 ("Istruzione e Ricerca") – Componente 2 ("Dalla Ricerca all'Impresa"), Investimento 3.3 ("Introduzione di dottorati innovativi che rispondono ai fabbisogni di innovazione delle imprese e promuovono l'assunzione dei ricercatori da parte delle imprese") – progetto M4C2 Investimento 3.3

Exome CNV Calling and Constellation Mapped Read for Clinical Structural Variant Analysis

Matteo Orlandi

PhD thesis

Verona, 09 December 2025

Table of Contents

Sommario

Nonostante l'ampio utilizzo del sequenziamento dell'esoma (WES) e del genoma (WGS) nella diagnostica delle malattie rare, molti pazienti rimangono privi di una diagnosi molecolare. Le pipeline di analisi routinarie sono altamente ottimizzate per la chiamata di varianti a singolo nucleotide (SNV) e per le piccole inserzioni o delezioni (indel), mentre le variazioni del numero di copie (CNV) e, più in generale, altri riarrangiamenti strutturali (SV) sono spesso analizzati con saggi separati oppure indagati solo in casi selezionati. Questa tesi si chiede fino a che punto il sequenziamento short read possa essere spinto verso una identificazione più completa delle varianti e quanto le tecnologie attuali possano avvicinarsi a un singolo saggio genomico in grado di catturare congiuntamente SNV, indel, CNV e SV complesse.

La prima parte del lavoro è incentrata sulle CNV da dati di sequenziamento dell'esoma. A partire dall'osservazione di insiemi di CNV instabili e difficili da gestire a livello clinico in esomi prodotti in casa, ho sviluppato una pipeline basata su Snakemake che integra tre software per la chiamata di CNV da esoma, ExomeDepth, ClinCNV e gCNV, all'interno di una fase di preprocessamento e di una successiva armonizzazione dei risultati. Il benchmarking sulla coorte di riferimento CNVPANEL01, che include CNV validate, e l'applicazione a una coorte eterogenea di esomi clinici dell'Ospedale Burlo Garofalo hanno mostrato che i tre strumenti raggiungono una sensibilità grossolanamente simile sugli eventi curati, in particolare per le delezioni, ma differiscono in modo marcato in termini di comportamento di segmentazione, numero di chiamate e robustezza rispetto alla composizione della baseline. ExomeDepth è altamente sensibile ma tende a produrre un maggior numero di chiamate a seguito di artefatti di copertura dipendenti dal contenuto GC e dalla scelta del pannello di normali. ClinCNV produce un numero minore di eventi, più continui, ed è tollerante rispetto all'eterogeneità della baseline. gCNV beneficia di una modellizzazione congiunta di coorti numerose, ma è computazionalmente più impegnativo. Una strategia di consenso che mantiene gli eventi supportati da almeno due software, seguita da annotazione clinica e filtraggio basato sulla frequenza, comprime il carico di

chiamate a una dimensione più compatibile con la revisione diagnostica, preservando al tempo stesso le CNV probabilmente patogene e patogene. Parallelamente, i risultati mettono in evidenza limiti intrinseci della chiamata di CNV da esoma, tra cui la dipendenza da pannelli di controlli normali, la natura discontinua del disegno di cattura e la copertura incompleta delle regioni regolatorie.

La seconda parte della tesi valuta il nuovo metodo Constellation mapped read di Illumina come saggio di genoma completo sensibile alla prossimità, che arricchisce il sequenziamento short read standard con informazione a lungo raggio derivata da cluster spazialmente prossimi sulla flow cell. In una coorte di sei genomi clinicamente caratterizzati, con riarrangiamenti strutturali noti e includendo un caso con WGS TruSeq PCR free appaiato, ho confrontato Constellation mapped read e il sequenziamento WGS standard in termini di copertura, frazione di genoma chiamabile, chiamata di varianti puntiformi, phasing e individuazione di SV. Constellation mapped read mantiene le prestazioni sulle SNV/indel del WGS PCR free, ma ricostruisce template lunghi che consentono un phasing in blocchi aplotipici che si estendono frequentemente per decine di megabasi. La chiamata di SV produce insiemi di CNV e SV più ampi e meglio supportati rispetto al dataset TruSeq appaiato e la combinazione di profondità di lettura, chiamate a livello di giunzione e matrici di colocalizzazione genome wide supporta ricostruzioni dettagliate di eventi complessi, inclusi duplicazioni in più fasi in SCN1A, traslocazioni bilanciate e microdelezioni focali.

Nel loro insieme, questi risultati mostrano che pipeline multi caller possono estrarre dai dati di esoma informazioni sulle CNV utili dal punto di vista clinico, pur in presenza di vincoli fondamentali, e che Constellation mapped read rappresenta una via promettente verso un saggio genomico integrato che preserva i punti di forza del sequenziamento del genoma completo PCR free e allo stesso tempo aggiunge phasing a lungo raggio e una risoluzione migliorata delle SV.

Abstract

Despite the widespread use of whole exome sequencing (WES) and whole genome sequencing (WGS) in rare disease diagnostics, many patients remain without a molecular diagnosis. Routine pipelines are highly optimized for calling single nucleotide variants (SNVs) and small insertions or deletions (indels), whereas copy number variants (CNV) and other structural rearrangements (SV) are often handled by separate assays or only investigated in selected cases. This thesis asks how far short read sequencing can be pushed toward comprehensive variant detection and how close current technologies can come to a single genome assay that jointly captures SNVs, indels, CNVs and complex structural variants.

The first part of the work focuses on CNV derived from WES. Motivated by unstable and clinically unmanageable CNV call sets in in house Twist exomes, I developed a Snakemake based pipeline that integrates three exome CNV callers, ExomeDepth, ClinCNV and gCNV, under harmonized preprocessing and configuration. Benchmarking on the CNVPANEL01 reference cohort with validated CNVs and application to a heterogeneous clinical exome cohort from Burlo Garofalo Hospital showed that the three tools reach broadly similar sensitivity on curated events, especially for deletions, but differ markedly in segmentation behavior, call volumes and robustness to baseline composition. ExomeDepth is highly sensitive but prone to call inflation driven by GC dependent coverage artefacts and by the choice of normal reference panel. ClinCNV produces fewer, more contiguous events and is more tolerant of baseline heterogeneity. gCNV benefits from joint modelling of large cohorts but is computationally demanding. A consensus strategy that retains CNVs supported by at least two callers, followed by clinical annotation and frequency based filtering, compresses the call burden to a size that is more compatible with diagnostic review while preserving the likely pathogenic and pathogenic CNVs. At the same time, the results highlight intrinsic limits of exome based CNV calling, including dependence on panels of normals, discontinuous target design and incomplete coverage of regulatory regions.

The second part of the thesis evaluates Illumina Constellation Mapped Read technology as a proximity aware whole genome assay that augments standard short read sequencing with long range information derived from spatially proximate clusters on the flow cell. In a cohort of six clinically characterized genomes with known structural rearrangements, including one case with matched TruSeq PCR free WGS, I compared Constellation Mapped Read and standard WGS in terms of coverage, callable genome, small variant calling, phasing and structural variant detection. Constellation maintains the small variant performance of PCR free WGS, but reconstructs long templates which enable phasing into haplotype blocks that frequently extend to tens of megabases. SV calling yields larger and more supported CNV and SV call sets than the matched TruSeq dataset, and the combination of read depth, junction level calls and genome wide colocation matrices supports detailed reconstructions of complex events, including multi step duplications at SCN1A, balanced translocations and focal microdeletions.

Taken together, these results show that carefully engineered multi caller pipelines can extract clinically useful CNV information from exome data, although fundamental constraints remain, and that Constellation mapped read is a promising path toward an integrated genome assay that preserves the strengths of PCR free WGS while adding long range phasing and enhanced structural resolution.

1. Introduction

1.1. Genomic medicine in the era of next generation sequencing

1.1.1. From Sanger sequencing to high throughput genomics

Before next generation sequencing, diagnostic genetics relied mainly on Sanger sequencing of one gene at a time. This approach provided highly accurate reads but required a strong a priori hypothesis and offered limited throughput, so clinical workflows focused on small panels of candidate genes and were often slow, expensive and inconclusive for disorders with heterogeneous or atypical presentations.

Massively parallel short read sequencing transformed this landscape by enabling millions of DNA fragments to be sequenced in a single run, with a dramatic reduction in cost per base and turnaround time. This class of technologies, often referred to as next generation sequencing, now underpins a continuum of assays that ranges from focused multi gene panels to whole exome sequencing and whole genome sequencing. In rare disease diagnostics, this flexibility allows laboratories to match test breadth to the clinical question. Narrowly targeted panels remain appropriate for well-defined phenotypes with limited locus heterogeneity, whereas exome and genome sequencing provide unbiased surveys of the coding and non-coding genome in patients with broad or complex presentations [1], [2].

Large scale population sequencing has catalogued tens of millions of single nucleotide variants (SNVs) and millions of small insertions or deletions (indels) across diverse human ancestries, providing an essential reference background for patient specific variant interpretation [3]. These resources underpin most modern pipelines for rare disease genomics and influence both the design of clinical tests and the interpretation of their results.

1.1.2. Diagnostic yield and current clinical practice

Whole exome sequencing (WES) has been widely adopted as a first line or early second line test in many diagnostic laboratories and national genomic medicine

programs. Large clinical series across mixed rare disease cohorts consistently report molecular diagnosis rates around 25-30%, with higher yields in some phenotypic groups such as neurodevelopmental disorders and early onset epilepsies [1], [2]. In most studies, a diagnosis corresponds to pathogenic or likely pathogenic variants in known disease genes that plausibly explain the patient's phenotype and that, in a substantial fraction of cases, lead to changes in management, surveillance or reproductive counselling [1].

Whole genome sequencing (WGS) is increasingly used either as a first tier test in selected settings or as a second tier investigation after non diagnostic exome sequencing. Comparative studies that reanalyze exome negative families with whole genome sequencing report an additional diagnostic yield of roughly 10-15 per cent. This gain reflects both technical advantages of WGS, in particular more uniform coverage of exons, improved detection of structural variations and the ability to interrogate intronic and regulatory regions [4], [5].

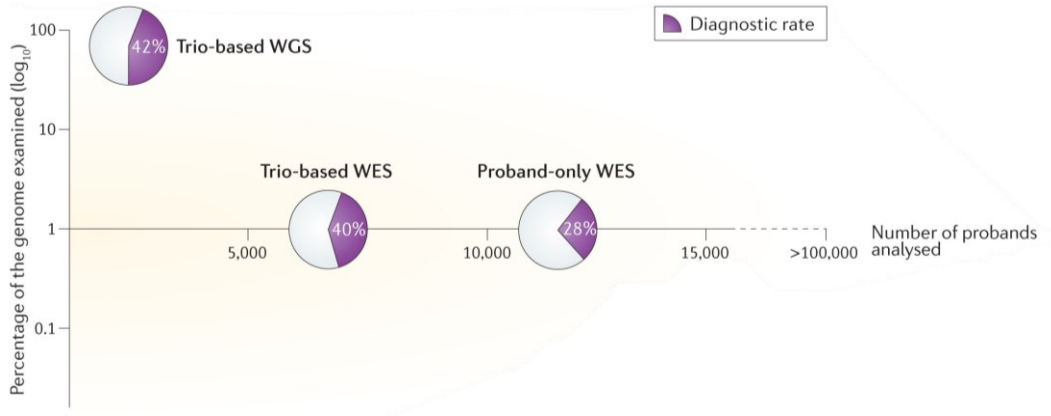

Figure 1.1: Whole-exome sequencing (WES) and whole-genome sequencing (WGS) diagnostic yield. Image adapted from [6].

Despite the WGS advantages, exome sequencing remains the predominant first line test in many health systems, because it has lower cost, more manageable data volumes and well established clinical workflows, while genome sequencing is often reserved for particularly complex phenotypes, critically ill patients or cases where non coding or structural variants (SVs) are strongly suspected.

1.1.3. Analytical workflows and limitations of short read sequencing

Short read sequencing data become clinically useful only after a multi-step analytical workflow. In widely adopted pipelines, raw reads are first aligned to the reference genome using algorithms such as BWA, followed by duplicate marking, local realignment or assembly around candidate variant sites, base quality recalibration and probabilistic variant calling with tools such as the Genome Analysis Toolkit or DRAGEN [7], [8]. Best practice recommendations describe end to end workflows from FASTQ files to high confidence variant calls for both exome and genome sequencing, and these frameworks have become de facto standards in research and diagnostics [8].

The analytical performance of single nucleotide variant and indel calling has been extensively benchmarked using high confidence reference truth sets from the Genome in a Bottle consortium, which integrates multiple sequencing technologies to define robust genotypes for well characterized samples. In uniquely mappable, well covered regions, both exome and genome sequencing can achieve very high sensitivity and precision for germline single nucleotide variants and small indels [9].

Several technical limitations remain. Exome capture introduces non uniform coverage because of variable hybridization efficiency and PCR amplification, which leads to subsets of exons that systematically fail to reach diagnostic depth and to differences between commercial kits in terms of which regions are well covered [9]. Both exome and genome sequencing perform less well in regions of low complexity, segmental duplications or extreme GC content, where mapping ambiguity and coverage artefacts can reduce sensitivity or increase false positives. In practice, this means that short read pipelines offer excellent performance for the easy fraction of the genome, but leave blind spots in technically challenging loci and in classes of variation that are inherently difficult to resolve with short reads, such as many structural variants. These constraints motivate complementary approaches and contribute to the residual diagnostic gap that persists after a standard exome or genome analysis.

1.2. Single nucleotide variants and small insertions/deletions in rare disease diagnostics

1.2.1. Spectrum, functional impact and detection

Single nucleotide variants change one base in the genome, either in the germline or, in the somatic context, in specific tissues. Small insertions or deletions, typically up to a few dozen base pairs, add or remove short stretches of sequence (Figure 1.2). Together, single nucleotide variants and small indels represent the most abundant class of genetic variation in humans. Population scale sequencing across diverse ancestries has revealed tens of millions of single nucleotide variants and millions of short indels, far outnumbering larger structural variants, and has provided a detailed map of background variation that is critical for clinical interpretation [3].

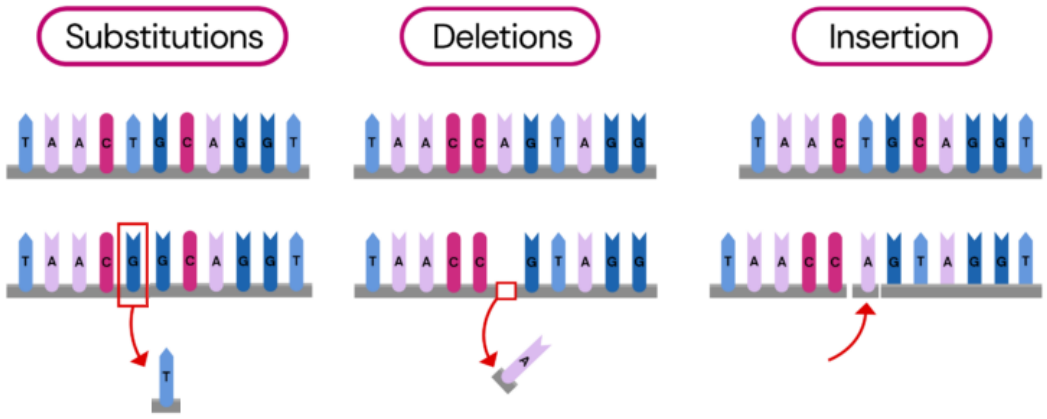

Figure 1.2: Single Nucleotide Variants (SNVs), small Deletions (DELs) and small Insertions (INSs). Adapted from [10].

In coding regions, these variants generate a wide spectrum of functional consequences. Missense changes alter amino acids and can perturb protein structure or interactions. Nonsense variants and frameshifting indels often introduce premature stop codons, which can trigger nonsense mediated decay or produce truncated proteins. Synonymous changes are frequently neutral but can affect splicing or translation in specific contexts. Variants at canonical splice sites, or within exonic and intronic regulatory motifs, can profoundly disrupt mRNA

processing. These classes of coding and splice altering variants constitute the core substrate of Mendelian disease gene discovery and routine diagnostic reporting.

Non coding single nucleotide variants and indels can also be pathogenic. Genome wide association studies and regulatory genomics have shown that many disease associated variants map to enhancers, promoters, insulators and other regulatory elements. Integrative analyses that combine association signals with functional maps from ENCODE and related projects demonstrate strong enrichment of disease variants in cell type specific regulatory regions [11], [12]. Mechanistic reviews have emphasized how non coding mutations can disrupt transcription factor binding, chromatin architecture or non-coding RNA function and thereby contribute to rare and common disease [13].

Short read exome and genome sequencing are particularly effective at detecting single nucleotide variants and small indels, provided that reads can be uniquely aligned and coverage is adequate. In the benchmark regions defined by Genome in a Bottle, germline single nucleotide variants are detected with near complete sensitivity and high precision, and performance for small indels has steadily improved with better algorithms and error models [9]. Nonetheless, non-uniform exome capture and difficult genomic contexts mean that even this well served class of variants is not captured perfectly, and clinically relevant bases can still fall below diagnostic depth in some capture designs.

1.2.2. Coding versus noncoding variation in disease

Turning exhaustive lists of single nucleotide variants and indels into clinically meaningful reports requires systematic integration of multiple layers of evidence. Variant annotation typically includes gene context, predicted consequence on transcripts and proteins, known disease associations and allele frequencies from large population datasets. In silico prediction tools such as PolyPhen 2 and CADD prioritize variants by combining conservation, structural features and a broad range of functional annotations into scores that correlate with likely pathogenicity [14], [15].

Despite these resources, interpretation remains a major bottleneck. Even apparently healthy individuals harbor large numbers of rare or private variants in genes linked to disease, which makes it difficult to distinguish benign background variation from truly pathogenic alleles. To address this problem, most clinical laboratories follow the standards and guidelines of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. These guidelines define a five tier classification, from benign to pathogenic, and specify how to weight evidence from population frequency, computational predictions, functional assays, segregation, de novo status and prior reports [16]. Computational tools that formalize these rules can improve consistency but cannot overcome the fundamental limitation that many variants lack sufficient data. As a result, a high proportion of changes detected by exome or genome sequencing are reported as variants of uncertain significance and may require periodic reevaluation.

Overall, single nucleotide variants and small indels are the best characterized class of genomic variation, and short read exome and genome sequencing detect them with high analytical performance in well behaved regions of the genome. Sophisticated pipelines for annotation and interpretation have matured to the point where, in many cohorts, the majority of diagnoses are driven by this variant class. At the same time, the incomplete coverage of regulatory elements, the challenges of interpreting non coding changes and the relative insensitivity of short reads to many structural variants and copy number changes leave a substantial fraction of patients without a molecular diagnosis. Closing this gap requires extending sequencing based diagnostics toward more complex forms of variation, such as copy number variants and other structural rearrangements, which are the focus of the subsequent sections of this thesis.

1.3. Structural variants as the “missing” component of diagnostic yield

Although SNVs and small indels explain a substantial fraction of Mendelian disease, a large proportion of patients with a strong clinical suspicion still remain without a molecular diagnosis after standard short-read NGS. This “diagnostic gap”

has directed attention toward structural variants, which represent a richer and more complex layer of genomic variation.

1.3.1. Definition and spectrum of structural variants

Structural variants (SVs) are typically defined as genomic rearrangements larger than 50 base pairs. They include deletions, duplications, inversions, insertions, translocations and more complex rearrangements that combine several of these event types (Figure 1.3).

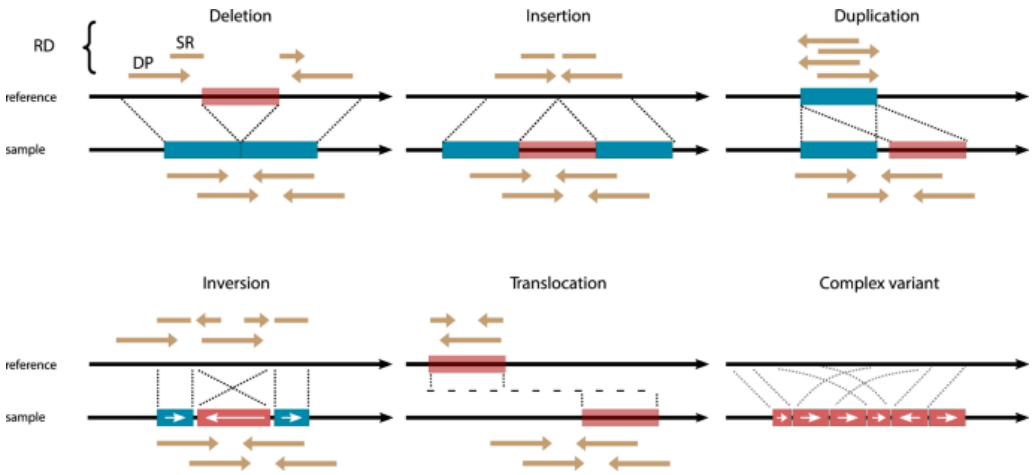

Figure 1.3: Most common SVs types: Deletion, the loss of a DNA segment; Insertion, the addition of a DNA segment; Duplication: the amplification of a DNA segment; Inversion, the flipping of a DNA segment; Translocation: a cut-and-paste of a DNA segment; Other Complex events which may derive from combination of the previously described events [17].

These variants reshape the local or large-scale architecture of chromosomes and therefore involve more DNA bases per event than SNVs and short indels. Population studies indicate that structural variation affects a substantial fraction of the human genome, and that the number of base pairs that differ between two individuals because of SVs exceeds the number contributed by SNVs [18].

Early genome surveys using array technologies and low-coverage whole-genome sequencing revealed thousands of common SVs per individual genome and showed that copy-number variable segments alone can cover several percent of the reference sequence. For example, high-density microarrays detected on the order of three to four thousand CNVs per human population and estimated that these regions account for approximately four to six percent of the genome [19]. More recent

sequencing-based maps have refined these estimates and confirmed that structural variation is pervasive and contributes substantially to human genomic diversity [20].

1.3.2. Mechanisms generating structural variants

Multiple molecular mechanisms can generate SVs, often influenced by local genome architecture. A first class of events arises through non-allelic homologous recombination (NAHR), in which misalignment and crossing over occur between segmental duplications or other low-copy repeats that share high sequence identity. This process typically produces recurrent deletions, duplications and inversions with stereotyped breakpoints, such as those observed at genomic disorder loci flanked by large repeat blocks [21].

A second class involves non-homologous end joining (NHEJ) and related error-prone repair pathways that join double-strand breaks with little or no sequence homology. These mechanisms are often implicated in nonrecurrent rearrangements and complex breakpoint patterns, particularly in cancer genomes where DNA damage and replication stress are frequent [22].

A third group of mechanisms is replication based. Fork stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR) have been proposed to explain structural variant architectures that combine duplications, deletions and inversions on the same haplotype. In these models, stalled replication forks repeatedly switch templates, guided by short stretches of microhomology, and can generate complex multi-step rearrangements in a single mutational event [23].

Finally, retrotransposition of mobile elements adds another route to structural variation. In humans, autonomous LINE-1 (L1) elements can copy and paste themselves to new genomic locations through target-primed reverse transcription. An L1 transcript is reverse transcribed at a staggered nick in the target site, and integration usually leaves a short direct target-site duplication flanking the new insertion. These events introduce several kilobases of new sequence, can be

accompanied by local deletions or rearrangements at the insertion site and therefore contribute both to copy number gain and to more complex genomic reorganization [23].

Together, NAHR, NHEJ, replication-based mechanisms and retrotransposition, as summarized in Figure 1.4, provide a unifying framework for understanding how both recurrent genomic disorders and heterogeneous non-recurrent rearrangements arise in the human genome.

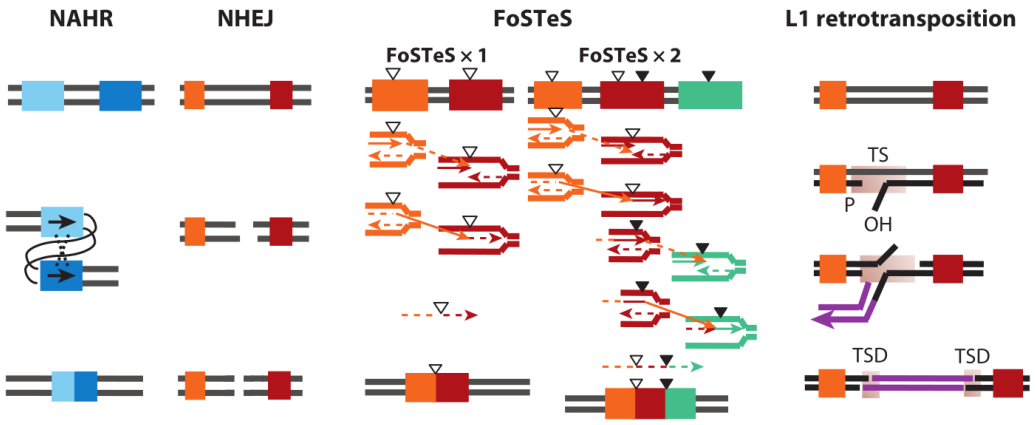

Figure 1.4: Four mechanisms generating human genomic rearrangements/CNVs: (i) non-allelic homologous recombination between repeats (LCRs/SDs, Alu, L1); (ii) non-homologous end-joining after double-strand breaks; (iii) fork stalling and template switching (single events → simple; multiple → complex); and (iv) retrotransposition, typically leaving a target site (TS) with target-site duplication (TSD) [23].

1.3.3. Structural variants in human disease and clinical diagnostics

The contribution of SVs to human disease is now well established. Dosage-sensitive genes disrupted by deletions or duplications underlie many genomic disorders, including microdeletion and microduplication syndromes associated with neurodevelopmental delay, congenital malformations and epilepsy. More subtle exonic rearrangements, such as multi-exon deletions or tandem duplications within single genes, are increasingly recognized as recurrent causes of monogenic disease, particularly in genes with complex local architecture [24].

In population genetics, SVs influence gene expression, local recombination, and selection, thereby shaping both individual phenotypes and evolutionary trajectories of human populations [18], [25].

Despite this functional importance, many clinical sequencing pipelines still prioritize SNVs and short indels, while SVs are either detected only partially or not at all. This creates a paradox in which the class of variants affecting the largest fraction of the genome is under-represented in diagnostic workflows. Integrating robust SV detection into routine analysis therefore represents a key opportunity to further increase diagnostic yield beyond that achievable with SNVs and indels alone.

1.3.4. Challenges of detecting SVs from short-read sequencing

Short-read NGS data provide the basis for most current clinical sequencing assays, yet they are intrinsically suboptimal for SV detection. Typical read lengths of 100-150 base pairs are insufficient to span many rearrangements and frequently cannot be mapped unambiguously in repetitive or segmentally duplicated regions. This limitation has motivated a rich ecosystem of SV calling algorithms that exploit multiple signatures, such as discordant read pairs, split reads, depth of coverage and local de novo assembly (Figure 1.5) [26].

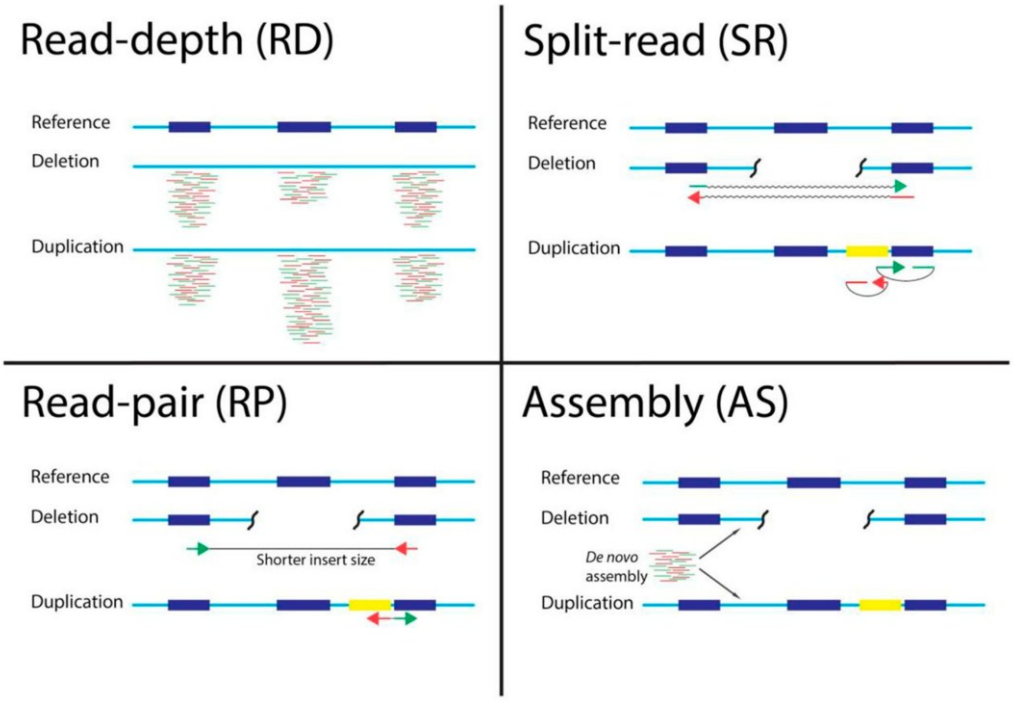

Figure 1.5: Overview of the four main algorithmic strategies for structural variant and CNV detection from NGS data. Read-depth (RD) methods infer deletions and duplications from local decreases or increases in sequencing coverage; split-read (SR) approaches detect breakpoints by identifying reads that align across junctions; read-pair (RP) strategies use discordant insert size and orientation of paired-end reads to support structural rearrangements; and assembly-based (AS) methods reconstruct sequences de novo to identify novel breakpoints and complex events [27].

Systematic evaluations using simulated and real whole-genome sequencing datasets have shown that no single short-read SV caller performs optimally across all variant types and size ranges. Recall decreases sharply in repetitive regions and for small to intermediate-sized SVs, whereas precision drops when sensitivity is pushed higher. Combining several algorithms can improve performance, but at the cost of increased complexity, computation time and challenges in merging partially concordant callsets [26], [28].

In addition, accurate breakpoint resolution is often difficult in complex loci, and SV genotyping across cohorts remains inconsistent. Recent comparative analyses that contrast short-read and long-read based SV detection have highlighted substantial blind spots of short-read methods, especially for insertions, inversions and SVs overlapping low-complexity or highly repetitive regions [29]. From a clinical perspective, these technical constraints imply that short-read WES and WGS

pipelines capture only a subset of clinically relevant SVs and that additional strategies are needed in order to systematically exploit this class of variation.

1.3.5. Copy-number variants: a structurally simple but clinically rich SV class

Among all structural variants, copy-number variants occupy a central position in clinical genetics, because they are both relatively accessible to current technologies and highly enriched for pathogenic events in many disorders.

1.3.5.1. Definition, genomic distribution and functional impact of CNVs

CNVs are typically defined as deletions or duplications of genomic segments larger than one kilobase, although technical definitions now often include events above 50 base pairs that alter the number of copies of a given region (Figure 1.6).

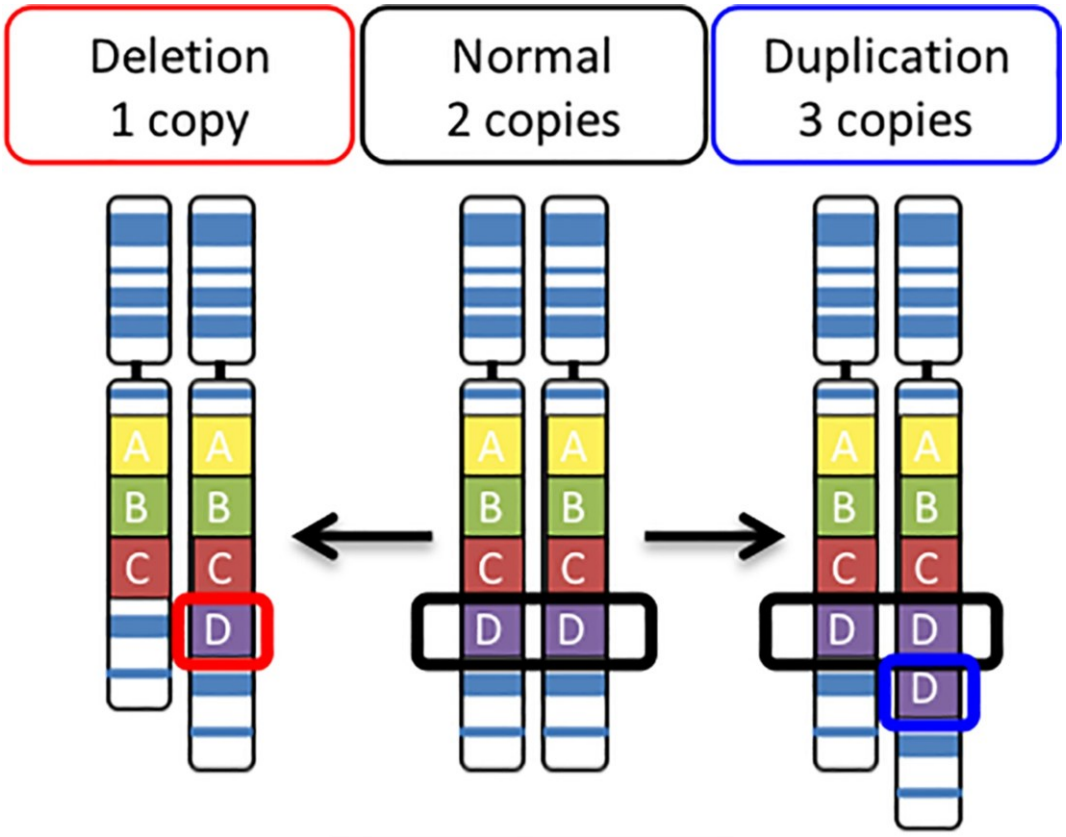

Figure 1.6: Copy Number Variants (CNV): Deletions (DEL) and Duplications (DUP) [30].

Early microarray and genome-wide surveys demonstrated that CNVs are ubiquitous and collectively cover several percent of the human genome, with each individual harboring hundreds to thousands of such events [19], [31].

CNVs can influence gene function through multiple mechanisms. Deletions may remove one or more exons, entire genes or regulatory elements, thereby causing haploinsufficiency or complete loss of function. Duplications can increase gene dosage, disrupt regulatory balance or create fusion transcripts when combined with other rearrangements. CNVs that do not overlap coding sequence can still modulate gene expression through effects on enhancers, insulators or higher-order chromatin architecture. Population-scale studies have shown that both common and rare CNVs contribute to adaptation, quantitative traits and disease susceptibility, and that many genes are recurrently affected by dosage-altering events [32].

In rare disease diagnostics, pathogenic CNVs frequently arise de novo, often with breakpoints in or near segmental duplications, and may span multiple genes or regulatory domains. Their size and genomic context can complicate interpretation, since relatively large events can encompass a mixture of dosage-sensitive and dosage-tolerant genes, and smaller events in apparently gene-poor regions may still disrupt long-range regulatory elements. These challenges have motivated the development of curated databases such as the Database of Genomic Variants (DGV), DECIPHER and ClinGen, which aggregate CNV frequency and clinical evidence in order to support interpretation [33], [34], [35].

1.3.5.2. CNVs in clinical diagnostics and the role of chromosomal microarrays

The clinical relevance of CNVs is reflected in the central role of chromosomal microarray analysis (CMA) in diagnostic workflows. Consensus guidelines from the American College of Medical Genetics and Genomics established CMA as the first-tier test for individuals with unexplained developmental delay, intellectual disability, congenital anomalies or autism spectrum disorder, based on a diagnostic yield of approximately fifteen to twenty percent, compared with about three percent for conventional karyotyping [36].

CMA platforms based on array comparative genomic hybridization or SNP arrays interrogate the genome at hundreds of thousands to millions of loci and can reliably

detect gains and losses at a resolution of a few hundred kilobases, or even down to single exons in custom designs. Clinical series across diverse populations have consistently confirmed the added value of CMA for patients with neurodevelopmental phenotypes, skeletal malformations and multiple congenital anomalies, and have documented its impact on clinical management and genetic counselling [37], [38].

Nevertheless, CMA has intrinsic limitations. It cannot reliably detect balanced rearrangements such as inversions or translocations, has reduced sensitivity in genomic regions with sparse probe coverage, and provides limited information on the exact breakpoints or allelic configuration of complex CNVs. In addition, targeted methods such as multiplex ligation-dependent probe amplification (MLPA) or quantitative PCR remain necessary for high-resolution interrogation of specific genes, for example in hereditary cancer panels or single-gene disorders.

1.3.5.3. CNV detection from short-read sequencing

Whole-genome sequencing (WGS) offers an attractive alternative to CMA for CNV detection, since it assays the genome in an unbiased manner and can, at least in principle, detect a broader spectrum of SVs. Read-depth based methods, often complemented by split-read and discordant-pair evidence, can identify deletions and duplications across a wide dynamic range of sizes, including small exonic events that are below the resolution of most microarrays. Benchmarking studies and large-scale reference projects have shown that high-coverage short-read WGS can recover most CNVs identifiable by CMA and additionally reveal many smaller or more complex events that were previously inaccessible [39].

However, as discussed above for SVs more broadly, WGS-based CNV detection still suffers from reduced sensitivity and specificity in repetitive or low-mappability regions.

Whole-exome sequencing occupies a unique position in this landscape because it is already widely used as a front-line test for SNV and indel discovery in genetically heterogeneous disorders. This has naturally prompted the question of whether WES data could also be exploited to detect clinically relevant CNVs, thereby maximizing the diagnostic information extracted from a single assay.

WES captures only the coding fraction of the genome and yields highly variable coverage driven by capture efficiency, GC content and local sequence context. CNV detection from WES is therefore typically based on read-depth information aggregated across exons [40].

However, independent benchmarking has highlighted important limitations. Tan et al. systematically evaluated four WES-based CNV tools and showed that sensitivity, specificity and breakpoint accuracy varied widely between methods and across CNV size ranges, with none of the tools achieving consistently high performance [41]. Zare et al. focused on WES CNV calling in cancer and reported moderate sensitivity but relatively high false discovery rates, particularly for small events and in noisy tumor data, underscoring the need for improved normalization and segmentation strategies [42]. Similarly, Gordeeva et al. benchmarked a broad set of germline exome CNV callers using an exon-level standard for NA12878 and observed low concordance between tools and marked precision–recall trade-offs, indicating that no single method delivers consistently robust performance across event types and size ranges [43]. Overall, these studies agree that while exome-based CNV calling is clinically useful, its performance is highly tool- and context-dependent.

1.3.5.4. Exome-specific challenges and the importance of study design

Several properties of WES data impose fundamental constraints on CNV detection that must be considered when designing pipelines and interpreting results. First, capture targets are discontinuous and restricted to exons, which means that intergenic and many intronic CNVs remain invisible, and that breakpoints are typically only coarsely localized to intervals between captured regions. Second, coverage is intrinsically heterogeneous. Differences in probe design, local GC content and hybridization efficiency create systematic biases that vary across both targets and sequencing runs. Robust CNV calling therefore requires careful normalization across a sufficiently large cohort of samples that have been processed with comparable library preparation and sequencing protocols [41], [42], [44].

Third, exome capture designs differ between laboratories, vendors and kit versions. This complicates the use of public reference cohorts and makes cross-study

comparison of CNV callsets difficult. In addition, the sparsity of exonic targets means that small CNVs affecting only one or two exons may be supported by very few data points, which increases both stochastic noise and the risk of artefacts due to mapping or local sequence features.

As a result, exome-based CNV detection is highly sensitive to study design choices, including the size and composition of the reference cohort, batch structure, capture kit, sequencing depth and mapping pipeline. Diagnostic laboratories that deploy WES CNV calling often invest substantial effort in defining internal quality control metrics, curating high-confidence training sets and re-validating calls with orthogonal methods such as CMA or MLPA, especially for small or borderline events.

1.4.Long molecule approaches for structural variant detection

Standard short read WES and WGS remain the backbone of diagnostic genomics, but they struggle with many classes of structural variation. Complex rearrangements, repeat mediated events and variants mapping to highly homologous loci are often only partially resolved or completely missed when reads are a few hundred bases long and scattered across the genome. This gap has motivated the development of sequencing and mapping technologies that preserve long range information, either by sequencing long DNA molecules directly or by imaging them at high resolution [45].

Long read sequencing (LRS) platforms such as Pacific Biosciences and Oxford Nanopore Technologies generate reads that span tens to hundreds of kilobases. These long fragments allow direct observation of large insertions and deletions, inversions, translocations and repeat expansions, and they support haplotype aware analysis because many heterozygous variants fall on the same read. Reviews of LRS consistently show that it improves the detection and interpretation of structural variants, enables assembly of nearly complete human genomes and helps resolve previously intractable regions such as centromeres and segmental duplications [45], [46].

Clinically, several studies now demonstrate that LRS increases diagnostic yield when applied to patients who remain undiagnosed after short read testing. In a recent rare disease cohort that had already undergone negative exome or genome sequencing, adding long read genome sequencing led to additional molecular diagnoses in about ten percent of cases, often through detection of complex structural rearrangements or repeat mediated events that short reads had missed [47]. This confirms that long molecule information can translate into clinically relevant findings. At the same time, long read platforms still come with higher per genome cost, more demanding DNA input requirements and dedicated instrumentation, which currently limit their use in high throughput diagnostic programs compared with short read WGS [46].

Optical genome mapping (OGM) takes a complementary approach. Instead of sequencing bases, ultra-high molecular weight DNA is fluorescently labelled and linearized in nanochannels. The resulting images are converted into genome wide maps of label positions, which can be compared with a reference to detect balanced and unbalanced chromosomal alterations, including large insertions and deletions, inversions, translocations and complex rearrangements. Recent reviews emphasize that OGM provides genome wide detection of structural and copy number variants at kilobase scale resolution and can consolidate several classical cytogenetic techniques into a single assay [48]. In parallel with OGM, several sequencing based strategies have been developed to enrich short read data with long range information. Linked read technologies such as the 10x Genomics Chromium system partition high molecular weight DNA into microdroplets, where molecules are tagged with a shared barcode before fragmentation and standard short read sequencing. After alignment, reads that share the same barcode can be grouped into synthetic long fragments of tens of kilobases, which supports haplotype phasing and structural variant discovery while still relying on Illumina sequencing chemistry [49]. Another group of methods exploits three dimensional genome organization to infer structural variation. Chromatin conformation capture and in particular Hi C generate genome wide contact maps that can be mined for signatures of balanced and unbalanced rearrangements. In tumor samples, Hi C contact matrices have been used to detect translocations and complex chromosomal

rearrangements, often with breakpoint level precision, while simultaneously deriving copy number profiles from the same data [50].

Collectively, these technologies illustrate a broad effort to move beyond classical short read WGS by embedding long range information into sequencing or imaging assays, although each of them has its own tradeoffs in terms of cost, throughput, variant classes covered and suitability for routine diagnostics.

1.5. Illumina Constellation mapped read as a bridge between short and long reads

Illumina Constellation mapped read technology was conceived precisely to address this gap. Instead of relying on conventional library preparation, Constellation mapped read performs tagmentation directly on the flow cell. Intact double stranded DNA is loaded into a modified NovaSeq X cartridge, where surface bound transposomes fragment and capture long template molecules within the nanowell array. Standard cluster generation and paired end 2 × 150 base sequencing are then performed, but the approximate physical position of each cluster on the flow cell is recorded and used by DRAGEN algorithms to infer which read pairs originate from the same original DNA molecule [51].

This spatial proximity information effectively reconstructs long interspersed templates from sets of short reads. Illumina reports that these templates routinely extend into the hundreds of kilobases and that proximity aware mapping allows reads in ambiguous or repetitive regions to be anchored by neighboring clusters that map uniquely Figure 1.7. As a result, coverage and variant calling performance improve in medically relevant genes that are difficult to interrogate with standard WGS [51].

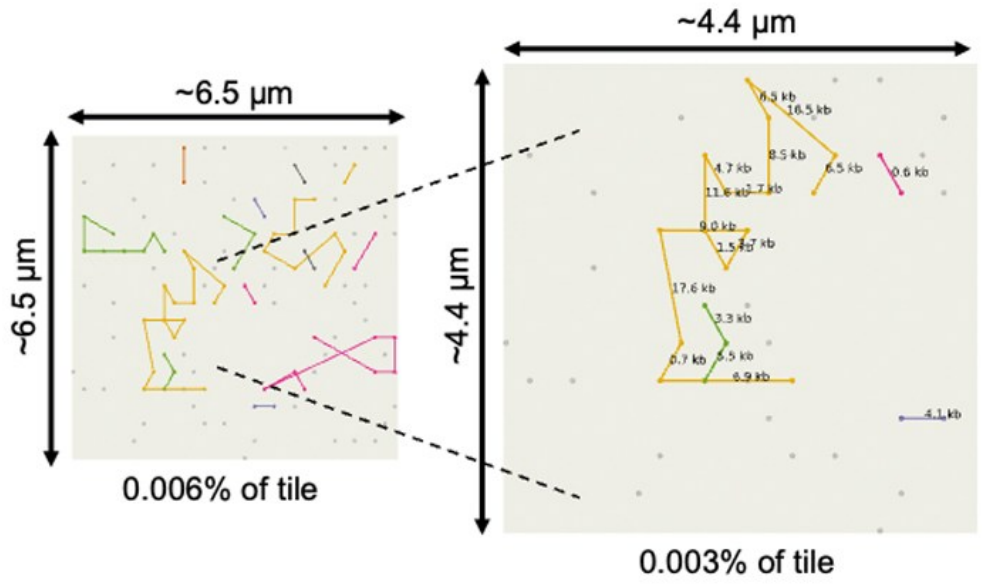

Figure 1.7: Schematic of the Constellation mapped read technology. Resulting clusters originating from the same DNA template remain physically close on the flow cell surface. Image adapted from [51].

A key consequence of this design is that Constellation addresses both small and large variant classes within a single whole genome workflow. For single nucleotide variants and short indels, the underlying data are still short read SBS. Benchmarking on Genome in a Bottle reference genomes shows that Constellation achieves very high F1 scores for small variants, including in challenging genomic contexts, and that proximity informed mapping markedly reduces both false positives and false negatives compared with conventional short read WGS using the same reference truth sets [51].

Structural variant calling also benefits from the added proximity signal. In internal benchmarking using DRAGEN v4.3 and Genome in a Bottle T2T-Q100 structural variant truth sets, Illumina reports that Constellation increases recall for structural variants larger than fifty base pairs from about 51.5 percent with standard short read WGS to roughly 87.8 percent, while following established GIAB best practices for SV evaluation. These performance gains extend beyond simple insertion and deletion events and include the ability to visualize complex rearrangements through so called colocation plots, which summarize proximity contacts between pairs of

genomic bins and reveal off diagonal patterns when the sample genome deviates from the reference [51].

In summary, Constellation mapped read bring long range information, improved mapping in difficult regions, high accuracy for small variants and substantially enhanced structural variant detection into a short read whole genome framework. This makes Constellation an attractive candidate for future diagnostic workflows aiming to unify variant discovery into one comprehensive test, and it provides the conceptual basis for comparing CNV detection from conventional WES with CNV and broader structural variant detection from Constellation WGS in the present thesis.

2. Aim

Despite the widespread use of short-read whole-exome (WES) and whole-genome sequencing (WGS), many patients with suspected genetic disorders still lack a molecular diagnosis. Routine pipelines are optimized for SNVs and small INDELs, whereas CNVs and other SVs are often assessed with separate assays or only in selected cases, fragmenting the workflow and leaving much structural variation underused. In our WES cohort, CNV calling generated large, unstable callsets that were difficult to validate and prioritize, so the first part of this thesis dissects these problems and reshapes exome-based CNV analysis into a smaller, more reliable set of clinically manageable events. In parallel, the thesis evaluates Illumina Constellation mapped read, a whole-genome approach that adds long-range information to standard short-read sequencing in order to move towards a single assay for SNVs, INDELs, CNVs and complex SVs on existing Illumina infrastructure.

The work is organized around two main aims.

1) Optimizing exome-based CNV calling for clinical use by:
    - systematically comparing different WES CNV callers on the same cohort and characterize their main discrepancies,
    - evaluating how different reference ("baseline") designs affect sensitivity, noise and reproducibility,
    - building a meta-caller that integrates tool outputs to enrich for high-confidence CNVs, and
    - implementing an annotation and prioritization strategy that reduces the number of CNVs requiring manual review while preserving maximal sensitivity.

2) Evaluating Illumina Constellation mapped read as a near "all-in-one" genome test by:

- comparing Constellation mapped read with standard PCR-free WGS in terms of coverage, callable genome and small-variant performance,
- quantifying phasing performance and the effective length of reconstructed haplotypes, and
- assessing CNV and SV detection from Constellation mapped read in a set of structurally characterized cases and explore whether proximity-based views help interpret complex rearrangements.

Together, these aims are intended to clarify how far short-read-based approaches can be pushed toward comprehensive variant detection and how close Constellation mapped read can come to a single, clinically useful genome assay.

3. Methods

3.1. Whole Exome Sequencing Dataset

3.1.1. CNVPANEL01 dataset

The CNVPANEL01 dataset analyzed in this study consists of two cohorts derived from publicly available Coriell Institute reference materials: a Normal cohort and the CNVPANEL01 cohort.

The Normal cohort (n = 12) includes well-characterized HapMap and 1000 Genomes Project samples (NA12877, NA12879, NA19017, NA18525, NA10851, NA18939, NA19625, NA20845, NA20502, HG00096, HG00268, NA12878). These samples serve as high-quality references for CNV calling.

The CNVPANEL01 cohort (n = 43) comprises lymphoblastoid cell lines with validated chromosomal abnormalities, selected from the NIGMS Human Genetic Cell Repository at Coriell [52]. This panel includes a diverse collection of genomes carrying known copy number alterations, such as partial or complete deletions, duplications, and unbalanced translocations. Each sample is accompanied by cytogenetic annotations in ISCN format, providing an orthogonal ground truth for CNV validation. The set includes the samples shown in Table 3.1.

SAMPLE ID	DISORDER	CNVS
GM01416	XXXX Syndrome	chrX:251797-155952677 DUP
GM05067	Aneuploid Chromosome Number - Trisomy 9	chr9:46586-39778619 DUP
GM05966	Derivative Chromosome	chr14:54502047-75681317 DUP
GM06226	Translocated Chromosome	chr1:247143229-248930177 DEL; chr16:35814-21378649 DUP
GM06870	Aneuploid Chromosome Number - Non-Trisomic	chr18:11542-15401752 DUP
GM06936	Chromosome Deletion	chr10:58486-12836926 DEL
GM07945	Adenosine Deaminase Deficiency With No Immunodeficiency	chr13:20228923-20460156 DEL; chr20:34910451-46231832 DEL
GM08331	Chromosome Deletion	chr13:97506715-109611222 DEL; chr21:25943808-28146869 DEL
GM09102	Chromosome Deletion	chr11:120620269-135074876 DEL
GM09216	Chromosome Deletion	chr2:10203410-26929010 DEL;

		chr4:143920938-144021781 DEL
GM09367	Duplicated Chromosome	chr6:107433158-142743017 DUP
GM09888	Trichorhinophalangeal Syndrome, Type II	chr8:106107809-118282364 DEL; chr14:50355382-51188531 DEL
GM10608	Chromosome Deletion	chr20:9891954-18050825 DEL
GM10636	Duplicated Chromosome	chr2:109704906-110869845 DUP; chrX:39949336-57930356 DUP
GM10800	Chromosome Deletion	chr4:69196130-94156942 DEL
GM10925	Greig Cephalopolysyndactyly Syndrome	chr7:38592416-54646811 DEL; chrX:372539-654184 DUP
GM10985	Chromosome Deletion	chr3:18654-10288693 DEL
GM10989	Gilles De La Tourette Syndrome	chr9:46586-11996831 DEL
GM11213	Chromosome Deletion	chr2:186245475-203738206 DEL
GM11419	Aneuploid Chromosome Number - Non-Trisomic	chr4:143899562-144129458 DUP; chrY:2782384-26653776 DUP
GM11672	Chromosome Deletion	chr10:48084407-73690949 DEL
GM12606	Chromosome Deletion	chr13:18471487-59667287 DUP
GM12662	Chromosome Deletion	chrY:22213783-26541710 DEL
GM13019	Turner Syndrome	chrX:251797-56869151 DEL; chrX:63123908-154822179 DUP
GM13464	Williams-Beuren Syndrome	chr7:73311765-74727754 DEL
GM13476	Smith-Magenis Syndrome	chr17:16860240-20492562 DEL
GM14164	Tetralogy Of Fallot	chr13:47227948-95062722 DEL
GM14485	Inverted Duplication Deletion	chr8:220289-7368769 DEL; chr8:12670906-43745225 DUP
GM14943	Chromosome Deletion	chr2:234368396-242147293 DEL
GM16362	Aneuploid Chromosome Number - Trisomy	chr22:16986520-19090414 DUP; chr22:21960704-22219583 DEL; chr22:42883607-50796015 DUP
GM16595	Cri-du-cha Syndrome	chr5:8633692-24036533 DEL
GM17867	Klinefelter Syndrome	chrX:251797-155699825
GM17942	DiGeorge Syndrome	chr22:18167914-21108322 DEL
GM20022	Duplicated Chromosome	chr3:134843252-195926102 DUP
GM20027	Turner Syndrome	chrX:251797-156004181 DEL
GM20556	Isodicentric Chromosome	chr15:19811062-32478247 DUP
GM21698	Chromosome Deletion	chr6:162519205-170610395 DEL
GM21699	Chromosome Deletion	chr3:18654-564371 DUP; chr6:163241020-170673434 DEL
GM21887	Angelman Syndrome	chr15:17167687-23199681 DEL
GM22601	Wolf-Hirschhorn Syndrome	chr4:65772-25980331 DEL
GM22624	Potocki-Shaffer Syndrome	chr11:40455217-46053197 DEL
GM22991	Chromosome 1P36 Deletion	chr1:817185-5250920 DEL

Table 3.1: Coriell Institute reference samples used for validation of the CNV calling. For each Sample ID, the associated disorder and known CNVs are reported with genomic coordinates in the format chr[chrom]:start-end. Variants are annotated as deletions (DEL) or duplications (DUP).

All samples were prepared using the Twist Library Preparation EF Kit 2.0, combined with the Twist Universal Adapter System and the Twist Target Enrichment Standard Hybridization v2 workflow. Libraries were sequenced on the Illumina NovaSeq 6000 platform, generating paired-end 150 bp (PE150) reads. Demultiplexed FASTQ files were provided directly by Twist Bioscience and were used as input for the downstream processing [53].

This dataset, encompassing a balanced mixture of normal reference and cytogenetically validated abnormal samples, provides an ideal framework for evaluating CNV detection accuracy and overall assay reproducibility within hybrid-capture sequencing workflows.

3.1.2. Burlo dataset

The second dataset analyzed in this study consisted of 87 samples, including 26 trios, 1 duo, and 6 single samples. All samples were processed at the Functional Genomics Laboratory following Twist Bioscience's protocols for library preparation and exome enrichment. Libraries were prepared using the Twist Library Preparation Enzymatic Fragmentation 2.0 Kit, which employs a controlled enzymatic fragmentation step to generate DNA fragments of optimal size for hybrid capture-based sequencing. This kit, combined with the Twist Universal Adapter System, provides an integrated workflow for efficient library construction with minimal hands-on time. Exome enrichment was performed using the Twist Exome 2.0 Plus Comprehensive Exome Spike-in panel (hereafter referred to as *Twist Exome*), which provides extended coverage of exonic regions. Hybridization and capture followed the Twist Target Enrichment Standard Hybridization v2 Protocol. Sequencing was performed on the Illumina NovaSeq 6000 platform, generating paired-end 150 bp (PE150) reads. Raw sequencing data were demultiplexed with Illumina bcl2fastq (v2.20) using unique dual indices (UDIs), allowing for up to one mismatch in barcode recognition [54].

3.1.3. WES alignment and alignment statistics

Sequencing adapters and low-quality bases were removed using fastp v0.21.0 [55]. Reads from each sample were aligned to the HG38 reference genome (chromosomes 1-22, X, Y, M) with BWA-MEM2 v2.2.1 [56], and the resulting SAM files were converted to BAM format using SAMtools v1.11 [57]. PCR and optical duplicates were marked and removed with Picard v2.17.11[58], and base quality score recalibration was performed with GATK BaseRecalibrator v4.5.0.0 [59]. The processing pipeline also included clipping of overlapping read pairs using BamUtil v1.4.14 [60] and computation of insert-size statistics with Picard CollectInsertSizeMetrics v2.17.10 [58].

Coverage and genotypability over the Twist target regions were assessed with bedtools coverage [61] and GATK CallableLoci v3.8 [8]. Hybrid-capture performance metrics, including Fold80 base penalty and the percentages of on-target, near-target and off-target bases, were calculated with Picard CollectHsMetrics v2.17.11 [58].

3.2.CNV calling software

3.2.1. EXCAVATOR2

EXCAVATOR2 is a read-depth-based CNV caller for WES and targeted sequencing that jointly uses on-target and off-target reads. After normalising read counts for technical biases, it segments the genome into regions of homogeneous copy number and classifies them as deletions or duplications [62]. In this thesis it is used only to provide an example batch of historical CNV calls and is not included in the systematic benchmarking, since it is no longer actively maintained and has been superseded by more recent exome CNV callers.

3.2.2. ExomeDepth

ExomeDepth is a Bayesian read-depth-based CNV caller designed specifically for targeted sequencing assays, including whole-exome sequencing (WES) and clinical gene panels. The method models the observed read counts across capture targets using a synthetic reference profile constructed from a set of normal samples [63].

For each target region, ExomeDepth fits a beta-binomial model that estimates the likelihood of the observed read depth under two competing hypotheses: copy-neutral state versus copy-number alteration. The beta-binomial formulation accounts for overdispersion typically present in WES data. CNV calling is then performed using a hidden Markov model (HMM) that segments the genome into contiguous intervals of constant copy number, smoothing local fluctuations and enforcing biologically plausible transitions between states (e.g., diploid → deletion → diploid) [63].

The output includes CNV coordinates, predicted copy-number state, Bayes Factors, and expected read-depth ratios in a CNV format.

3.2.3. ClinCNV

ClinCNV is a multi-sample CNV caller tailored for clinical diagnostics and large targeted sequencing cohorts [64]. ClinCNV leverages information from the entire dataset to estimate expected coverage behavior and detect departures indicative of copy-number abnormalities. The algorithm operates in several stages.

First, clinCNV performs GC-content correction, library-size normalization, and variance stabilization across all samples to reduce technical biases inherent to targeted sequencing. It then models the normalized coverage using multi-sample statistical learning: for each target region, the distribution of expected read depth is estimated from the cohort, allowing the identification of samples that deviate significantly from the population baseline.

Detection of CNVs is performed using a robust segmentation algorithm that integrates deviations across neighboring targets and controls for local variability. The method supports both paired and unpaired designs, but is particularly effective in large cohorts where the distribution of copy-number states can be inferred directly from the data. ClinCNV also incorporates a probabilistic mixture model to refine copy-number state assignments, especially in regions affected by systematic noise or variable capture efficiency [64].

3.2.4. gCNV (GATK Germline CNV Caller)

gCNV, part of the GATK suite, is a generative probabilistic model for germline CNV detection that relies on hierarchical Bayesian inference. It is designed to operate on large cohorts of WES or targeted sequencing samples and uses latent factors to model systematic sources of variation in read depth [65].

The method begins by constructing a read-count matrix across all samples and targets, followed by a series of normalization steps, including GC correction and sample-specific scaling. gCNV then learns a linear latent representation of coverage variability using probabilistic matrix factorization. These latent factors capture batch effects, capture-efficiency fluctuations, and other technical noise sources that cannot be eliminated through standard normalization alone [65].

CNV detection is performed by comparing each sample's latent-normalized read depth to the inferred reference distribution, using a Bayesian hidden Markov model to segment the genome into discrete copy-number states. The hierarchical nature of the model allows information sharing across samples, improving robustness in the presence of heterogeneous cohorts or variable sequencing quality. gCNV produces posterior probabilities for each state, allowing confident discrimination between deletions, duplications, and diploid regions [65].

3.2.5. Sensitivity and coverage correlation

The evaluation of WES-derived CNVs was constrained by the structure of the available gold standard. For most samples only one or two CNVs had been validated with orthogonal methods, which did not allow a reliable estimation of false positives, specificity or precision. For this reason, performance was quantified primarily in terms of sensitivity (recall), using two complementary definitions: base-level sensitivity and target-level sensitivity.

$$Sensitivity = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Base-level sensitivity was defined as the proportion of genomic bases within validated CNVs that were recovered by the caller. For each validated event, I calculated the number of bases overlapping at least one called CNV of the same type (deletion or duplication) and divided this by the total length of the validated CNV. Aggregating these values across all events provided an overall base-level sensitivity. This metric captures how much of each validated event is covered by the calls and how closely the detected CNVs approach the true breakpoints, both at the single-event and at the global level.

$$\text{Base} - \text{level Sensitivity} = \frac{\text{True Positive bases}}{\text{True Positive bases} + \text{False Negative bases}}$$

Target-level sensitivity was defined with respect to the exome capture design. For each validated CNV, I identified all target regions (e.g. capture intervals) that lay within the validated coordinates and counted how many of these targets were correctly assigned to a CNV of the appropriate type. Target-level sensitivity was computed as the fraction of validated targets that were recovered. While the base-level sensitivity focuses on the continuity and breakpoint accuracy of each event, the target-level sensitivity reflects how completely the callable portion of each validated CNV (i.e. the covered targets) is correctly classified. The two metrics are therefore broadly complementary: base-level sensitivity is more informative about breakpoint proximity, whereas target-level sensitivity summarizes how much of the "callable content" of each CNV is captured.

$$\text{Target} - \text{level Sensitivity} = \frac{\text{True Positive targets}}{\text{True Positive targets} + \text{False Negative targets}}$$

In addition to these sensitivity measures, Pearson correlation of read-depth profiles was used to evaluate and optimize baseline (reference) selection strategies. For each sample, we computed the Pearson correlation between its vector of read depth across targets and that of each candidate reference sample, and selected the X "nearest neighbors" with the highest correlation to construct the baseline. The resulting distributions of within-baseline correlations, and their variability across

runs and strategies, were compared to identify baseline designs that produced more homogeneous and stable coverage profiles, which are expected to be more favorable for read-depth-based CNV calling.

$$r = \frac{\sum_i (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

3.2.6. Quantification of GC-associated coverage variability

To quantitatively assess the contribution of GC content to coverage variability, normalized coverage was modeled independently for each sample as a function of target GC content. Given the clearly non-linear relationship observed in exploratory analyses, a third-degree polynomial regression was adopted as a parsimonious model capable of capturing the broad curvature of the GC-dependent trend while avoiding overparameterization. For each sample, normalized coverage values were regressed against the GC content of target regions using a third-degree polynomial model. The coefficient of determination (R2) of the fitted model was used as a sample-level estimate of the proportion of normalized coverage variability explained by GC content. This approach was intended as a descriptive summary of GC-associated coverage variability rather than a mechanistic model of GC bias, enabling direct comparison of the strength of GC-dependent effects across samples and datasets.

3.3.Snakemake

Snakemake was used as the workflow management system to orchestrate all bioinformatic analyses in this thesis, providing a reproducible and scalable framework that links individual tools into a directed acyclic graph of rules, where each rule specifies its input and output files together with the command line needed to transform the former into the latter [66]. This rule based representation allows Snakemake to infer dependencies automatically from filenames, to parallelize independent jobs across available CPU cores or on a high performance computing cluster, and to resume interrupted runs without recomputing completed steps [67].

In this work, separate but coordinated workflows were implemented to handle whole exome and whole genome data, including read alignment, quality control, variant calling, copy number analysis and downstream summarization, while a shared configuration file and consistent naming of samples and targets ensured that the same pipeline could be applied uniformly across different cohorts, capture designs and sequencing assays.

3.4.Constellation Mapped Read Whole Genome Sequencing dataset

3.4.1. Study cohort and known structural variants

A second whole genome sequencing dataset used in this thesis was generated with the Illumina Constellation Mapped Read technology [51]. The cohort comprised six clinically characterised genomes from Meyer Children's Hospital (Florence, Italy). For each case, a structural alteration had been previously identified in routine diagnostics and was used here as a case specific truth set for benchmarking. The known events were: a complex rearrangement at *SCN1A* in X7670, including an approximately 56.6 kb tandem duplication with an internal deletion of about 2.6 kb within one of the duplicated segments; a heterozygous deletion of about 2 kb involving *STXBP1* in X7673; a balanced translocation t(19;22)(q13.4;p11.2) together with a multi megabase duplication spanning 19q13.33-q13.43 in X7674; a balanced t(3;8)(q21.3;q22.1) in X7675; a tandem duplication on chromosome 18 on haplotype 2 in X7676; and, in X7677, a composite interchromosomal rearrangement between chromosomes 2 and 4 consisting of insertion of about 2 Mb from chromosome 2 into chromosome 4, a balanced reciprocal t(2;4), and a 17.7 Mb deletion on chromosome 4.

Where gene level information was available from prior clinical reports, alterations involved *SCN1A* (X7670) and *STXBP1* (X7673). The remaining events were described at cytoband or chromosome level due to their size and structural complexity. All alterations had been established by orthogonal assays, including MLPA, array CGH, karyotyping, whole exome or whole genome sequencing and long read approaches, and were not discovered de novo in this study. They were used exclusively to evaluate the ability of Constellation to detect, refine and phase clinically relevant structural variants.

3.4.2. Sample collection, DNA extraction and quality assessment

Genomic DNA was extracted from whole blood using either the QIAsymphony DSP kit (Qiagen) or the Bionano Ultra High Molecular Weight (UHMW) protocol, following the manufacturers' instructions [68], [69]. DNA concentration was measured with Qubit dsDNA assays (Thermo Fisher Scientific) [70]. Integrity and long fragment content were evaluated by pulsed field gel electrophoresis and by Agilent Genomic DNA TapeStation, which was used to quantify the fraction of fragments longer than 60 kb, as recommended for Constellation [51], [71]. For optimal proximity metrics the target was high molecular weight DNA with at least 70% of fragments above 60 kb, while samples with at least 10% of fragments above 60 kb were still accepted. For each sample and lane we prepared 350 ng of high molecular weight genomic DNA, with an acceptable range of roughly 250-450 ng; inputs down to 100 ng were tolerated, at the cost of reduced proximity depth according to Illumina's standard operating procedure.

3.4.3. Constellation sequencing on NovaSeq X Plus and run planning

Sequencing was performed at the Illumina Solution Center (Milan, Italy) on a NovaSeq X Plus instrument using 10B flow cells and a 300 cycle kit, with a Constellation specific custom recipe for 2×150 bp reads. Run planning was carried out in BaseSpace Sequence Hub Cloud Run Planner by selecting NovaSeq X Series and read lengths of 151, 0, 0, 151 [72]. One sample was loaded per lane, up to a maximum of eight lanes per flow cell, and adapter sequences were specified according to the run planner instructions.

On the instrument, the planned BaseSpace Sequence Hub run was selected and the Constellation custom recipe matching the NovaSeq X software version was uploaded. Upon completion of each run, data were automatically transferred to Illumina Connected Analytics (ICA) under the corresponding BaseSpace managed project [73]. Conversion from BCL to FASTQ and demultiplexing were performed

with Illumina DRAGEN BCL Convert 4.3.16, as defined in the BaseSpace configuration [7].

3.4.4. Constellation secondary analysis on Illumina Connected Analytics

FASTQ files were processed on ICA using the Constellation pipeline in development (TRIP-APP-KOL-Constellation_0-3_rc6) with default settings and the GRCh38 reference genome bundle [73], [74]. In brief, the pipeline exploits proximity information by linking reads that originate from nearby clusters on the flow cell and uses this information during alignment with the DRAGEN mapper. It then constructs a genome wide colocation matrix, calls small variants with the DRAGEN germline caller, performs haplotype phasing of reads and variants with WhatsHap, and finally runs DRAGEN modules for structural variant and copy number variant calling on haplotagged BAM files that incorporate proximity context.

The Constellation analysis produced, for each sample, haplotagged BAMs annotated with HP and PS phasing tags and Constellation specific tags, phased small variant VCFs, SV VCFs, CNV VCFs, and an additional CNV_SV VCF optimized for increased sensitivity below 10 kb. The pipeline also outputs a genome wide colocation matrix and a consolidated metrics report. For manual inspection in IGV, five files were used: the haplotagged BAM with its index, the phased VCF with its index, and a GTF file describing the phase blocks [74].

Constellation augments the BAM files with tags that encode proximity and phasing information, including xq (mapping quality with proximity linking), ps (pair score), PX (template identifier) and the standard HP and PS tags used by WhatsHap [75]. Metrics produced by the pipeline include the proximity rate at different quality thresholds (Q20, Q25, Q30), the mean_phred_link_quality, percentiles of template length and genomic span, the number of templates and subpairs, and a fit_phred_rmse statistic, with values below one indicating a good fit [74].

3.4.5. Standard PCR free whole genome sequencing and DRAGEN analysis

For methodological comparison, one sample from the Constellation cohort (X7675) had previously been sequenced as standard Illumina TruSeq DNA PCR free whole genome sequencing on a NovaSeq 6000 instrument. For this dataset, primary and secondary analysis were performed in BaseSpace with the DRAGEN_Germline_Whole_Genome 4.4.4 application [72], [75]. The workflow included mapping and alignment, duplicate marking, calling of single nucleotide variants and indels with VCF output, and calling of structural and copy number variants, including an additional CNV_SV output. Resulting VCFs and BAMs were downloaded from BaseSpace and used for head to head comparisons against the matched Constellation dataset.

For the evaluation of phasing performance, small variants in the TruSeq dataset were phased with WhatsHap version 2.8 [75], since the DRAGEN germline workflow does not provide variant phasing by default.

3.4.6. Quality control metrics, variant filtering and phasing evaluation

Quality control was carried out at both run and sample level by combining DRAGEN and Constellation metrics. Standard whole genome metrics included mean coverage, breadth of coverage at 1× and 10×, uniformity of coverage and callable genome fraction. For Constellation we additionally quantified the number of templates, the distributions of template length and genomic span, and the number of subpairs per template.

Insert size distributions for Constellation samples were computed with Picard CollectInsertSizeMetrics version 3.1.1 and visualized with a custom Python script based on matplotlib [58], [76]. Template level statistics, including the fraction of templates longer than 10, 20 and 60 kb, template length distributions and subpair counts, were extracted from the Constellation reports, in particular from the template_stats_all.csv file, and plotted with a custom Python script using seaborn.

Coverage, uniformity and variant level statistics were obtained from DRAGEN reports for both TruSeq and Constellation.

Phasing performance was evaluated using the tabular output of whatshap stats for both technologies. For all comparative analyses we considered only structural variants of length at least 50 bp, in line with common practice and with standard definitions used in the structural variation literature.

3.4.7. Visualization and manual review

Regional evidence for small variants, structural variants and phase block continuity was inspected in IGV using the haplotagged BAM and index, the phased VCF and index, and the phase block GTF file [77]. The genome wide colocation matrix produced by Constellation was explored with Higlass version 1.13.4 to support the interpretation of complex rearrangements and to relate proximity patterns to the underlying structural variation, following the recommendations in the Constellation visualization documentation [78].

4. Results

4.1.CNV from WES

4.1.1. Anomalous CNV Landscape in In-House Twist Exome Batch

In the March 2023 batch, 59 sequenced in-house samples with the Twist Exome 2.0 Plus Comprehensive Exome Spike-in design (hereafter, Twist Exome), I encountered an immediately unusual CNV landscape (Figure 4.1). Across 59 samples, EXCAVATOR2 was used as a CNV caller; this read-depth method explicitly exploits both on-target and off-target reads from WES to achieve genome-wide CNV resolution [62].

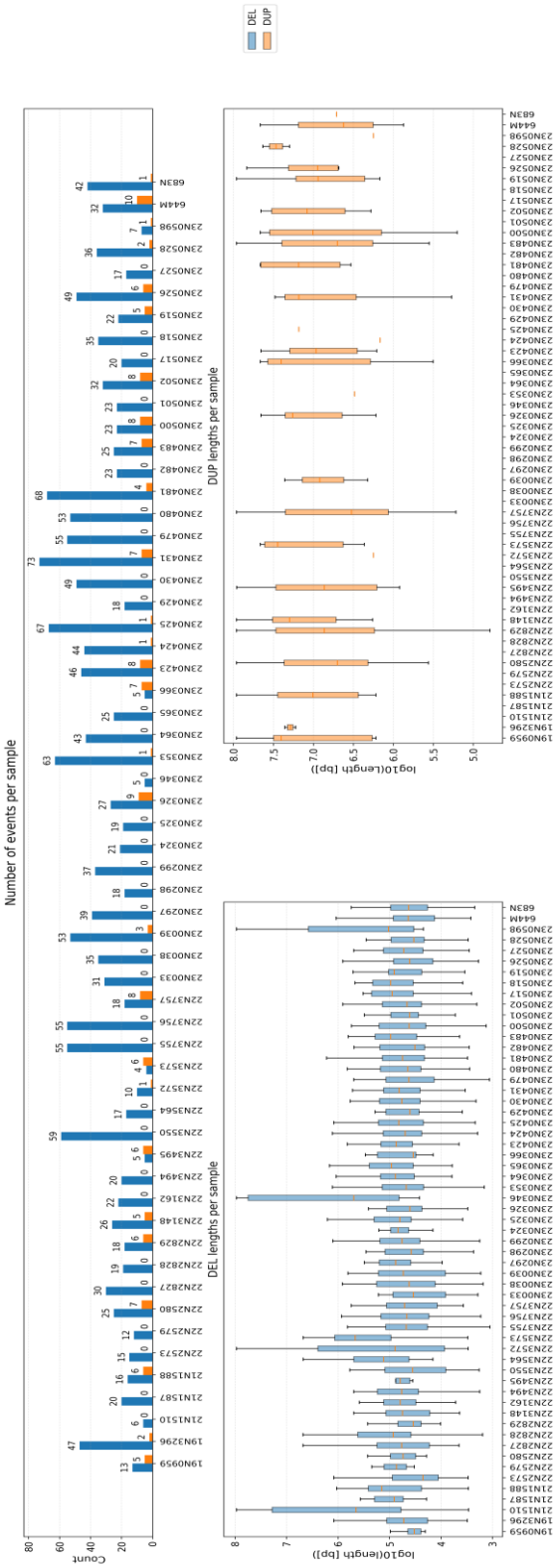

Figure 4.1: CNVs call performed with EXCAVATOR2 on an in-house WES batch captured with Twist Exome 2.0 Plus Comprehensive Exome Spike-in (HG19).

Aggregating all calls, I observed 1,792 deletions vs 141 duplications (a 12.7:1 DEL:DUP ratio), with a median of 25 deletions per sample (Inter Quartile Range, IQR: 18-43.5) but a median of 0 duplications (IQR 0-5.5). Fully 52.5% (31/59) of samples had no duplications at all, and among the 28 samples with any DUPs the median DUP count was 6. Length scales reinforced the asymmetry: per-sample DEL medians centered around 55.8 kb (IQR 41.9-76.5 kb), whereas per-sample DUP medians (where present) clustered around 12.55 Mb (IQR 5.00-18.57 Mb), i.e., ~188× longer than co-sample deletions on a median-to-median basis. Extremes illustrate the pattern: 23N0431 carried the most deletions (n=73), and ten samples exceeded 50 deletions; at the other end, some samples had few but very long deletions (e.g., 23N0346, n=5, mean 29.97 Mb; 21N1510, n=6, mean 25.51 Mb). Duplications, when present, were rare but extremely long: the largest per-sample DUP median reached 31.44 Mb (23N0528), and 5/28 duplication-bearing samples had DUP medians >20 Mb (12/28 had DUP means >20 Mb). Even the sample with the most duplications (644M, n=10) showed multi-megabase DUP segments (median 4.27 Mb).

Critically, the same broad CNV events recur across probands and their unaffected parents, and the skewed "few DUP, many DEL, both long" signature replicated across additional clinical batches (not shown). Together, these lines of evidence mark this figure as the first clear observation of a systematic issue and motivate the deeper investigations that follow.

4.1.2. Transition to ExomeDepth: A More Balanced yet Inflated CNV Landscape

Moving from the EXCAVATOR2 read-depth caller, where our March-2023 cohort had shown a clinically unhelpful pattern, we adopted ExomeDepth to analyze subsequent WES batches [63]. EXCAVATOR2 (first released in 2016) leverages off-target reads to extend genome-wide coverage from exomes, but the codebase is dated and, after contacting the original authors, we could not obtain ongoing support [62].

ExomeDepth, by contrast, models read counts with a beta-binomial; calls are then merged along targets (HMM) and accompanied by a Bayes Factor (BF) as a per-

event quality score, features that are explicitly documented in the original method paper, although no suggestions are provided on any filtering thresholds to use [79]. This choice is also consistent with recent literature: in 2022, ExomeDepth was selected as a screening tool for its well-balanced performance with high sensitivity in a clinical WES pipeline, and in 2023 a Genes cohort study reported the highest performance and accuracy for ExomeDepth among evaluated WES CNV callers (and subsequently applied it to >450 cases) [80], [81]. I analyzed 87 samples described in Table 3.1, which were processed in-house and coming from the Burlo Garofalo Hospital (Trieste, Italy).

	#fragments	**insert_size MEAN**	**%dupl**	**MEAN Coverage**	**%1X**	**%10X**	**%PASS**	**fold80**	**Uniformity of coverage (Pct > 0,2*mean)**
Mean	44.751.424	296,78	17,96	98,94	99,72	99,52	96,74	1,38	97,77
Median	48.366.906	296,32	17,92	105,00	99,78	99,48	96,72	1,40	97,78
St.Dev.	12.600.604	7,20	1,95	23,66	0,10	0,24	0,11	0,05	0,06

Table 4.1: : The average insert size, the percentage of duplicated reads, the percentage of genomic bases covered by at least 1X and 10X, the percentage of calleble bases, the fold80 (computed on non-zero coverage bases) and the Uniformity of coverage (% of bases coverage at least 20% of the mean coverage).

ExomeDepth showed a fundamentally changed the landscape: I observed 27,935 deletions and 14,496 duplications in total (DEL:DUP = 1.93:1), with per-sample medians of 199 deletions (IQR 162-314.5) and 158 duplications (IQR 127.5-191.5). Fully 66/87 samples had more deletions than duplications, but 21/87 showed the reverse, and the median per-sample DEL:DUP ratio was 1.39 (IQR 1.03-2.59), far more balanced than the EXCAVATOR2 profile (12.7:1, with >50% of samples having zero DUP). Event lengths collapsed from Mb-scale to kb-scale: across samples, median deletion length ≈1.03 kb (IQR 0.48-1.66 kb) and median duplication length ≈1.18 kb (IQR 0.74-1.80 kb), versus the earlier EXCAVATOR2 batch where typical medians were tens of kilobases for deletions and tens of megabases for duplications. This came with a cost: calls were much more

numerous, creating a review burden that was infeasible in a diagnostic workflow without stringent filtering. Moreover, I saw clear count outliers that dominate the call set, e.g., 22_3161 (1,127 DEL; 378 DUP), 22_3250 (1,151; 365), 23_1576 (1,195; 372), 23_535 (1,156; 328), 23_540 (1,206; 345), 23_529 (1,833; 111), 22_1870 (870; 102), 22_1869 (622; 108), 23_2060 (794; 131), 23_827 (549; 167), and 30 (601; 135), all above Tukey's upper fence for deletions (>543) and, for a subset, also for duplications (>288). Length outliers were rarer but notable (e.g., duplication median up to ~25.2 kb in one sample; deletion medians >4-7 kb in a few others), suggesting sample- or batch-specific noise/segmentation that I will dissect next (Figure 4.2).

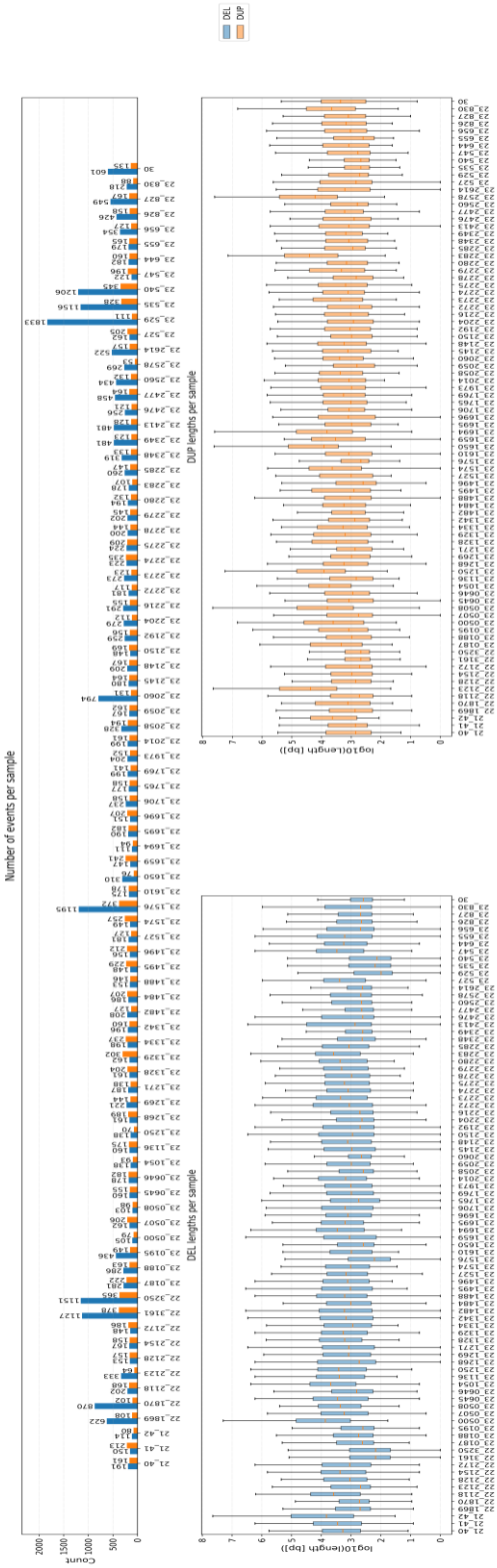

Figure 4.2: CNVs call performed with ExomeDepth on an in-house WES batch captured with Twist Exome 2.0 Plus Comprehensive Exome Spike-in (HG38).

Taken together, the ExomeDepth switch trades long, sparse, megabase events for many short, kilobase-scale calls with improved DEL/DUP balance, a profile that is methodologically consistent given the ExomeDepth beta-binomial modeling, but that demands robust QC and prioritization (e.g., cohort frequency filters/panel-of-normals, per-target QC, capture-uniformity checks) to be clinically actionable; the sheer number of CNVs precludes exhaustive manual curation, and the outlier samples with hundreds to thousands of events warrant targeted investigation of coverage, batch effects, and reference-panel composition.

4.1.3. GC-Bias as a Driver of Batch-Specific CNV Instability

The shift from EXCAVATOR2 to ExomeDepth reduced the prevalence of megabase-scale, and yielded a more balanced DEL/DUP profile, but it also exposed a different and likely more primary signal: coverage oscillations tightly coupled to genomic GC content. As illustrated by the capture-wise curves in Figure 4.3 (normalized coverage vs %GC over the design's regions), individual captures display distinct GC response functions, some peaking in GC-rich bins, others dipping, superimposed on the same target GC distribution.

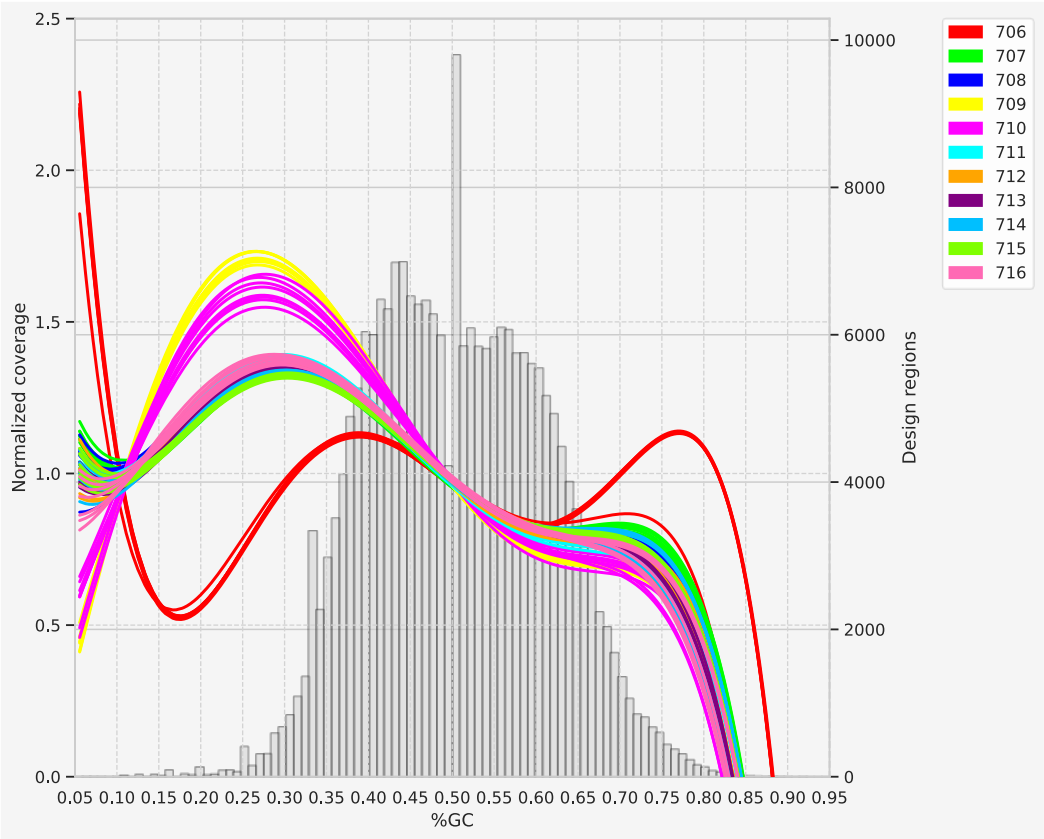

Figure 4.3: Normalized coverage across Twist target regions as a function of GC content (%GC). Colored curves represent individual capture experiments (IDs 706–716), while the grey histogram shows the distribution of design regions across GC bins (right axis).

This heterogeneous enrichment/penalization across the GC spectrum is sufficient to create systematic depth "waves" that may mimic CNV segmentation, thereby inflating call counts and pushing certain samples into outlier territory. In practice, when enrichment differs between captures or runs, the reference panel selected by a read-depth caller may introduce bias, and short kilobase-scale events accumulate precisely where the coverage-GC curve deviates most from the cohort mode. Our discussion with the Twist manufacturer confirmed that these patterns are a known multi-factor phenomenon, linked to library complexity, hybridization/wash conditions, and protocol variants, along with practical levers to mitigate them (e.g., increasing effective library complexity, testing alternative hybridization chemistries and wash implementations, and considering instrumented temperature control for washes). The key point is that there are wet-lab adjustments worth piloting as well as computational controls.

4.1.4. Quantitative assessment of GC bias across cohorts

To complement the graphical assessment of GC bias, I quantified the fraction of normalized coverage variability explained by GC content at the sample level. For each sample, normalized coverage was modeled as a third-degree polynomial function of target GC content, and the resulting R^2 was used as a summary measure of GC-associated variability. In the Burlo cohort, GC content explained a substantial fraction of normalized coverage variability, with a median R^2 of 0.237 (mean 0.245, IQR 0.220–0.258, range 0.074–0.388). In contrast, in the CNVPANEL01 (Twist) cohort, the corresponding values were markedly lower, with a median R^2 of 0.040 (mean 0.042, IQR 0.036–0.047, range 0.031–0.063) (Figure 4.4).

The difference in GC-bias score distributions between the two cohorts was highly significant (Mann–Whitney U test, $p << 0.001$).

These results indicate that GC-associated effects account for a substantially larger fraction of coverage variability in the heterogeneous clinical cohort compared to the more homogeneous reference dataset. This quantitative difference is consistent with the increased instability of CNV calls observed in the Burlo cohort and supports the hypothesis that technical factors, including GC bias and batch-related variability, contribute substantially to coverage distortions in clinical exome data.

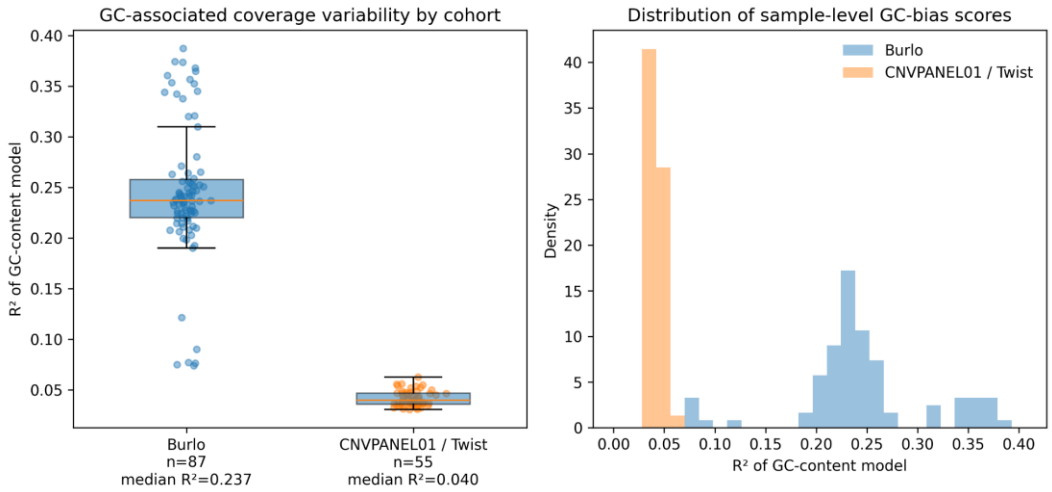

Figure 4.4: Sample-level GC-bias scores were computed as the coefficient of determination (R^2) of a third-degree polynomial regression modelling normalized coverage as a function of target GC content. The left panel shows the distribution of R^2 values across samples in the Burlo clinical cohort

and in the CNVPANEL01/Twist reference cohort. The right panel shows the corresponding score distributions.

4.1.5. Evaluation of ExomeDepth on the CNVPANEL01 Reference Cohort

To separate GC-driven artefacts from more general caller behavior, I next applied ExomeDepth to the CNVPANEL01 cohort provided by the Twist company. These 55 reference samples form a deliberately favorable use case, with highly correlated coverage profiles across targets and no obvious batch discontinuities, so one would expect a relatively stable CNV landscape if ExomeDepth were mainly reacting to gross mismatches in library quality or GC composition (Figure 4.5).

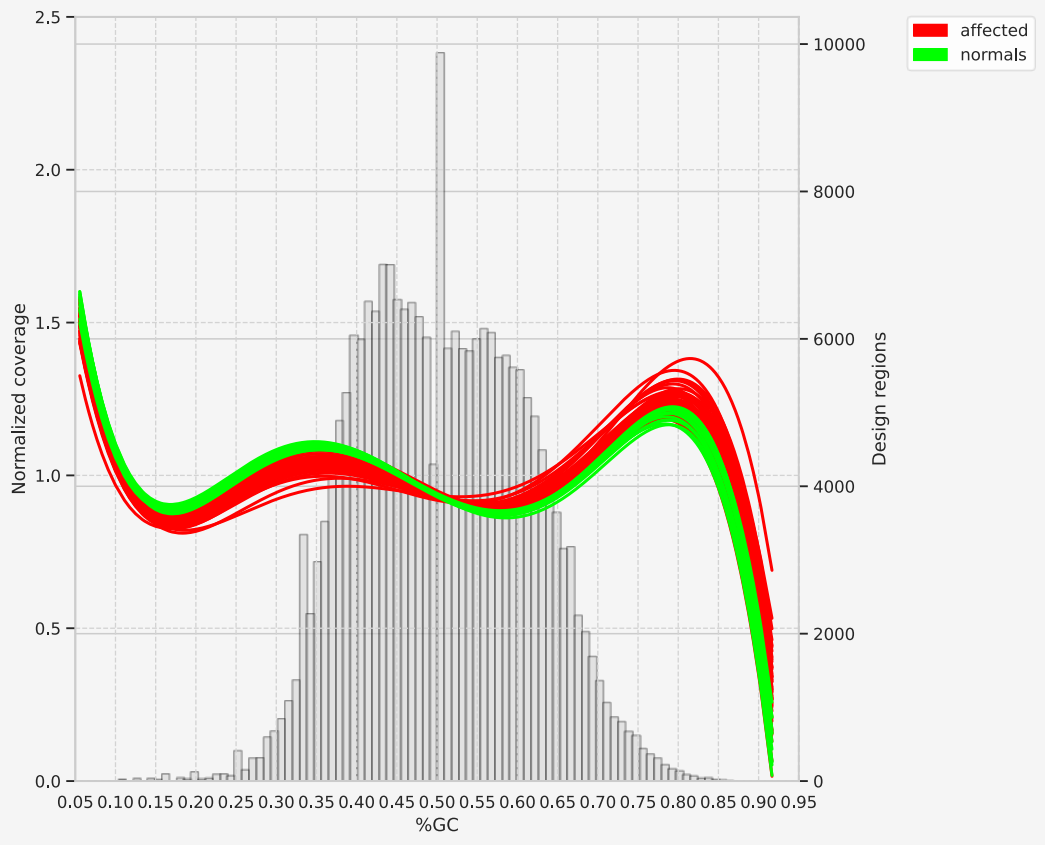

Figure 4.5: Normalized coverage across Twist target regions as a function of GC content (%GC). The colors represent the affected samples (red) and the healthy samples (green).

The samples alignment statistics are described in Table 4.2.

	#fragments	insert_size MEAN	%dupl	MEAN Coverage	%1X	%10X	%PASS	fold80	Uniformity of coverage (Pct > 0,2*mean)
Mean	66.288.245	256,33	15,71	128,65	99,69	99,58	96,37	1,27	97,77
Median	68.367.790	257,58	15,80	133,01	99,76	99,64	96,39	1,27	97,78
St.Dev.	15.672.523	5,49	1,72	28,70	0,10	0,10	0,05	0,02	0,06

Table 4.2: The average insert size, the percentage of duplicated reads, the percentage of genomic bases covered by at least 1X and 10X, the percentage of callable bases, the fold80 (computed on non-zero coverage bases) and the Uniformity of coverage (% of bases coverage at least 20% of the mean coverage).

Indeed, the call burden was more regular than in the earlier clinical cohort: ExomeDepth reported 10,516 deletions and 8,211 duplications overall (DEL:DUP ≈ 1.28:1), with per-sample medians of 183 deletions (IQR 168.5-210) and 154 duplications (IQR 136-167). Median event lengths remained in the kilobase range and were tightly clustered across samples, with a median of medians of ~1.56 kb for deletions (IQR 1.05-2.24 kb) and ~1.62 kb for duplications (IQR 1.20-2.38 kb), as illustrated in the lower panels of Figure 4.6. In terms of counts, only a single library was an outlier for deletions (GM16595, 456 DEL; Tukey upper fence 272) and one for duplications (GM21699, 235 DUP; upper fence 213.5), confirming that the bulk of the panel sits in a compact band around the cohort medians. Length outliers were more frequent but still interpretable: a subset of samples shows markedly elevated median segment sizes, with deletion medians up to ~7.5 kb and duplication medians up to ~6.8 kb (e.g., GM01416, GM16595, GM17867, GM20027), suggesting localized segmentation related to the underlying validated events.

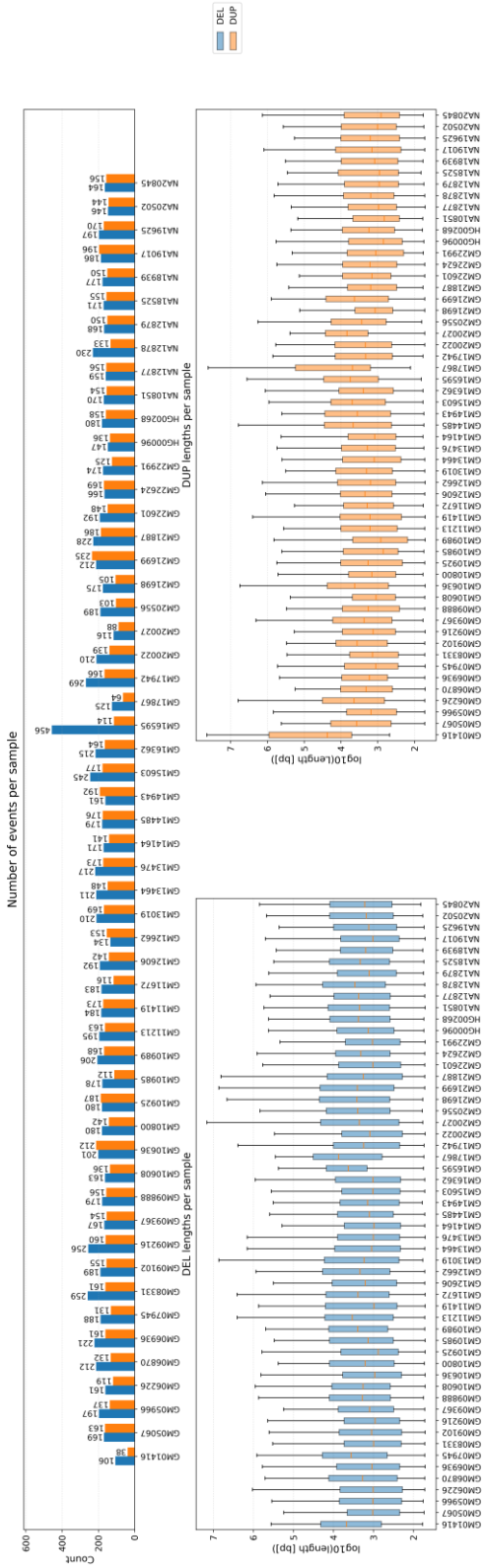

Figure 4.6: CNV calls performed with ExomeDepth on the CNVPANEL01 samples captured with Twist Exome 2.0 Plus Comprehensive Exome Spike-in (HG38).

Taken together, these Twist-derived data showed that, even with strongly correlated reference profiles, ExomeDepth continued to generate a wide dispersion of copy-number counts, including few but clear outliers. The instability observed in our clinical batches therefore cannot be attributed solely to idiosyncrasies of our wet-lab implementation or to extreme GC-mismatches in a handful of libraries; rather, it reflects more general interactions between capture design, library properties and the sensitivity of read-depth callers.

4.1.6. Correlation Helps, but It Does Not Solve the Problem

The Twist CNVPANEL01 analysis showed that high inter-sample correlation markedly stabilizes ExomeDepth's behavior compared to heterogeneous clinical batches, but it did not eliminate the underlying variability of read-depth-based CNV calling. Even in a manufacturer-controlled panel with tightly co-varying coverage profiles, ExomeDepth still produced a broad dispersion of call counts.

To further test whether part of this instability reflected poorly matched normals rather than intrinsic caller behavior, I then pruned the ExomeDepth baseline for the Burlo dataset to a more homogeneous pool of controls and let the software recompute, for each case, its internal reference set. This optimization reduced the burden to 17,668 deletions and 13,838 duplications (31,684 CNVs) across 87 Burlo samples, with medians of 172 deletions (IQR 147.5-212.5), 153.5 duplications (IQR 135.5-170.75) and about 332 CNVs per library (IQR 307.25-375). The number of clear count outliers fell to seven samples above the deletion fence (more than about 310 events) and two above the duplication fence (more than about 224 events), but strongly inflated libraries persisted, for example 23_529 with 1,295 deletions and 1,390 CNVs in total and 22_3161 with 620 duplications (Figure 4.7).

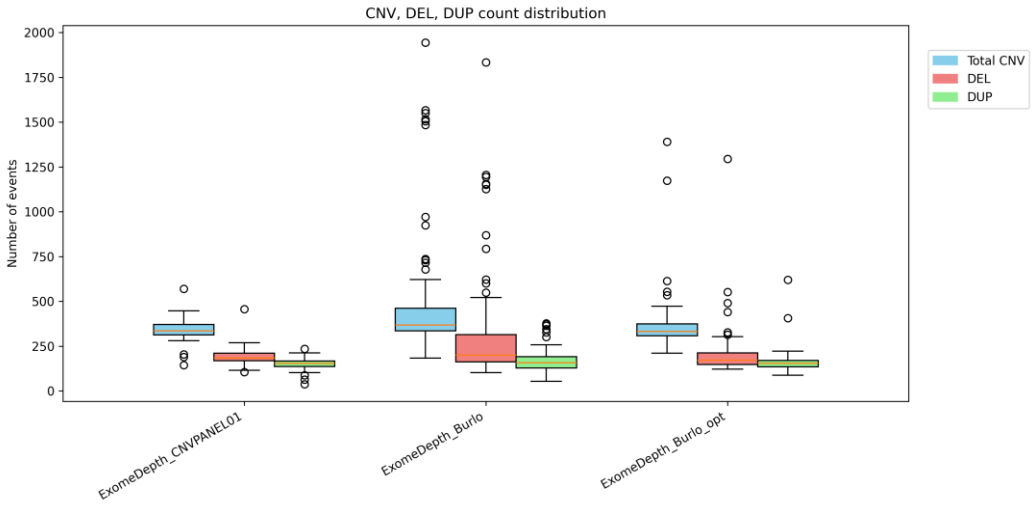

Figure 4.7: Per sample distribution of ExomeDepth CNV counts in the Twist CNVPANEL01 reference set and in the Burlo cohort before and after optimization of the normal baseline.

Across all three datasets, median event sizes remained in the kilobase range, with sample-specific median deletion lengths increasing from 942 bp (IQR 590-1,348) in Burlo to 1,063 bp (IQR 934-1,454) in the optimized Burlo subset and 1,599 bp (IQR 1,208-1,963) in CNVPANEL01, and analogous patterns for duplications (1,033, 1,110 and 1,564 bp respectively). At the same time, the composition of the baseline cannot be fully audited, because ExomeDepth automatically selects an "optimal" subset of normals for each case according to correlation but does not report which controls, or how many, are actually used. Taken together, the homogeneous CNVPANEL01 panel and the partial optimization of the Burlo cohort show that improving inter-sample correlation and refining the pool of normals clearly stabilizes ExomeDepth and attenuates the most extreme explosions of call volume, yet they still leave hundreds of short CNVs per sample and a non-trivial set of outliers. This burden is not compatible with exhaustive manual review in a diagnostic workflow and motivates the next step of this work, namely the construction of a reproducible multi-caller pipeline that (i) benchmarks ExomeDepth, gCNV and ClinCNV under harmonized preprocessing, (ii) makes reference-panel composition explicit, (iii) derives consensus call sets, and (iv) overlays an annotation and filtering layer that compresses the call burden to a clinically manageable shortlist.

4.2. Implementation of the CNV Detection and Benchmarking Pipeline

4.2.1. Overview of the Pipeline Architecture

A fully automated and modular workflow was implemented using the *Snakemake* workflow management system to orchestrate the entire process of copy number variant (CNV) detection, benchmarking, and comparative analysis across multiple algorithms and experimental conditions. The pipeline was designed to operate on whole-exome sequencing (WES) data, supporting the simultaneous processing of large sample cohorts and enabling the integration of several CNV calling tools, including ExomeDepth, clinCNV, and gCNV from the GATK suite. Each tool was incorporated as a separate submodule within the workflow, maintaining its specific requirements while ensuring standardization of final outputs.

The general philosophy behind this pipeline was to achieve a balance between flexibility and reproducibility, tailoring the available computational resources with the workload. By relying on a configuration-driven approach, all parameters, including reference genomes, target designs, scatter-gather strategies, and other options, are stored in YAML configuration files, which can be easily modified without altering the core Snakefile. The top-level Snakefile defines the high-level execution logic and coordinates the inclusion of modular rule files located under the *rules/* directory. Each module handles a specific stage of the analysis (e.g., preprocessing, CNV calling, and post-processing, output normalization, statistics evaluation, etc.). This modular design not only improves maintainability but also enables partial or full re-runs and parallelized execution across the computer environments.

From a computational standpoint, the workflow leverages Snakemake's rule-based dependency graph to optimize task scheduling and resource allocation, making it well suited for high-performance clusters. Moreover, by specifying input-output relationships rather than procedural steps, the pipeline guarantees full reproducibility and traceability of results. Not least, the *Snakemake* pipeline and the

underlying *conda environment* allow full and easy portability of the pipeline across different computing platforms.

4.2.2. Input Data Management and Configuration

At the core of the workflow lies a robust data management layer that dynamically constructs all the necessary inputs based on configuration files and sample metadata. The pipeline reads a PED file, a standard tab-delimited format used in human genetics, containing information about family structure, individual IDs, parental relationships, sex, and phenotype. This file provides the essential mapping between biological samples and their corresponding sequencing data, enabling the correct propagation of sample identifiers throughout all downstream analyses.

Configuration files in YAML format define all global parameters required by the workflow (e.g. *config/config_hg38.yaml*). These include paths to the reference genome, the results directory, input data location, the number of scatter-gather intervals for flexible splitting strategies, and the selection of CNV detection tools to be executed. When the scatter mode is set, the pipeline automatically determines the number of intervals in which the whole reference genome must be split to speed up the computation for *gCNV*, the slowest of the software; otherwise, the reference genome can be used without segmenting the analysis.

Such flexibility allows the pipeline to be readily adapted to different reference assemblies, capture designs, and sequencing strategies. For instance, the same architecture can accommodate both WES designs and targeted designs datasets, by merely adjusting the configuration YAML. The use of structured configuration files ensures that every run is fully documented and reproducible, while also simplifying the integration of metadata from different sequencing centers or platforms. This level of abstraction was a critical design choice to enable large-scale comparative studies without manual reconfiguration or code modifications.

4.2.3. Workflow Structure and Execution Logic

The main orchestration of the pipeline is handled through a top-level *Snakefile*. The file could allow independent or combined execution of subsets of tools, which proved useful for benchmarking and method comparison.

The *rule all* section of the *Snakefile* defines the expected final outputs for a successful workflow completion, using Snakemake's *expand()* function to dynamically generate file paths based on the list of samples, tools, and intervals. This includes intermediate files such as BAM symlinks, reference interval annotations, normalization matrices, and the final CNV callsets in TSV or VCF format. By defining the entire analysis through explicit input-output relationships, the workflow ensures that any missing file triggers the automatic re-execution of its corresponding rule, a feature essential for reproducibility in complex multi-step analyses (Figure S. 1).

Each CNV detection tool is implemented in a dedicated submodule located under the *rules/* directory.

4.2.3.1. Normal Sample Selection

Before entering tool-specific pipelines, the workflow creates a tool-specific baseline layer under *workflow/rules/baseline_generation/*. The *normals_optimization.smk* file builds the lists of "normal" samples used for reference modeling in each caller.

The PED file is first loaded with the standard six-column header, providing a consistent representation of sample identities, family relationships, sex and phenotype. For each sample and each tool, the workflow can either:

- reuse a static definition of normals (e.g. pre-defined lists), or
- compute an optimized set of reference samples, depending on flags such as *reference_optimization, drop_relatives_<tool>, split_sex_<tool>, and drop_affected_<tool>*.

The Python script *select_normals.py* implements this optimized normal selection strategy. It reads the PED file, constructs an in-memory representation of individuals and their relationships, and applies a series of filters based on options (e.g. *drop_relatives_ExomeDepth=True, split_sex_ClinCNV=True*). Qualitatively, the logic:

- optionally excludes affected individuals for callers configured to use healthy references;
- optionally removes relatives of the index sample to avoid biasing the reference with shared CNVs;
- optionally stratifies by sex, building sex-matched reference sets (required by ExomeDepth).

On top of these pedigree- and phenotype-based filters, a dedicated Python clustering module uses coverage-derived features to refine the choice of normals. For the candidate reference pool, it computes a Pearson correlation matrix of coverage profiles and applies correlation-based clustering (and/or ranking) to identify those samples whose coverage patterns are most similar to the index sample. The final normals list for each tool is obtained by selecting the highest-correlated, unrelated, sex-matched and unaffected individuals (depending on the user's choice), up to a configurable maximum number of reference samples. This approach ensures that the reference cohorts are not only biologically appropriate but also technically well matched, which is particularly important for methods such as ExomeDepth that are sensitive to coverage heterogeneity.

These normals lists are thus tailored for each tool and each sample.

4.2.3.2. ExomeDepth module

The ExomeDepth module embodies a compact yet complete implementation of the ExomeDepth workflow. Coverage extraction is performed through a dedicated Snakemake rule that processes each BAM file together with the BED target list, producing count tables suitable for input into the ExomeDepth R scripts. These scripts, included in the repository and automatically invoked by Snakemake, build the Bayesian model that represents each sample as a linear combination of reference

individuals. The module also handles the automatic selection of reference samples and converts the raw ExomeDepth output into harmonized TSV files. A consistent naming policy and shared directory structure facilitate direct comparison with the outputs of other callers.

4.2.3.3. ClinCNV module

The clinCNV module is organized around three sequential phases that reflect the native structure of clinCNV. A preprocessing step adapts the BED file to the exact format required by the tool. A second phase computes target-wise coverage matrices and prepares the normalization inputs needed for segmentation. The final phase executes the clinCNV calling routine itself. The workflow also manages the auxiliary files that control the selection of normal samples, relying on automatically generated lists stored in dedicated directories. These lists allow clinCNV to tailor its normalization to the structure of the cohort, and internal consistency checks ensure that missing or incompatible normals are detected early in the process.

4.2.3.4. gCNV module

Among the supported tools, gCNV is the most computationally demanding, and its implementation reflects this complexity. The workflow first determines the target coverage with the *preprocess* tool and, subsequently, the contig-specific ploidy using the *DetermineGermlineContigPloidy* tool. Depending on the configuration, the reference may then be partitioned into multiple intervals to enable scatter-gather execution. For each subregion, the pipeline runs the *GermlineCNVCaller*, generating intermediate models capturing both sample-level and interval-level variance. These models are post-processed with *PostprocessGermlineCNVCalls* to derive interval-level copy-number segments and sample-wise posterior probabilities. Finally, the gathered outputs are merged into unified VCF files, ensuring consistent annotation and ordering across all contigs. By leveraging Snakemake's parallelization capabilities, the module distributes the workload efficiently across available resources, reducing runtimes for large cohorts.

4.2.4. Harmonization of Outputs, Merging Callers, and Summary Statistics

Once each caller has produced its own native output format (CSV/TSV/VCF), the workflow collapses them onto a standard BED representation and merges calls across tools.

In *workflow/rules/common/csv2bed.smk*, a set of rules convert ExomeDepth, ClinCNV and gCNV outputs into BED files:

- *convert_csv_to_bed_exomedepth* reads *ExomeDepth/raw_exom_out/{sample}.csv* and writes *ExomeDepth/bed_raw/{sample}.bed*.
- Similar rules handle ClinCNV and gCNV outputs (gCNV VCFs are converted to BED format), normalizing them into a common layout: chr, start, end, CNV_type, sample, and caller-specific annotations in extra columns.

Further processing (sorting and length filtering) produces *.../bed/{sample}.bed* and *.../bed_min{min_len}/{sample}.bed*, where *min_len* is the configured length threshold. This harmonized calling space is crucial for direct comparison between algorithms.

Next, *merge_callers.smk* implements a consensus-based merging of callsets using a custom Python code (in *workflow/scripts/merge_cnv_calls.py*). The rule defines:

- a union of all CNV segments reported by any caller,
- regions supported by at least two callers (intersect_2),
- regions supported by all three callers (intersect_3).

Length filters are then applied, ensuring only CNVs above *min_len* are propagated to final merged datasets.

Eventually, two rule files, *csv_stats.smk* and *merged_csv_stats.smk*, generate summary metrics:

- For each caller (ExomeDepth, ClinCNV, gCNV), *stats_bed* and *stats_bed_minlen* compute per-sample metrics (counts, length distribution) and append them into per-tool aggregate files (*stats_min{min_len}/merged.tsv*).
- For merged callsets (union, intersect_2, intersect_3), analogous rules under *merged_csv_stats.smk* generate statistics describing the consensus call sets.

A small Python script *calculate_stats.py* is used to implement these metrics in a consistent way across all BED inputs.

4.2.5. Benchmarking and Validation Modules

An essential component of the pipeline is its benchmarking system, implemented in the *rules/benchmark/* directory. These modules perform systematic comparisons between detected CNV calls and curated gold-standard datasets, such as validated CNV lists in BED format. Using *bedtools*-based intersections, the workflow quantifies the number of true positives (TP), for each caller.

The benchmarking process is structured in several levels of granularity. First, intersections are computed separately for each sample and tool, generating per-sample performance summaries. In particular, for each sample is evaluated:

- The proportion of target design regions overlapping the validated events and the CNV called (for target-sensitivity evaluation);
- The number of bases overlapping the validated events and the CNV called (for base-sensitivity evaluation).

Subsequently, aggregated statistics are derived to provide cohort-level statistics. Additional rules allow for the intersection of merged CNV calls (across tools) against reference sets, assessing the benefit of consensus-based CNV discovery.

This evaluation framework is complemented by custom scripts such as *intersect_TP.py* and *calculate_stats.py*, which further refine the comparison metrics, taking into account CNV type (deletion or duplication), overlap thresholds, and length distributions. As a result, the pipeline not only performs automated

variant calling but also generates a detailed quantitative assessment of the reliability and concordance of each CNV caller under study.

4.2.6. Annotation

The final step on the variant side uses AnnotSV via *workflow/rules/annotation/annotsv.smk*. The rule *annotsv_tsv* takes as input the merged BED files (especially union and intersect_2/intersect_3 after length filtering) and passes them to AnnotSV with parameters taken directly from the YAML configuration:

- annotsv_path: executable location.
- annotsv_refgenome: genome build (e.g. HG38).
- overlap_annotsv, annotationMode: overlapping strategy and annotation mode.

AnnotSV produces per-sample TSV files enriched with gene annotations, overlap with clinical databases, and predicted functional impact. These are the most clinically interpretable outputs of the pipeline, and they form the natural bridge between purely technical benchmarking and downstream variant interpretation.

4.2.7. Scalability, Parallelization, and Reproducibility

One of the defining features of this implementation is its ability to scale efficiently across computing environments. By leveraging Snakemake's native support for scatter-gather paradigms, the pipeline divides computationally intensive tasks, such as gCNV's cohort analysis, into smaller scatter intervals corresponding to sub-chromosomal regions. Each sub-task runs independently and in parallel, significantly reducing runtime on high-performance clusters.

The pipeline's rule definitions are fully reproducible and version-controlled, ensuring that any analysis can be exactly replicated given the same configuration files and input datasets. Additionally, Snakemake's built-in logging and DAG visualization tools allow users to track the progress of the workflow, identify failed rules, and re-run only the necessary steps. The use of checksum-based job tracking

further guarantees that intermediate files are only recomputed when necessary, preventing unnecessary duplication of effort and computation time.

This architecture proved particularly advantageous in large-scale analyses, where hundreds of samples were processed simultaneously. By managing concurrency and dependencies automatically, the pipeline enables high-throughput CNV calling while preserving full traceability and transparency of results.

4.3.Application of the CNV Detection and Benchmarking Pipeline

The development of an optimized workflow for Copy Number Variant (CNV) detection from Whole-Exome Sequencing (WES) data required the systematic evaluation of the performance, reproducibility, and interpretability of existing tools under realistic experimental conditions.

Four independent but interconnected analytical phases were designed to address this objective (Figure 4.8):

1. Benchmarking of CNV calling algorithms on validated datasets to assess inter-tool variability;
2. Optimization of baseline composition, investigating how reference-sample selection strategies affect CNV call stability;
3. Refinement and integration of CNV results, aiming to consolidate multi-tool outputs into a single, high-confidence call set;
4. Annotation and prioritization of the integrated CNV results.

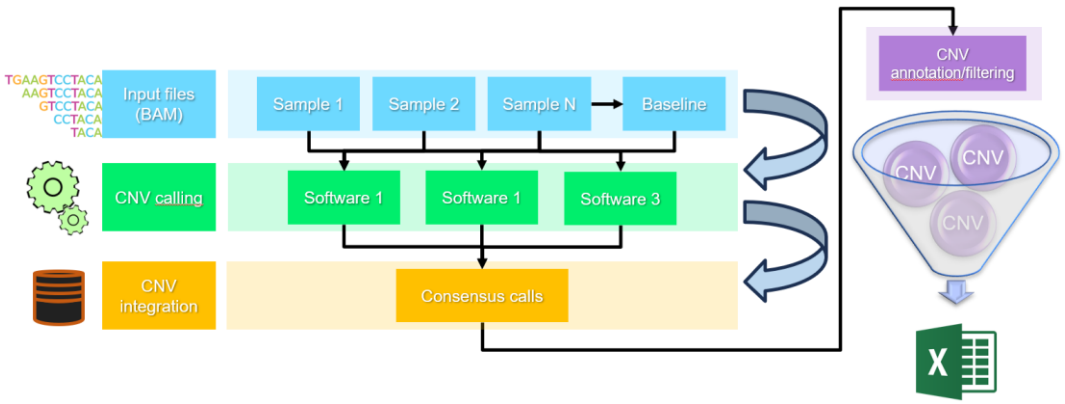

Figure 4.8: Snakemake-based pipeline overview used in this thesis.

These analyses were performed on two exome sequencing datasets differing in coverage uniformity and experimental quality, thereby representing both best- and real-case scenarios for CNV detection.
Benchmarking was primarily conducted using the CNVPANEL01 dataset, which exhibits highly correlated coverage profiles and contains validated CNV events used as a truth set.
Optimization and refinement steps were then also applied on a clinical dataset characterized by heterogeneous sequencing conditions, to test the pipeline's robustness in real diagnostic contexts.

ExomeDepth, ClinCNV, and gCNV (GATK) were evaluated as core components of the pipeline. Their methodological diversity (binomial modeling, HMM segmentation, and Bayesian inference, respectively) enabled a comparative assessment across different statistical paradigms for coverage-based CNV detection.

The results presented in this chapter focus on the analytical and computational outcomes of this multi-step approach, illustrating how each phase contributes to the enhancement of CNV detection from WES data.

4.3.1. CNVPANEL01 benchmark

4.3.1.1. CNV calling

Copy Number Variant (CNV) detection was performed on the CNVPANEL01 samples using the Twist Exome Design. Analyses were carried out using the three CNV detection tools integrated into the pipeline (ExomeDepth, clinCNV, and gCNV).

Restricting the analysis to CNV events longer than 50 bp, ExomeDepth identified on average 340 CNVs per sample, clinCNV reported approximately 79 CNVs, and gCNV ~202 CNVs (Figure 4.9).

The comparison highlighted marked differences in the call density and length distribution among tools. ExomeDepth and gCNV consistently generated a higher

number of CNV predictions per sample, indicating greater sensitivity but also a higher likelihood of false positives. Conversely, clinCNV produced a substantially lower number of calls, often spanning larger genomic intervals, suggesting a more conservative approach that prioritizes well-supported events (Figure 4.10).

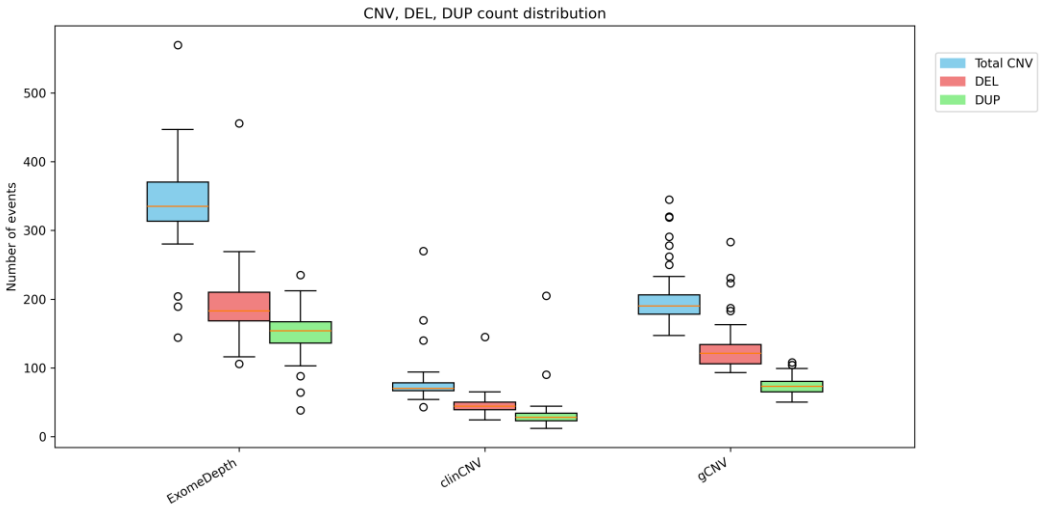

Figure 4.9: Per sample distribution of ExomeDepth, clinCNV and gCNV CNV counts in the Twist CNVPANEL01 reference set.

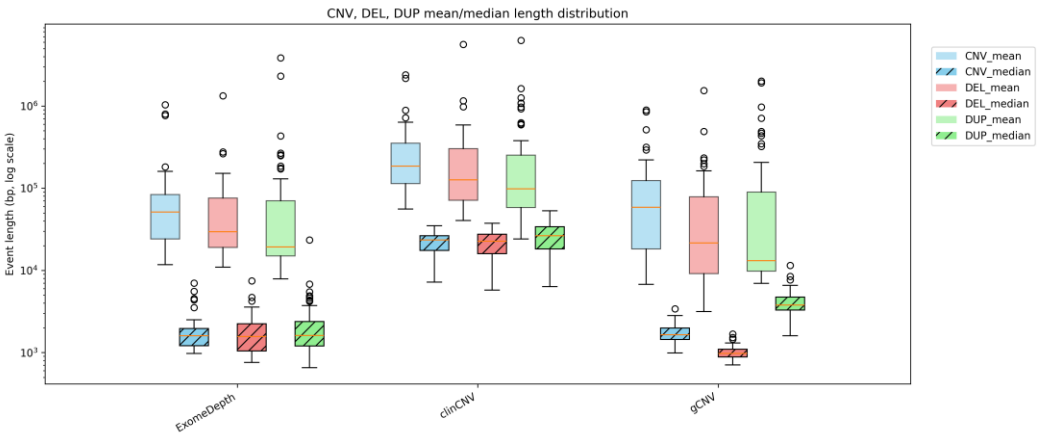

Figure 4.10: Per sample distribution of ExomeDepth, clinCNV and gCNV CNV lengths in the Twist CNVPANEL01 reference set.

4.3.1.2. Base-level performance

Before comparing callers at the target level, we first assessed how effectively they recovered the base-level extent of the 55 validated CNVs from the CNVPANEL01

affected samples. In an event-based view, a CNV was considered correctly detected when at least 75% of its validated bases were covered by one or more predicted segments (Methods 3.2.5). Under this definition, ExomeDepth correctly recovered 44/55 events (80.0%), while clinCNV and gCNV each reached 43/55 events (78.2%), indicating broadly similar event-level performance when using the Twist Exome design.

I then moved from events to all bases encompassed by the validated CNVs (1,295,732,537 bp), and computed base-level sensitivity as the fraction of these bases assigned to the expected copy-number state. Across all regions, ExomeDepth and gCNV recovered 86.5% and 87.8% of validated bases, respectively, whereas ClinCNV reached 80.6%. When stratifying by variant type, all three callers performed very well on deletions (559,820,086 bp validated), with base-level sensitivities between 93.8% and 95.9%, confirming that loss events are consistently well captured by read-depth segmentation. Duplications (735,912,451 bp validated) were more challenging: ExomeDepth and gCNV recovered 80.3% and 83.4% of validated bases, while clinCNV showed lower base-level sensitivity (68.9%), reflecting its tendency to produce fewer, more conservative segments (Table 4.3).

Tool	**TP covered > 75% of bases**	**Base-level sensitivity (%)**	**Base-level sensitivity (%) DEL**	**Base-level sensitivity (%) DUP**
ExomeDepth	44/55	86.5	93.9	80.3
ClinCNV	43/55	80.6	95.9	68.9
gCNV	43/55	87.8	93.8	83.4

Table 4.3: The number of TP identified, the base-level sensitivity and the base-level sensitivity split by DEL and DUP.

Together, these results show that base-level recovery of validated CNVs is generally high, especially for deletions, while duplications display greater variability across callers.

4.3.1.3. Target-level performance

Similar to before, sensitivity was first evaluated at the event level using the 55 validated CNVs from the CNVPANEL01 affected samples, defining a true positive (TP) as an event for which at least 75% of its design targets were assigned to the correct copy-number state (Methods 3.2.5). Under this criterion, ExomeDepth detected 51/55 events (93%), while ClinCNV and gCNV each detected 50/55 events (90%), confirming that all three callers achieve broadly comparable performance.

To complement the event-level assessment, sensitivity was also measured at the region level by aggregating all validated design regions across the 55 CNVs and quantifying the proportion correctly classified by each tool (93,301TP regions). This region-level analysis revealed highly consistent trends: ExomeDepth and gCNV reached 93.5% and 93.2% sensitivity, respectively, whereas clinCNV achieved 86.3%. Stratification by variant type showed uniformly high performance for deletions (98.9-99.7% across tools, 39,613 TP DEL regions), while duplications (53,688 TP DUP regions) remained more challenging, with sensitivities of 89.6% for ExomeDepth, 89.0% for gCNV, and 76.4% for clinCNV (Table 4.4).

Tool	**TP covered > 75% of bases**	**Target-level sensitivity (%)**	**Target-level sensitivity (%) DEL**	**Target-level sensitivity (%) DUP**
ExomeDepth	51/55	93.5	98.9	89.6
ClinCNV	50/55	86.3	99.7	76.4
gCNV	50/55	93.2	99.0	89.9

Table 4.4: The number of TP identified, the target-level sensitivity and the target-level sensitivity split by DEL and DUP.

4.3.2. Overlap of true positive bases across callers

To characterize how true positive bases are shared among callers, I summarized the base level overlaps within validated regions using global Venn diagrams for duplications and deletions. For deletions, the union of true positive bases amounts to about 540 Mb, and nearly all of them lie in regions supported by multiple tools. The triple intersection between ExomeDepth, ClinCNV and gCNV accounts for 509 Mb, that is approximately 94% of the union, while pair specific overlaps contribute only about 5% in total and caller specific regions are almost negligible, with at most 1-2% of bases unique to a single tool (Figure 4.11).

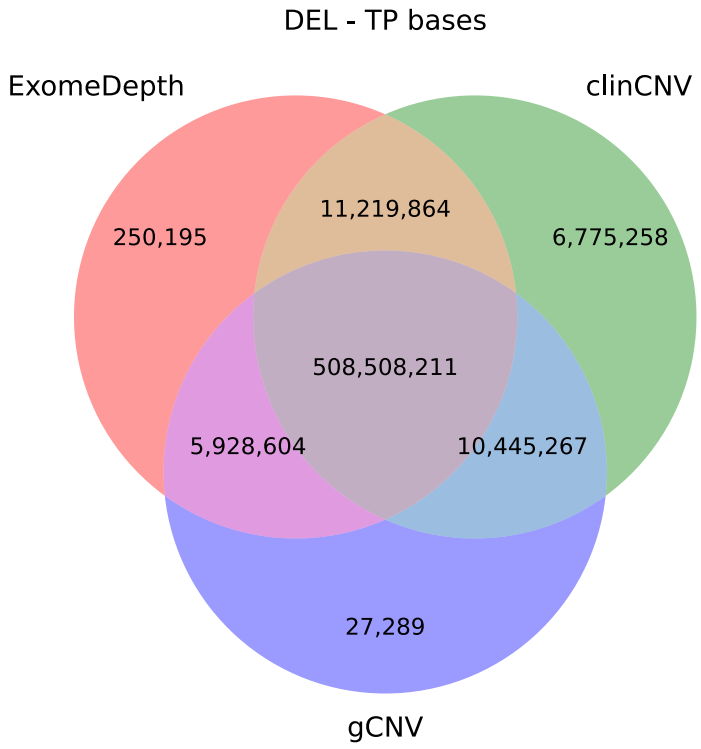

Figure 4.11: Venn diagram summarizing, for all validated deletions, the number of bases correctly identified by ExomeDepth, ClinCNV and gCNV, highlighting the large triple intersection and the smaller caller specific contributions.

For duplications, the agreement remains high but is more heterogeneous. The union comprises about 670 Mb, of which 423 Mb, roughly 63%, fall in the triple intersection. A substantial fraction of true positive duplication bases, about 151 Mb or 23% of the union, is shared only between ExomeDepth and gCNV, whereas

bases unique to individual callers represent small but non zero contributions, around 6% for ClinCNV and about 1% for each of the other two (Figure 4.12).

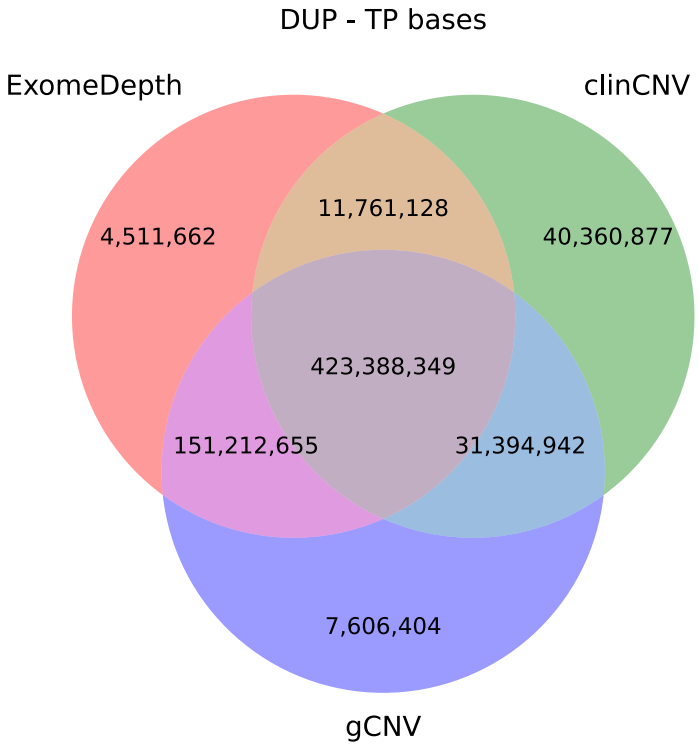

Figure 4.12: Venn diagram summarizing, for all validated duplications, the number of bases correctly identified by ExomeDepth, ClinCNV and gCNV, highlighting the large triple intersection and the smaller caller specific contributions.

These patterns confirm that, in the space of validated CNVs, the three tools largely converge on the same events, especially for deletions, but each caller retains a specific sensitivity profile that contributes additional true positive bases.

4.3.3. Impact of Baseline Composition on CNV Detection

4.3.3.1. Rationale

To assess how reference composition influences CNV detection, I compared three callers (ExomeDepth, ClinCNV, and gCNV) under different baseline configurations. The first baseline consisted exclusively of samples from CNVPANEL01, a curated and homogeneous cohort. The second baseline combined CNVPANEL01 with the Burlo samples, producing a heterogeneous cohort (CNVPANEL01+Burlo) that better approximates real clinical laboratory

conditions, where normals may originate from diverse sequencing runs, batches, and quality profiles. Finally, I evaluated the effect of selecting a subset of reference normals based on Pearson correlation, using 5 to 30 coverage-matched samples drawn from the heterogeneous cohort.

As before, for all callers, performance was quantified along three axes:

- CNV counts (total, deletions, duplications);
- base-level coverage;
- target-level sensitivity.

The combination of these metrics allows distinguishing between precision (call volume and fragmentation) and sensitivity (ability to detect validated CNVs), particularly when transitioning from an idealized to a realistic baseline.

4.3.3.2. Coverage similarity landscape

When considering the full heterogeneous baseline (all 142 normals), the two cohorts showed measurable differences in coverage similarity. For Burlo samples, the mean Pearson correlation to all other normals ranged from 0.947 to 0.971, with a median of 0.967. The worst pairing for each Burlo sample ("min over all partners") was typically around 0.91 (median 0.910, range 0.865-0.942). CNVPANEL01 samples were systematically less similar to the rest of the cohort: their mean correlation to all other samples had a median of 0.958 and ranged from 0.930 to 0.965, and their minimum correlation per sample ranged from 0.865 to 0.929 with a median of 0.907. In practical terms, more than half of the CNVPANEL01 normals (34/55) had an overall mean correlation below 0.96, compared with 16/87 Burlo samples. Thus, once the two cohorts are pooled, the baseline is clearly not perfectly homogeneous; if one takes 0.98 as a conservative threshold for "well-matched" coverage profiles, the average similarity experienced by a typical sample is already noticeably below that target.

The top-K analysis (Figure 4.13) shows how much this improves when reference normals are explicitly selected by correlation and how quickly similarity deteriorates as K increases. When each sample is paired only with its five most correlated neighbors, the median mean correlation is ≈0.990 (range 0.970-0.993

across all samples), and 138/142 samples have all of their top-5 partners at $r \geq 0.98$. As K increases from 5 to 30, the median mean correlation gradually decreases from 0.990 to 0.987, and the lower whisker drifts away from 0.98: for K=30 the mean correlation spans ≈0.962-0.990, and seven samples have an average correlation to their 30 "best" references below 0.98. The effect is slightly more pronounced for CNVPANEL01, where the median mean correlation to the top-30 neighbors is 0.986 (range 0.962-0.989), compared with 0.988 (0.973-0.990) for Burlo samples. These results indicate that, although the heterogeneous CNVPANEL01+Burlo baseline contains many highly correlated normals, the choice of K matters: small K (5-15) allows almost all samples to be matched to references with $r \geq 0.98$, whereas pushing K towards 20-30 inevitably pulls in less similar profiles, especially for the most atypical samples.

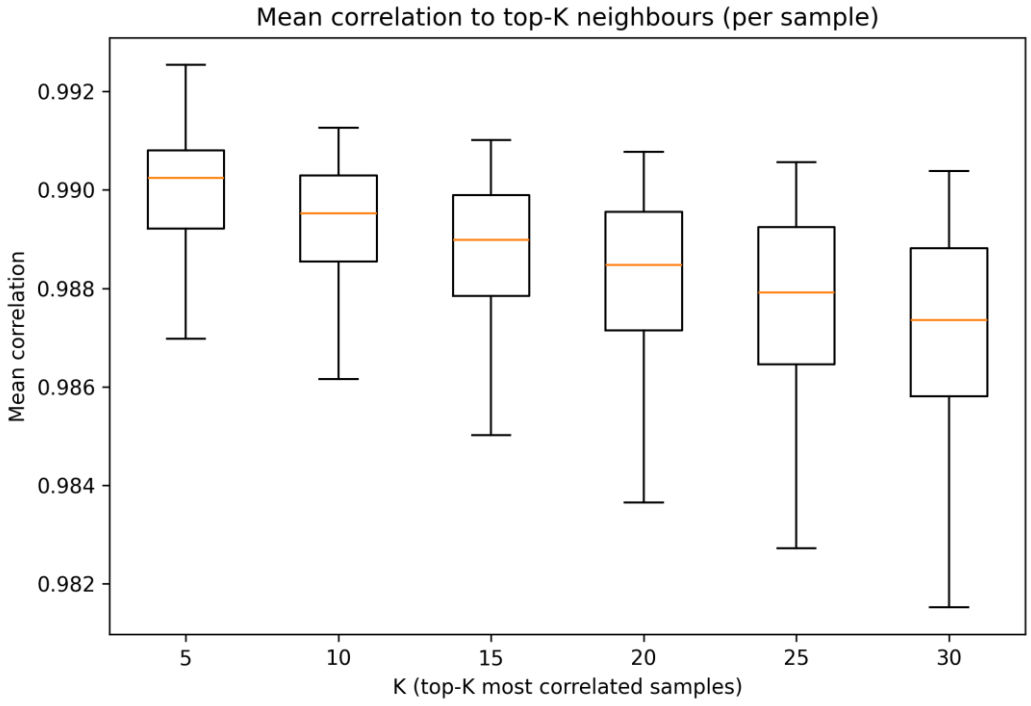

Figure 4.13: Mean coverage Pearson correlation coefficient across samples.

4.3.3.3. ExomeDepth

ExomeDepth showed the strongest dependence on baseline composition. When switching from CNVPANEL01 to the heterogeneous CNVPANEL01+Burlo baseline, the number of detected CNVs increased dramatically. Total calls nearly

doubled (from ~340 to ~688 average events), with the most pronounced inflation occurring in duplications (149 → 445), while deletions also showed a substantial rise (191 → 243). This inflation reflects increased susceptibility to coverage variability rather than biological signal. Also, base-level and target-level sensitivity dropped (93.6% vs 88.6% and 86.2% vs 63.2%).

Correlation-based selection of normals reversed this effect entirely. Selecting 5-15 highly correlated normals restored CNV counts to values comparable to the CNVPANEL01 baseline, both for deletions and duplications, while maintaining or slightly improving sensitivity (peak 95.1%) and base-level coverage (peak 87.1%). Increasing the number of selected normals (20-30 samples) preserved this stabilized profile (Figure 4.14).

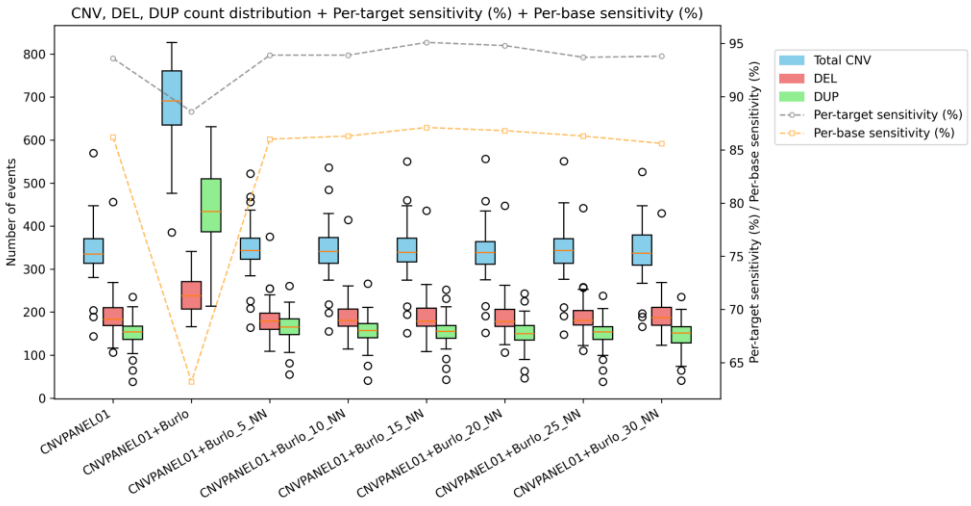

Figure 4.14: Per sample distribution of ExomeDepth CNV counts in the Twist CNVPANEL01 reference set with different baselines. Also, the per-base and per-target sensitivity are reported.

Overall, ExomeDepth benefits substantially from tailored baselines: in realistic settings (CNVPANEL01+Burlo), baseline optimization is essential to prevent marked inflation in CNV calls while preserving correct detection of validated events.

4.3.3.4. clinCNV

ClinCNV displayed a markedly different pattern. When moving from CNVPANEL01 to the heterogeneous CNVPANEL01+Burlo baseline, the caller did not inflate the total number of CNVs. Instead, CNV counts decreased (79 → 62),

and the distributions became slightly more compact. Similar trends were observed for deletions and duplications. This suggests that clinCNV's multi-sample normalization integrates additional variability rather than amplifying it, effectively dampening noise in the reference panel.

However, both base-level and target-level sensitivity slightly decreased when using the heterogeneous cohort (base-level: 80.6% → 77.5%; target-level: 86.3% → 85.4%). Correlation-based selection of normals did not fully restore the metrics to the CNVPANEL01 baseline: it made the situation worse. Nevertheless, it reduced the impact of some outliers.

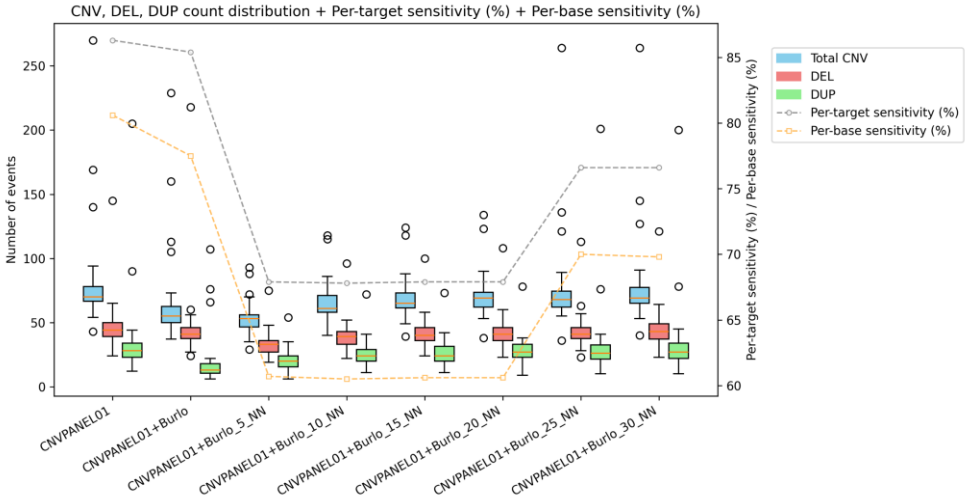

Figure 4.15: Per sample distribution of clinCNV CNV counts in the Twist CNVPANEL01 reference set with different baselines. Also, the per-base and per-target sensitivity are reported.

In summary, clinCNV is robust to heterogeneous baselines and produces stable call volumes, but its accuracy (both base- and region-level) is somewhat affected by baseline composition. Correlation-selected baselines offer moderate benefits primarily by mitigating outlier behavior rather than globally improving performance.

4.3.3.5. gCNV

gCNV exhibited yet another distinct behavior. Unlike ExomeDepth, the transition to the CNVPANEL01+Burlo baseline did not produce call inflation. Instead, call counts decreased substantially (202 →162), and distributions became tighter, indicating improved stability when using a more diverse reference cohort. This trend was consistent across deletions and duplications.

Moreover, gCNV maintained nearly constant base-level sensitivity (87.9% → 87.7%) and target-level sensitivity (~93%). These results indicate that gCNV benefits from larger and more variable reference cohorts, consistent with its hierarchical Bayesian framework, which stabilizes latent coverage components when fed with broader population-level variability.

Correlation-based selection had negative effects. Selecting small numbers of normals (<15 samples) increased CNV counts and introduced variability, likely due to reduced information available to the model (Figure 4.16).

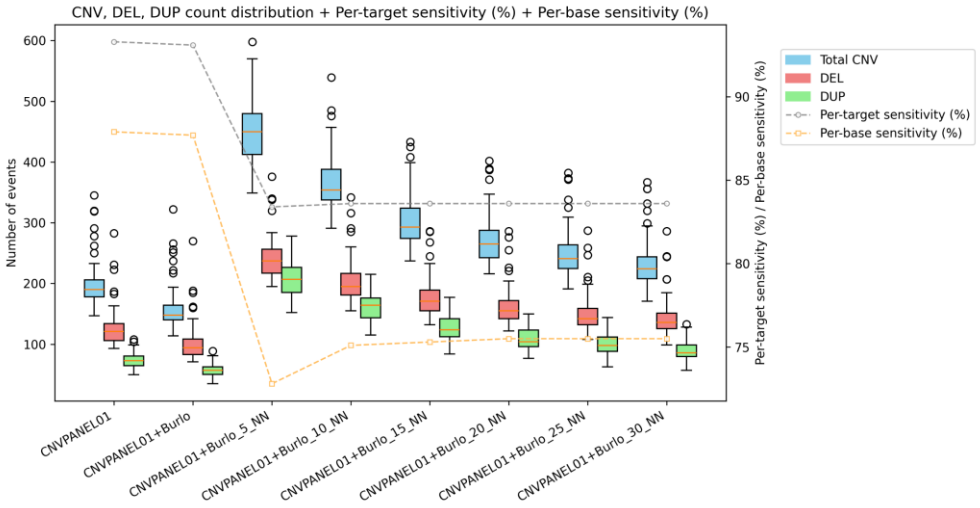

Figure 4.16: Per sample distribution of gCNV CNV counts in the Twist CNVPANEL01 reference set with different baselines. Also, the per-base and per-target sensitivity are reported.

This indicates that gCNV is sensitive primarily to baseline size, not baseline similarity: the caller performs best with broad cohorts and shows limited to none benefit from correlation-based filtering.

4.3.3.6. Integrated Interpretation

Taken together, these results show that baseline composition affects callers in fundamentally different ways.

- ExomeDepth mirrors real clinical variability: in heterogeneous conditions (CNVPANEL01+Burlo), it loses sensitivity and produces inflated and fragmented call sets. Correlation-based optimization is therefore necessary to restore realistic performance.

- ClinCNV is intrinsically robust to baseline heterogeneity. The heterogeneous cohort does not inflate calls but reduces global events, slightly reduces accuracy. Baseline optimization benefits mainly outliers, indicating that this tool requires less baseline tuning in practice.
- gCNV benefits from large, heterogeneous cohorts. It maintains stable accuracy and displays improved call stability when more variability is present in the baseline. Correlation-based selection is generally unnecessary.

These differences highlight the importance of caller-specific baseline strategies rather than applying uniform approaches across tools. The CNV counts and the base-level sensitivity according to the baseline selection are reported in Table S. 1 and Table S. 2.

4.3.4. Integration of CNV Calls

4.3.4.1. Consensus-Based Strategy

Given the complementary strengths and weaknesses of different CNV callers, a final refinement step was implemented to consolidate results across tools. Three integration levels were tested:

1. Union of all calls, maximizing sensitivity at the cost of including low-confidence events.
2. Intersection of calls detected by at least two tools, providing a balance between sensitivity and precision.
3. Intersection of calls shared by all three tools, maximizing specificity but reducing sensitivity.

This integrative approach aimed to exploit the statistical independence among algorithms, where recurrently detected events are more likely to represent true biological variants.

4.3.4.2. Evaluation of Consensus Call Sets

To evaluate whether integrating CNV calls across tools could provide a more robust and clinically meaningful result set, I quantified base-level and region-level performance for three consensus strategies: the union of all events, the intersection of events detected by at least two callers, and the strict intersection of all three.

For this analysis, I selected for each tool the configuration that yielded the highest accuracy in the previous sections: ExomeDepth with 15 correlation-selected normals, and ClinCNV and gCNV using the heterogeneous CNVPANEL01+Burlo baseline, which more accurately reflects real sequencing variability. These configurations provide the most reliable individual call sets and constitute an appropriate starting point for consensus integration.

As shown in Table 4.5, the individual tools exhibit notable differences: ExomeDepth achieves the highest region-level sensitivity (95.1%) with strong base-level performance (87.1%), gCNV provides a similarly balanced profile (93.1% region-level, 87.7% base-level), while clinCNV shows reduced accuracy (85.4% and 77.5%, respectively) but maintains exceptional precision. When combining results across callers, these strengths become complementary.

Tool	**TP covered > 75% of bases**	**TP covered > 75% of target**	**Base-level sensitivity (%)**	**Target-level sensitivity (%)**	**CNV counts (mean/median)**
ExomeDepth_15_nn	45/55	51/55	87.1	95.1	342/339
ClinCNV_all	44/55	51/55	77.5	85.4	62/55
gCNV_all	43/55	50/55	87.7	93.1	162/148

Table 4.5: The number of TP identified according to base- and target-level 75% cutoff, the base-level sensitivity, the target-level sensitivity and the CNV counts (mean/median).

The union of all calls unsurprisingly achieves the highest region-level sensitivity (96.6%), but at the cost of producing a relatively large number of events (~366 per sample), limiting its utility in clinical or translational analyses. Conversely, the strict intersection of all three callers yields excellent specificity but suffers a marked loss in sensitivity (82.7%) and produces a small call set (60 events on average), providing an incomplete representation of the validated CNV landscape (Table 4.6).

Tool	TP covered > 75% of bases	TP covered > 75% of target	Base-level sensitivity (%)	Target-level sensitivity (%)	CNV counts (mean/median)
Union	47/55	54/55	92.0	96.6	366/358
At least 2	46/55	51/55	89.5	94.3	136/127
All	41/55	47/55	71.0	82.7	60/56

Table 4.6: The number of TP identified according to base- and target-level 75% cutoff, the base-level sensitivity, the target-level sensitivity and the CNV counts (mean/median).

The intermediate strategy, retaining only events detected by at least two tools, achieves the most balanced and practically valuable outcome. This approach maintains exceptionally high region-level sensitivity (94.3%) and the highest base-level performance among all strategies (89.5%), surpassing what any single caller achieves individually. At the same time, the resulting call volume (136 events on average, 127 median) remains relatively compact with single callers (except clinCNV) and with the union, but it is still too high to be exhaustively reviewed in routine clinical practice. These characteristics indicate that agreement between at least two callers is a good indicator of true biological signal and provides a useful compromise between sensitivity and precision, while still requiring additional prioritization and annotation layers to reach a clinically manageable shortlist.

Taken together, these results demonstrate that a consensus-based strategy requiring concordance between at least two tools offers the highest-quality call set, combining the complementary strengths of the three callers while minimizing their

respective weaknesses. This integrated panel maximizes the reliability of the detected events and provides a solid foundation for downstream analyses.

4.3.5. CNV Annotation

Copy-number calls from the three callers (ExomeDepth using the 15 most correlated reference samples by Pearson correlation, clinCNV on the full cohort, and gCNV on the full cohort) were harmonized and annotated with AnnotSV v3.5.3. For downstream analyses I focused on the at-least-two-caller consensus, defined as CNVs supported by ≥2 callers for a given sample. Within the at-least-two-caller consensus, I summarized CNVs per sample and per variant type (deletion versus duplication), and stratified them by AnnotSV ACMG-based classification (ACMG 4-5 = likely pathogenic/pathogenic, ACMG 3 = variants of uncertain significance, ACMG 1-2 = benign/likely benign). I then re-computed these counts after applying internal cohort frequency thresholds of ≤5% and ≤1% on the allele frequency reported in the population.

4.3.5.1. Cohort-level CNV annotation across caller consensus and frequency strata

In this consensus set, the overall CNV calls per sample is similar between affected and unaffected individuals, and the informative contrast arises from ACMG class composition and frequency strata rather than from raw counts. Among affected samples, the at-least-two-caller consensus contains 3,908 deletions and 1,865 duplications (medians of ~87 deletions and ~43 duplications per sample). In unaffected samples, the corresponding numbers are 1,138 deletions and 573 duplications (medians ~86 and ~42 per sample).

For deletions in affected individuals, ACMG classification already produces a clinically meaningful spread: out of 3,908 consensus deletions, 272 (7.0%) are annotated as LP/P, 1,018 (26.0%) as VUS, and 2,618 (67.0%) as benign/likely benign. In unaffected samples, LP/P deletions are fewer in absolute and relative terms (46 of 1,138, 4.0%), with a larger benign fraction (~71%). For duplications, both affected and unaffected samples are dominated by VUS, but LP/P calls are

clearly enriched in cases. Affected individuals carry 1,865 consensus duplications, of which 56 (3.0%) are LP/P, 1,633 (87.6%) VUS, and 176 (9.4%) benign/likely benign. In unaffected samples, there are 573 duplications with only 3 LP/P (0.5%) and a similar VUS fraction (~88%). This pattern shows that the at-least-two-caller consensus, combined with ACMG classification, already concentrates LP/P calls in affected individuals while keeping their number in controls very low (Figure 4.17, Figure 4.18).

Allele-frequency filtering further sharpens this contrast and helps to focus review on rare, plausibly actionable events. In affected samples, restricting to ≤5% internal frequency retains 2,071 of 3,908 deletions and 1,036 of 1,865 duplications. For deletions, the LP/P fraction increases from 7.0% in the unfiltered set to 10.7% in the ≤5% bin (221 LP/P out of 2,071 events), and for duplications from 3.0% to 5.2% (54 LP/P out of 1,036 events). Tightening to ≤1% frequency has a stronger effect: for affected deletions the total count drops to 1,522 (a 61% reduction relative to the full consensus), yet 197 LP/P deletions (72% of the original LP/P) and 995 VUS deletions (98% of VUS) are retained, while the benign component is reduced from 2,618 to roughly 330 events. For affected duplications, the ≤1% filter keeps 762 events (a 59% reduction), retaining 54 of 56 LP/P duplications (96%) but eliminating ~60% of VUS and the majority of benign/likely benign calls. In unaffected samples, rare deletions at ≤1% are fewer (399 events, LP/P fraction ~4.8%) and rare duplications even more so (216 events, LP/P fraction <1%), underscoring that most rare high-confidence LP/P CNVs accumulate in the affected cohort (Figure 4.17, Figure 4.18).

Overall, these results show that the at-least-two-caller consensus, complemented by ACMG classification and modest frequency filtering, achieves the desired compression of the candidate space. In affected individuals, the combination of consensus and frequency filters removes a large portion of clearly benign CNVs while preserving most LP/P events and a substantial subset of VUS that warrant clinical review. In unaffected samples, the same procedure yields very few LP/P calls, particularly among duplications, which supports the specificity of the

approach and suggests that the observed LP/P enrichment in affected samples is unlikely to be an artefact of the annotation pipeline.

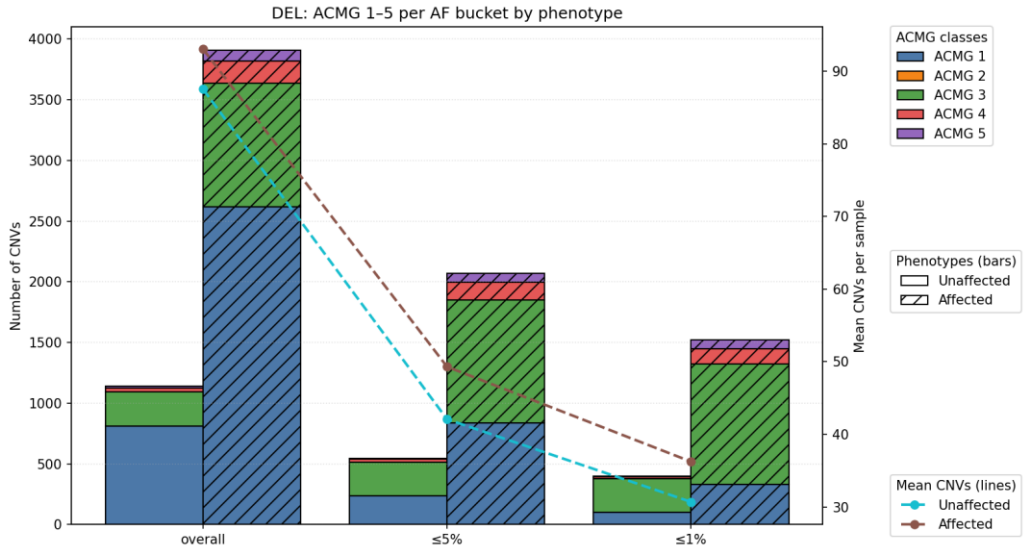

Figure 4.17: Distribution of deletion CNVs (DELs) of the CNVPANEL01 samples across ACMG classes 1-5, stratified by allele-frequency (AF) bucket and phenotype. Stacked bars show the total number of CNVs in unaffected and affected (hatched) individuals for all variants (overall), variants with AF ≤ 5%, and variants with AF ≤ 1%. Bar colors indicate ACMG pathogenicity classes as reported in the legend. Dashed lines (right y-axis) represent the mean number of CNVs per sample for unaffected (cyan) and affected (brown) individuals.

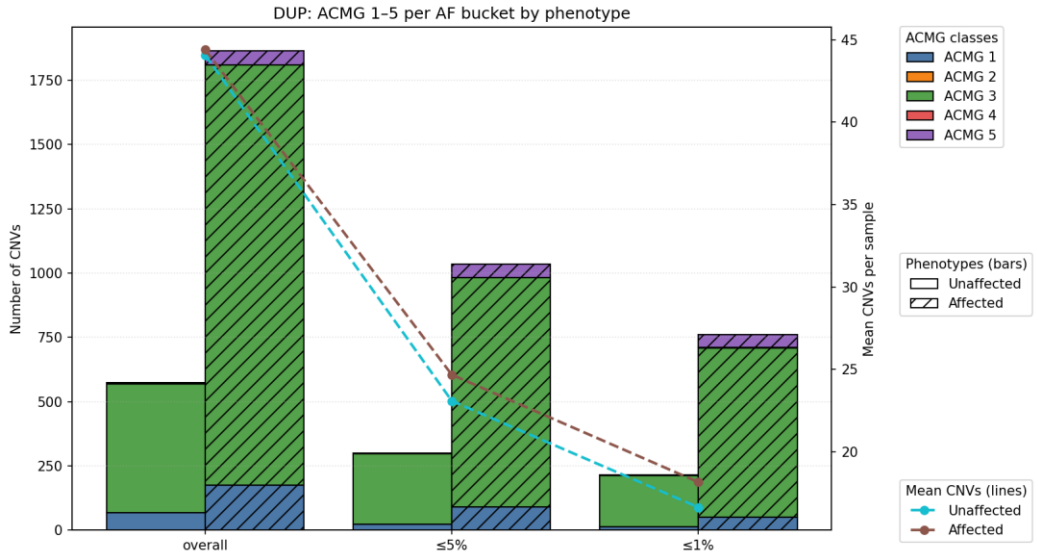

Figure 4.18: Distribution of duplications CNVs (DUPs) of the CNVPANEL01 samples across ACMG classes 1-5, stratified by allele-frequency (AF) bucket and phenotype. Stacked bars show the

total number of CNVs in unaffected and affected (hatched) individuals for all variants (overall), variants with AF ≤ 5%, and variants with AF ≤ 1%. Bar colors indicate ACMG pathogenicity classes as reported in the legend. Dashed lines (right y-axis) represent the mean number of CNVs per sample for unaffected (cyan) and affected (brown) individuals.

4.3.5.2. CNV annotation within validated regions (goldset overlap)

To evaluate how well this annotation and filtering strategy behaves on experimentally confirmed CNVs, I repeated the analysis after restricting calls to validated CNVPANEL01 regions (TP-only set). For each affected sample and caller, CNVs were intersected with the panel's validation targets; only overlapping events were retained, then consolidated with the at-least-two-caller consensus and annotated with AnnotSV. As above, I summarized deletions and duplications by ACMG class and then re-computed the counts after applying the ≤5% and ≤1% frequency filters.

Within these true-positive regions, the at-least-two-caller consensus is strongly enriched for CNVs that AnnotSV labels as LP/P or VUS. For validated deletions, the consensus includes 58 events in affected samples: 40 (69.0%) are LP/P, 2 (3.4%) are VUS, and 16 (27.6%) are benign/likely benign (ACMG 1-2). For validated duplications, there are 142 events, of which 40 (28.2%) are LP/P, 91 (64.1%) VUS, and 11 (7.7%) benign/likely benign. Thus, the vast majority of true-positive CNVs in these regions fall into the ACMG 4-5-3 classes, while a smaller tail (~25-30% for deletions and ~8% for duplications) is annotated as benign/likely benign. From a technical perspective, these ACMG1-2 events are still genuine CNVs (they are in the validation set), but current annotation resources consider them unlikely to be pathogenic (Figure 4.19, Figure 4.20).

Applying frequency filters inside the TP-only subset shows how much clinical triage costs in terms of raw sensitivity. For validated deletions, restricting to ≤5% frequency retains 50 of 58 events, and all 40 LP/P deletions and both VUS deletions survive this filter; only 8 benign/likely benign deletions are removed. Tightening to ≤1% further reduces the set to 48 deletions, but still keeps all 40 LP/P and both VUS deletions, trimming the benign tail down to 6 events. In other words, in the

deletion TP set, frequency filtering primarily removes true-positive CNVs that AnnotSV already considers benign, while fully preserving LP/P and VUS deletions (Figure 4.19).

For validated duplications, the behavior is similar but slightly less clean. The ≤5% filter retains 137 of 142 duplications, including all 40 LP/P and 88 of 91 VUS events, while the benign tail shrinks from 11 to 9 duplications. At ≤1%, 134 duplications remain: again all 40 LP/P are preserved, alongside 86 VUS and 8 benign/likely benign events. Thus, in true-positive regions, the combination of ACMG classification and frequency thresholds successfully concentrates CNVs into LP/P and VUS without substantially eroding the LP/P signal. A small fraction of validated CNVs is nonetheless classified as ACMG1-2 and progressively removed by frequency filters; these likely represent technically real but clinically neutral or poorly supported events, and their loss is quantitatively limited compared with the gain in interpretability (Figure 4.20).

Taken together, the TP-only analysis supports the use of the at-least-two-caller consensus as the main clinical set. On validated deletions and duplications, AnnotSV predominantly assigns ACMG 4-5 or 3, and the ≤5%/≤1% frequency filters act mostly on the benign tail while leaving all LP/P events intact. The fact that a small number of true-positive CNVs fall into ACMG 1-2 and are removed by filtering highlights the intrinsic limitations of current CNV annotation and frequency resources, but it does not overturn the overall strategy: for the vast majority of CNVs in validated regions, the consensus+ACMG+frequency pipeline behaves exactly as expected, concentrating clinically relevant classes (ACMG 4-5 and, to a lesser extent, 3) and discarding likely neutral variation.

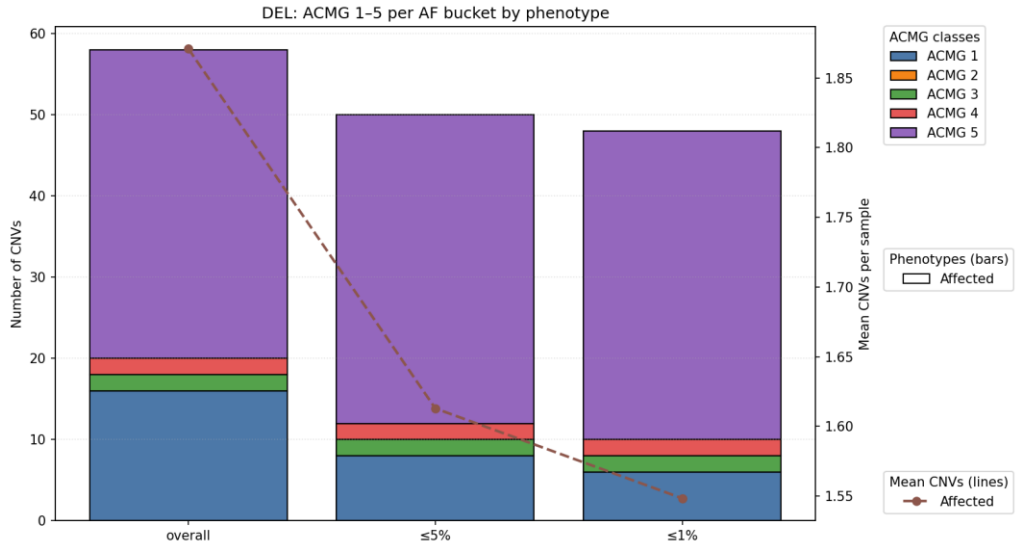

Figure 4.19: Distribution of deletion CNVs (DELs) within the TP regions only of the CNVPANEL01 samples across ACMG classes 1-5, stratified by allele-frequency (AF) bucket and phenotype. Stacked bars show the total number of CNVs in affected individuals for all variants (overall), variants with AF ≤ 5%, and variants with AF ≤ 1%. Bar colors indicate ACMG pathogenicity classes as reported in the legend. The dashed line (right y-axis) represents the mean number of CNVs per sample for affected (brown) individuals.

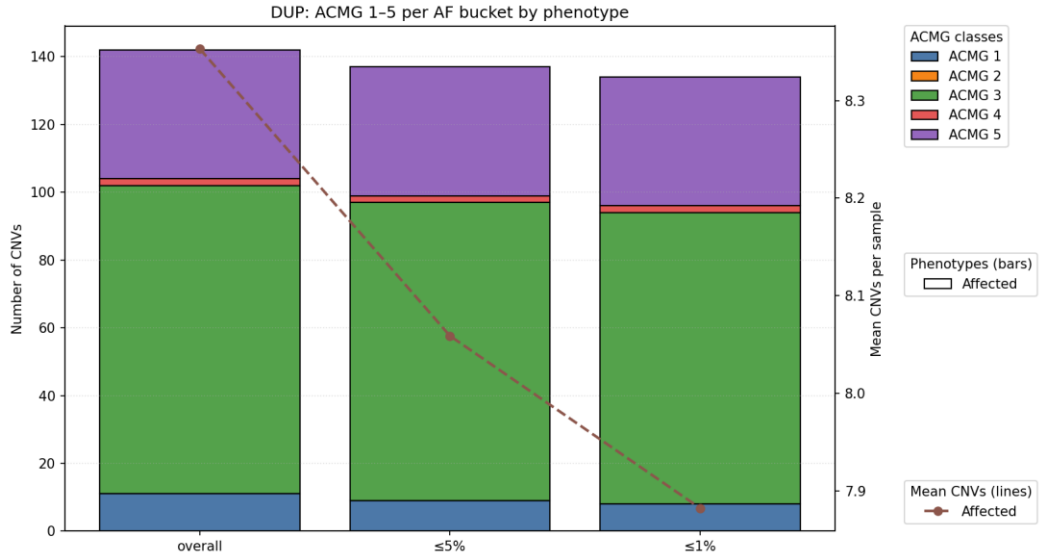

Figure 4.20: Distribution of duplication CNVs (DUPs) within the TP regions only of the CNVPANEL01 samples across ACMG classes 1-5, stratified by allele-frequency (AF) bucket and phenotype. Stacked bars show the total number of CNVs in affected individuals for all variants (overall), variants with AF ≤ 5%, and variants with AF ≤ 1%. Bar colors indicate ACMG

pathogenicity classes as reported in the legend. The dashed line (right y-axis) represents the mean number of CNVs per sample for affected (brown) individuals.

In summary, filtering after intersecting the results maintained base-level and target-level sensitivity close to the unfiltered consensus and comparable to the best individual callers, while substantially reducing the number of events requiring review.

Tool	Base-level sensitivity (%)	Target-level sensitivity (%)	Counts (average/median)
At least 2	89.5	94.3	136/127
At least 2, AF ≤ 5%	89.2	94.1	74/67
At least 2, AF ≤ 1%	89.2	93.9	55/48

Table 4.7: Base- and target-level sensitivity and the CNV counts (mean/median).

4.3.5.3. Exploratory annotation in the Burlo dataset without a validated truth set

I applied the same pipeline used for CNVPANEL01 to the Burlo cohort, using ExomeDepth with the 15 most correlated reference samples, clinCNV on the full cohort, and gCNV on the full cohort, followed by harmonization and AnnotSV v3.5.3 annotation. As for CNVPANEL01, I focused on the at-least-two-caller consensus, defined as CNVs supported by ≥2 callers in each sample. The Burlo dataset comprises 36 affected individuals and 51 unaffected samples. Since no validated truth set is available, this analysis is exploratory: it describes how ACMG classes and frequency filters behave in this cohort, without claiming sensitivity or positive predictive value.

In the at-least-two-caller consensus, the overall CNV load per sample is similar between affected and unaffected individuals. Among affected samples, the consensus set contains 5,459 deletions and 2,164 duplications, with median values of ~69 deletions and ~39 duplications per sample. Among unaffected samples, the corresponding numbers are 5,793 deletions and 2,390 duplications, with medians of ~74 deletions and ~42 duplications per sample.

The ACMG distribution provides a more informative contrast. For deletions in affected individuals, the at-least-two-caller consensus includes 548 likely pathogenic or pathogenic (LP/P, ACMG 4-5), 2,607 variants of uncertain significance (VUS, ACMG 3), and 2,304 benign/likely benign (ACMG 1-2) out of 5,459 total events. This corresponds to ~10% LP/P, ~48% VUS, and ~42% benign. In unaffected samples, deletions show a similar but slightly more benign-skewed profile: 466 LP/P, 2,441 VUS, and 2,886 benign out of 5,793 total deletions (~8% LP/P, 42% VUS, 50% benign). For duplications, both groups are dominated by VUS, but LP/P calls are clearly enriched in affected individuals. Affected samples carry 2,164 duplications in the consensus set, of which 256 (11.8%) are LP/P, 1,693 (78.2%) are VUS, and 215 (9.9%) are benign/likely benign. Unaffected samples, in contrast, have 2,390 duplications but only 51 LP/P (2.1%), with 2,058 VUS (86.1%) and 281 benign (11.8%). In other words, LP/P duplications are ~5-6 times more frequent in affected than in unaffected individuals, while LP/P deletions show only a modest enrichment. This pattern is qualitatively consistent with CNVPANEL01: duplications remain the most ambiguous class overall, but when they are both consensus-supported and labelled ACMG 4-5 they tend to cluster preferentially in affected patients (Figure 4.21, Figure 4.22).

Frequency filtering acts as a second, orthogonal layer that compresses the candidate space and shifts the ACMG composition towards LP/P and VUS, particularly in affected samples. For affected deletions, restricting to ≤5% internal frequency reduces the total from 5,459 to 4,189 events (~23% reduction), while retaining 517 of 548 LP/P deletions (94%); the LP/P fraction rises from 10.0% to 12.3%, and the benign fraction drops from 42.2% to 25.5%. Tightening to ≤1% reduces deletions further to 3,789 events (~31% reduction), with 504 LP/P deletions (92% of the original LP/P) and 2,593 VUS still present; the benign component falls to 18.3% of events. In unaffected deletions, the same filters produce similar relative reductions and a smaller increase in LP/P fraction (from 8.0% to 11.3% at ≤1%), reinforcing that frequency alone does not fully separate cases from controls but preferentially removes clearly benign calls in both groups (Figure 4.21).

For duplications, the filters have an even stronger impact on VUS and benign calls while leaving most LP/P events in place. In affected samples, the ≤5% filter reduces duplications from 2,164 to 1,475 (~32% reduction) and increases the LP/P fraction from 11.8% to 17.3% (255 LP/P), with VUS falling from 1,693 to 1,119 and benign calls from 215 to 101. At ≤1%, affected duplications shrink to 1,245 events (~42% reduction), yet 253 of 256 LP/P duplications (99%) are retained and the LP/P fraction rises further to 20.3%, while benign calls drop to 80 and VUS to 912. In unaffected samples, duplications are pruned even more aggressively: the ≤1% filter removes ~60% of consensus duplications (from 2,390 to 953) and leaves only 44 LP/P duplications (4.6%) among predominantly VUS calls. Taken together, these observations indicate that consensus + ACMG + frequency filters behave in the intended direction: they maintain most LP/P events (and a substantial subset of VUS) in affected individuals, sharply reduce benign calls in both groups, and substantially down-weight rare duplications in unaffected samples (Figure 4.22).

In the absence of a validated truth set, these results cannot be turned into formal performance metrics. However, the pattern is coherent with expectations from the CNVPANEL01 analyses: (i) unaffected individuals can carry many consensus CNVs, but the fraction of ACMG 4-5 events, particularly among duplications, is systematically higher in affected patients; (ii) the at-least-two-caller consensus controls technical noise without collapsing the callset; and (iii) frequency filters act mainly to trim the benign tail and to modestly enrich LP/P and VUS, especially in rare duplications in affected individuals. For clinical triage in this dataset, a pragmatic funnel is therefore to focus on at-least-two-caller LP/P CNVs in affected samples, then to review rare (≤1%) VUS, prioritizing those that overlap ClinGen dosage-sensitive genes or morbid OMIM loci and that match the patient's phenotype, while using the distribution in unaffected samples as an internal sanity check on the specificity of the annotation.

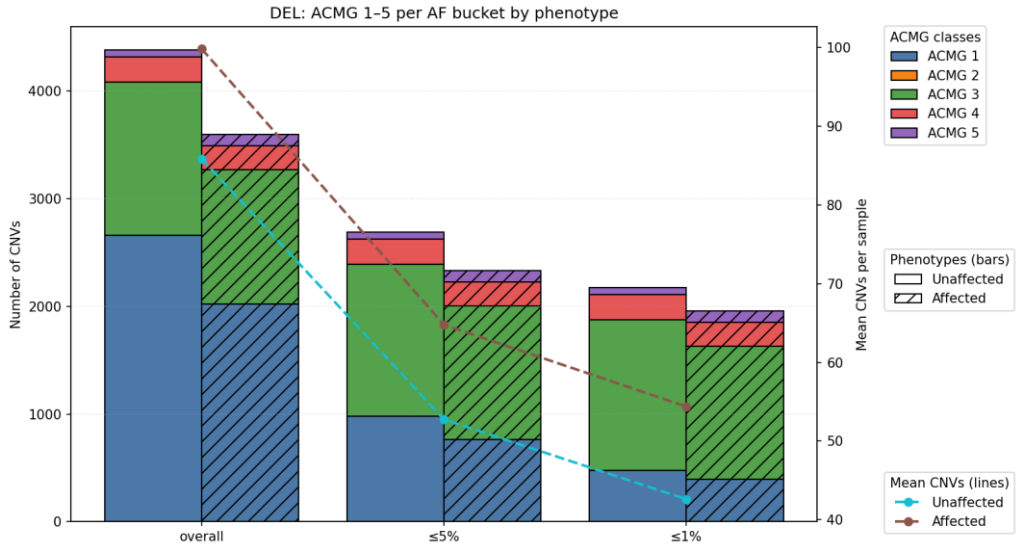

Figure 4.21: Distribution of deletion CNVs (DELs) of the Burlo samples across ACMG classes 1-5, stratified by allele-frequency (AF) bucket and phenotype. Stacked bars show the total number of CNVs in unaffected and affected (hatched) individuals for all variants (overall), variants with AF ≤ 5%, and variants with AF ≤ 1%. Bar colors indicate ACMG pathogenicity classes as reported in the legend. Dashed lines (right y-axis) represent the mean number of CNVs per sample for unaffected (cyan) and affected (brown) individuals.

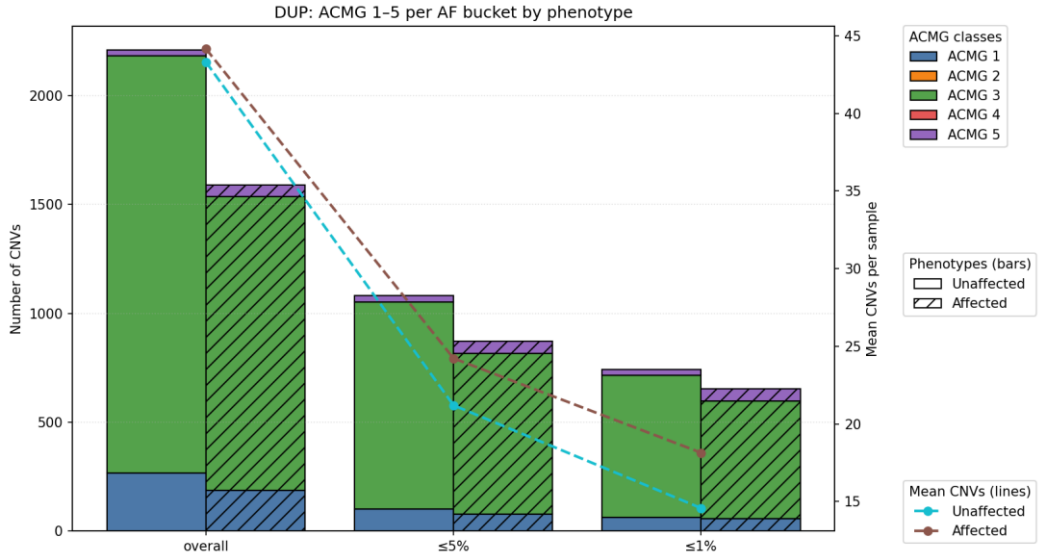

Figure 4.22: Distribution of duplication CNVs (DUPs) of the Burlo samples across ACMG classes 1-5, stratified by allele-frequency (AF) bucket and phenotype. Stacked bars show the total number of CNVs in unaffected and affected (hatched) individuals for all variants (overall), variants with AF ≤ 5%, and variants with AF ≤ 1%. Bar colors indicate ACMG pathogenicity classes as reported in the legend. Dashed lines (right y-axis) represent the mean number of CNVs per sample for unaffected (cyan) and affected (brown) individuals.

4.4.Constellation mapped read: results

4.4.1. Study design and overview of the Constellation mapped read dataset

In parallel to the exome-based CNV analyses, I investigated whether a single Illumina assay could deliver, in one shot, small variants, copy-number changes, and structurally complex events, while also providing long-range information for phasing. To this end, I analyzed six human genomes sequenced with the Constellation mapped read protocol on NovaSeq X, all carrying previously characterized structural rearrangements that served as per-sample truth sets. For one of these genomes (X7675), a matched TruSeq PCR-free WGS dataset was available, allowing a direct comparison between Constellation mapped read and a conventional high-quality short-read workflow.

All samples were processed through the same Constellation mapped read secondary analysis pipeline, which reconstructs long "templates" from spatially co-located read clusters on the flow cell, performs standard mapping and small-variant calling, phases variants and reads, and finally runs DRAGEN modules for CNV and SV detection on haplotagged BAM files. In the following sections I first examine how DNA extraction impacts template reconstruction, then assess coverage uniformity and callability relative to PCR-free WGS, and finally describe small-variant, phasing performance, and structural variant discovery in the six clinical cases.

4.4.2. Impact of DNA extraction on molecular integrity and template reconstruction

The six genomes were extracted using two different high-molecular-weight (HMW) DNA protocols: three samples from a QIAsymphony-based workflow and three from a Bionano ultra-HMW extraction. Both methods satisfied the minimal Constellation requirement in terms of long-fragment content, with at least 10% of the DNA above 60 kb by TapeStation; however, Bionano consistently retained a larger fraction of very long molecules, in some cases exceeding 50% of fragments above 60 kb. Across all runs, raw sequencing throughput was highly comparable,

with about 1.8 billion read pairs per sample and average insert sizes in the expected 300 bp range (Table 4.8).

Sample ID	DNA extraction	Tape Station % > 60kb	# Raw reads	Mean insert size
X7670	QIAsymphony	50.55	1,862,859,514	295
X7673	QIAsymphony	43.8	1,872,062,540	298
X7674	Bionano	56.82	1,827,136,126	333
X7675	Bionano	53	1,801,502,818	346
X7676	Bionano	41.61	1,768,363,984	368
X7677	Bionano	57,29	1,864,966,112	314

Table 4.8: For each sequenced sample, the percentage of DNA sequences above 60kb length the number of sequenced reads and the average insert size.

The insert-size distributions were clearly right-skewed, with long tails that inflate the mean relative to the modal value, consistent with the presence of rare long fragments captured by the on-flow-cell tagmentation process (Figure 4.23).

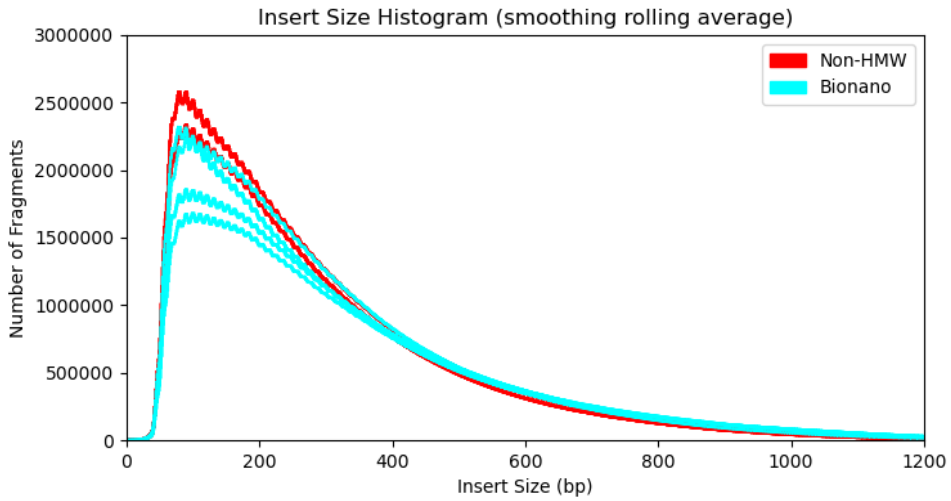

Figure 4.23: Histogram distribution of insert sizes.

When reads were grouped into templates, Bionano-derived samples showed a marked enrichment for long reconstructed molecules. For QIAsymphony genomes, the bulk of templates remained below 50 kb and the 75th percentile was around 40

kb, with only a modest fraction extending beyond 100 kb. In contrast, Bionano samples displayed a much broader tail, with a substantial proportion of templates above 60 kb and individual molecules stretching into the megabase range, with maximum reconstructed spans above 1 Mb in the best cases (Figure 4.24).

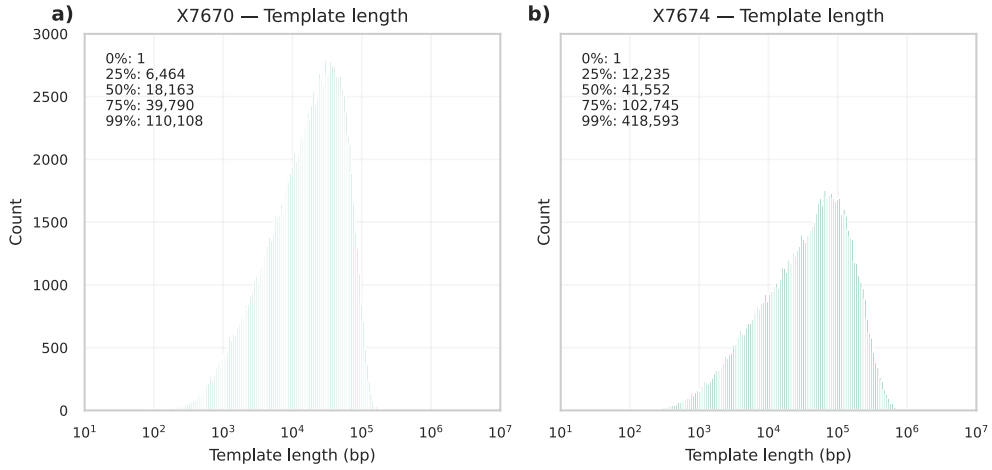

Figure 4.24: Log-scaled histograms of template lengths for X7670 (a, QIAsymphony) and X7674 (b, Bionano), showing the percentile distribution of template sizes.

A complementary view came from counting the number of "subpairs" per template, that is, the number of read pairs assigned to the same reconstructed molecule in representative genomic windows. For both extraction protocols the distributions were heavily skewed, with many templates supported by only a few read pairs and a long right tail of more densely sampled molecules. QIAsymphony templates reached a few tens of subpairs at most, while Bionano templates extended to much higher values, again reflecting the underlying enrichment for very long DNA fragments (Figure 4.25).

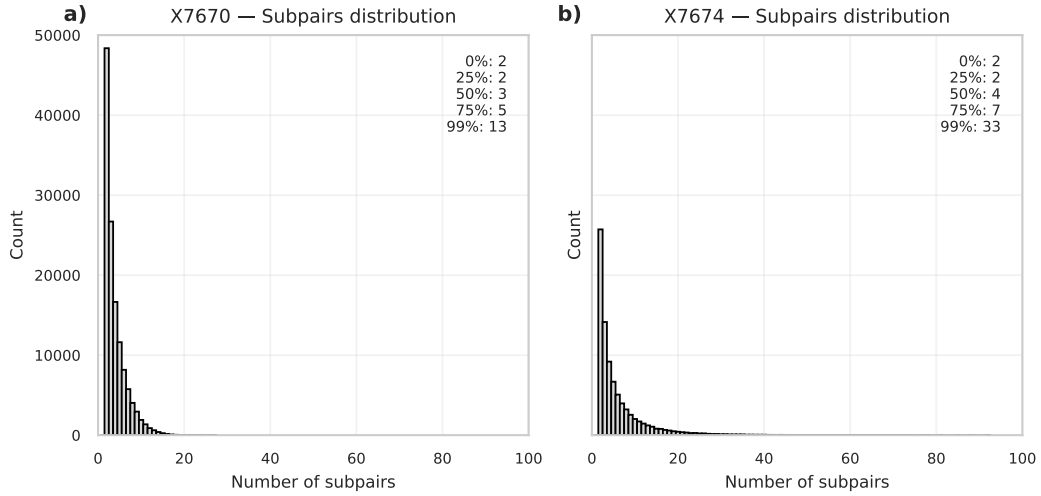

Figure 4.25: Distribution of subpair counts in X7670 (a, QIAsymphony) and X7674 (b, Bionano), showing the number of templates associated with each subpair class together with their percentile ranges.

When plotting template length as a function of subpair count, the two quantities scaled together as expected, but the average genomic distance between adjacent subpairs remained roughly constant at around 10 kb, indicating that Constellation tends to decorate long molecules with read clusters at a quasi-regular spacing rather than saturating them in a localized fashion (Figure 4.26).

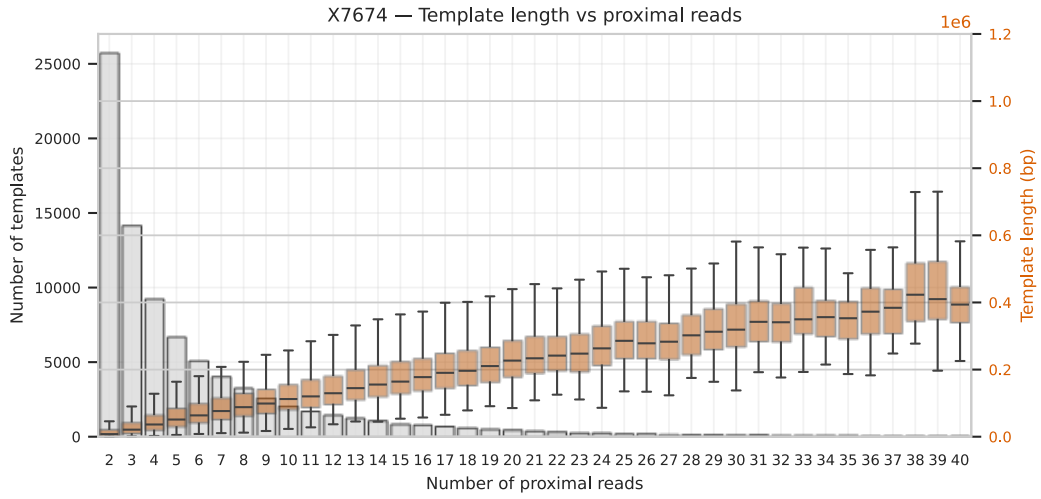

Figure 4.26: Joint distribution of template length and subpair count in X7674 (Bionano), restricted to templates with up to 40 subpairs. The bar plot shows the number of templates per subpair class, while the overlaid boxplots summarize the corresponding template length distributions.

Overall, these results confirm that Constellation faithfully propagates the information content present in the input DNA: when HMW extraction yields truly long molecules, the protocol can reconstruct correspondingly long templates and thereby provide the substrate for extended phasing and long-range SV interpretation. Conversely, when DNA integrity is more modest, the approach still behaves like a high-quality short-read WGS, but its long-range potential is naturally reduced.

4.4.3. Coverage, uniformity and callable genome fraction

A key requirement for integrating Constellation into the same analytical space as standard PCR-free WGS is that it preserves the typical breadth and uniformity of Illumina short-read data. To evaluate this point, I derived classical WGS metrics from the DRAGEN reports for all six Constellation samples and for the matched TruSeq PCR-free dataset of sample X7675.

Constellation runs achieved a mean mapped coverage of roughly 56-65× across the six genomes, with very limited variability between samples. The fraction of bases covered by at least one read was above 95% in all cases, and the fraction covered at ≥10× remained close to 95%, indicating that the higher mean depth translated into an evenly distributed coverage rather than being driven by a subset of highly covered regions. Coverage uniformity, quantified as the percentage of the genome with depth above 20% of the mean, was consistently around 94-95%. The callable genome fraction, defined as the proportion of bases meeting DRAGEN's standard criteria for small-variant calling, was close to 99% for all Constellation datasets (Table 4.9).

The matched TruSeq PCR-free WGS for X7675 showed a lower mean coverage (about 32×) but comparable breadth and uniformity, with coverage ≥1× and ≥10× in the mid-90% range and coverage uniformity above 94%. In other words, once differences in nominal depth are taken into account, the basic mapping behavior of Constellation is essentially indistinguishable from that of a conventional PCR-free WGS library (Table 4.9).

	Sample ID	MEAN coverage	%1X	%10X	% callability	Uniformity of coverage (Pct > 0,2*mean)
Constellation mapped read	X7670	63.99	96.59	95.43	99.47	95.22
	X7673	65.20	95.79	94.71	98.68	94.49
	X7674	61.75	96.56	95.49	99.46	95.29
	X7675	58.62	95.83	94.79	98.65	94.68
	X7676	56.57	95.86	94.81	98.65	94.70
	X7677	63.17	95.83	94.87	98.65	94.71
TruSeq PCR-Free	X7675	31.99	95.24	93.79	98.12	94.09

Table 4.9: The average mapped coverage of the genome, the percentage of genomic bases covered by at least 1X and 10X, the percentage of calleble bases and the percentage of bases covered by at least 20% of the average genome coverage.

Inspection of representative genomic windows in IGV confirmed this impression visually: read alignments from the two technologies produced very similar coverage profiles, without obvious systematic dips or spikes specific to Constellation (Figure S. 2).

Taken together, these observations indicate that embedding spatial information at the flow-cell level and reconstructing templates does not introduce measurable coverage artefacts at the genome scale. Constellation behaves as a standard short-read WGS in terms of depth, breadth, and callability, which is a prerequisite for any fair comparison of small-variant and SV performance against the PCR-free baseline.

4.4.4. Small-variant concordance and long-range phasing

On top of coverage considerations, a single-assay strategy must not compromise single-nucleotide and indel calling. To quantify this, I compared small-variant statistics between Constellation mapped read and the TruSeq PCR-free dataset. For each Constellation genome, DRAGEN identified approximately 5.1 million small variants, of which about 4.1 million were SNVs and roughly 0.97 million were indels, figures that fall within the expected range for high-coverage human WGS.

The matched PCR-free library for X7675 yielded a very similar total variant count, again around 5.17 million calls, confirming that Constellation does not inflate or suppress the overall small-variant burden.

Where Constellation diverges from standard WGS is in its phasing performance. Because templates link together reads originating from long HMW molecules, Whatshap can propagate phase information over extended genomic distances. In the six Constellation genomes, around 97% of all small variants were assigned to a phase block, with average block lengths in the hundreds of kilobases and maximum spans reaching tens of megabases in Bionano-extracted samples. Block NG50 values, computed over the distribution of phased segments, frequently exceeded 10-20 Mb for the best Bionano genomes, illustrating the ability of Constellation to reconstruct chromosome-scale haplotypes from standard 2×150 bp reads.

The PCR-free library of X7675, phased with the same Whatshap pipeline but without access to proximity information, showed a very different behavior. Although the total number of variants was essentially the same, only about 86% could be phased, and the resulting blocks were much more fragmented: hundreds of thousands of short segments with average lengths of the order of 1-2 kb and no long-range continuity. In practice, this means that while both technologies detect the same single-nucleotide changes, only Constellation provides a haplotypic scaffold that can bridge across large structural events and support allele-specific interpretation (Table 4.10).

	Sample ID	Total	SNV	Inde	% variants phased	Nr phased blocks	Avg phased blocks length	Max phased blocks length	Phased blocks NG50
Constellation mapped read	X7670	5,079,741	4,113,184	966,557	96.75	8,448	261,877	13,230,583	1,932,982
	X7673	5,123,694	4,144,632	979,062	96.50	11,300	205,173	9,566,299	1,407,597
	X7674	5,110,332	4,134,627	975,705	97.22	4,257	542,728	75,518,919	26,376,015
	X7675	5,168,049	4,181,037	987,012	97.23	3,431	657,690	91,106,493	31,365,187

	X7676	5,114,844	4,138,433	976,411	97.27	3,522	636,191	73,368,490	20,969,197
	X7677	5,105,641	4,131,944	973,697	97.10	4,858	459,009	49,382,676	10,188,283
TruSeq PCR-Free	X7675	5,174,432	4,185,369	989,063	86.30	411,251	1,481	2,368,810	N.D.

Table 4.10: Small variant calling and phasing performance for both Constellation mapped read and TruSeq PCR-free samples. Reported metrics include total small variants, SNVs, and Indels, phasing percentage, and phased block statistics (number of blocks, average block length, maximum block length, and NG50).

These results highlight a central message of this section: Constellation maintains the small-variant performance of PCR-free WGS but upgrades it with long-range phasing, and the extent of this upgrade is tightly coupled to the quality of the input DNA, with ultra-HMW extraction yielding the most impressive block lengths.

4.4.5. Structural variant and copy-number calls in Constellation versus PCR-free WGS

Given the upstream focus of this thesis on CNVs and SVs, an obvious question is how Constellation mapped read behaves in terms of structural variant discovery. The Constellation pipeline produces three main callsets: a read-depth-based CNV VCF, a junction-based SV VCF (including deletions, insertions, duplications and breakends), and an integrated CNV_SV set that retains only events supported by both depth and breakpoints, with a particular emphasis on shorter CNVs below the nominal detection limit of pure read-depth methods.

Across the six Constellation genomes, the junction-based SV caller reported on the order of 27-29 thousand events per sample, dominated by deletions and insertions, with a smaller number of tandem duplications and breakend pairs. The CNV module contributed around 130-180 copy-number events per genome, largely multi-exonic or multi-kilobase gains and losses. The integrated CNV_SV callset contained roughly 1.7-1.8 thousand events per Constellation sample, representing the subset for which depth and junction evidence were consistent (Table 4.11).

For the individual sequenced with both technologies (X7675), these numbers can be directly compared. In the Constellation dataset, the CNV caller reported 127

events, the SV caller about 28.7 thousand junction-level variants, and the CNV_SV module 1,779 integrated events. In the PCR-free dataset, using the same DRAGEN version, the CNV caller identified 102 copy-number events, the SV caller about 20 thousand variants, and the CNV_SV module 1,283 integrated events. Thus Constellation produced slightly more CNVs, a substantially larger SV callset, and a higher number of events with concordant depth and junction support (Table 4.11).

		SV caller					CNV caller			CNV_SV callers		
	Sample ID	**TOTAL**	**Deletions**	**Insertions**	**Duplications**	**Breakend pairs**	**TOTAL**	**Deletions**	**Duplications**	**TOTAL**	**Deletions**	**Duplications**
Constellation mapped read	**X7670**	27,37	9,258	15,645	134	2,333	160	79	81	1.765	1.556	209
	X7673	27,566	9,262	15,886	116	2,302	136	71	65	1.642	1.466	176
	X7674	28,851	9,488	16,240	152	2,971	180	62	118	1.834	1.571	263
	X7675	28,686	9,595	16,271	110	2,71	127	58	69	1.779	1.605	174
	X7676	28,634	9,408	16,188	139	2,899	139	68	71	1.781	1.578	203
	X7677	27,896	9,314	15,863	153	2,566	126	52	74	1.732	1.508	224
TruSeq PCR-Free	**X7675**	19,964	6,492	11,395	166	1,911	102	60	42	1,283	503	157

Table 4.11: Number of SV/CNV events per caller, reported as total and by variant class (DEL, INS, DUP, BND).

Given that both datasets were analyzed with the same software and reference, these differences are unlikely to reflect trivial pipeline discrepancies. Rather, they suggest that proximity-informed mapping and haplotagging provide additional sensitivity for detecting and refining structural events, particularly in regions where conventional short reads struggle with mapping ambiguity. In the context of this thesis, this is relevant because it offers a complementary perspective to the earlier WES CNV analyses: while exome CNV callers required careful GC-aware normalization and extensive reference tuning to keep false positives under control, Constellation starts from a more stable WGS substrate and enriches it with long-range context to improve SV and CNV resolution.

4.4.6. Clinical case analysis

To assess how Constellation performs on clinically relevant structural variants, I re-analyzed six diagnostic genomes from Meyer Children's Hospital, each carrying a previously established rearrangement used as a per-sample truth set. These alterations had been originally resolved by combinations of MLPA, array-CGH, karyotyping, FISH, long-read sequencing, optical mapping, short-read WGS or WES, and were not discovered de novo in this work. Instead, Constellation was evaluated on its ability to retrieve, refine and phase these events using only short reads augmented with proximity information.

4.4.6.1. Sample X7670 - Complex rearrangement at the SCN1A locus

Sample X7670 contains a complex structural variant centered on the SCN1A gene. Using Constellation, I combined read-depth information, junction-based SV calls, short-range variant phasing and long-range proximity maps to reconstruct the structure of the rearranged allele.

For interpretation, the region was divided into five consecutive segments, named A to E. The CNV caller showed two copy-number gains at segments B and D, separated by a copy-neutral segment C. This pattern is compatible with an allele in which B and D are both duplicated and interleaved with a retained copy of C. The more sensitive CNV_SV caller, designed to rescue smaller CNVs supported by breakpoints, additionally reported a ~2.6 kb deletion confined to one of the duplicated B segments. Finally, the SV caller identified a tandem duplication of ~56.5 kb spanning the entire B-D interval, together with two deletion calls corresponding to the internal loss within B and to the segment C interval (Figure 4.27).

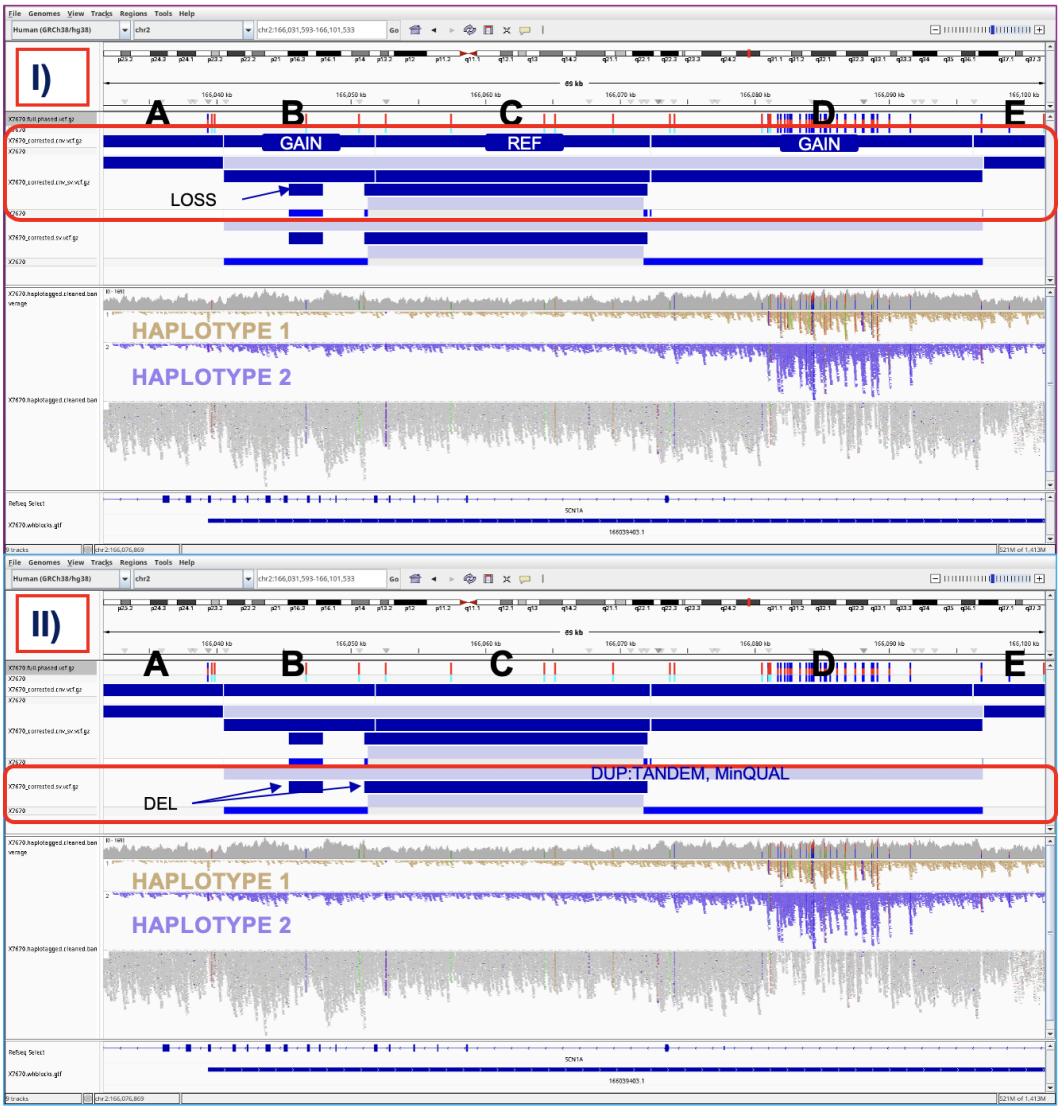

Figure 4.27: Sample X7670: IGV screenshot of a 69kb window, of haplotyped alignment file, together with phased variant calling file, CNV, CNV_SV, and SV variant tracks. I) The read-depth based CNV variant calling identified two copy-number gains corresponding to segments B and D, interspersed with reference C, outlining the overall duplicated architecture of the region. The CNV_SV variant caller detected a 2.6 kb internal deletion within segment B. II) The junction-based SV variant calling algorithm detected a MinQUAL 56.6 Kb tandem duplication encompassing the entire rearranged locus and 2 deletion events, one whitin B segment and one associated to C region. Haplotype architecture resulted precisely delineated in the D segment where there is a concentrate grade of genetic variation.

Haplotype-resolved views at SCN1A showed that segment D, which harbours a cluster of heterozygous variants, falls within a well-defined phase block, allowing the rearranged configuration to be assigned to a single chromosome. When

alignments were grouped as chimeric reads in IGV, six breakpoint junctions could be inspected directly. Split reads connect the beginning of B to the end of D and vice versa, and additional reads support the ~2.6 kb deletion within one of the B copies. Together, these observations support a model in which one allele carries a B-D-B configuration, with only one B copy harbouring the internal deletion (Figure 4.28).

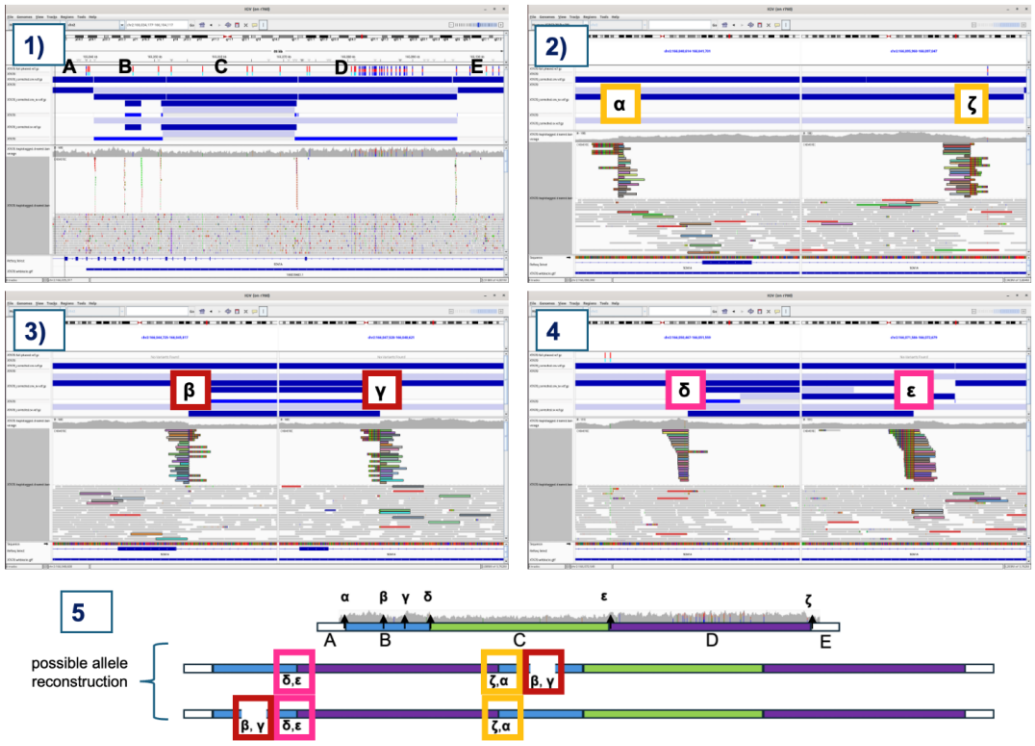

Figure 4.28: Sample X7670: Structural variant reconstruction based on split reads. IGV visualization was performed by setting Group by > Chimeric, and using View Chimeric Alignments by Split Screen for each split read, allowing direct inspection of sequence reads that align to two distinct genomic regions. 1) Overview of the six breakpoint junctions identified within the rearranged locus. In this view, each split read is represented as a read whose sequence is partially clipped relative to its original alignment, corresponding to the breakpoint position. 2) Alignment of the starting portion of segment B (α) with the terminal portion of segment D (ζ). 3) Alignment of the terminal part of B (δ) with the starting region of D (ε). 4) Split-read alignments supporting the 2.6 kb deletion within segment B, confirming the internal structural discontinuity within one of the duplicated copies. 5) Most likely allele reconstructions.

Proximity maps add an orthogonal spatial perspective on the same event. In colocation matrices limited to the SCN1A target region, bins corresponding to

segments B and D exhibit a strong off-diagonal signal, indicating frequent contacts between these non-contiguous genomic intervals. Within this block, the portion of B that is deleted on one copy shows a relative drop in contact density, consistent with the loss of material from only one allele. Combining copy-number changes, breakpoint calls, haplotype information and proximity patterns yields a consistent allele-level model of the SCN1A rearrangement that would be difficult to derive from standard short-read WGS alone (Figure 4.29).

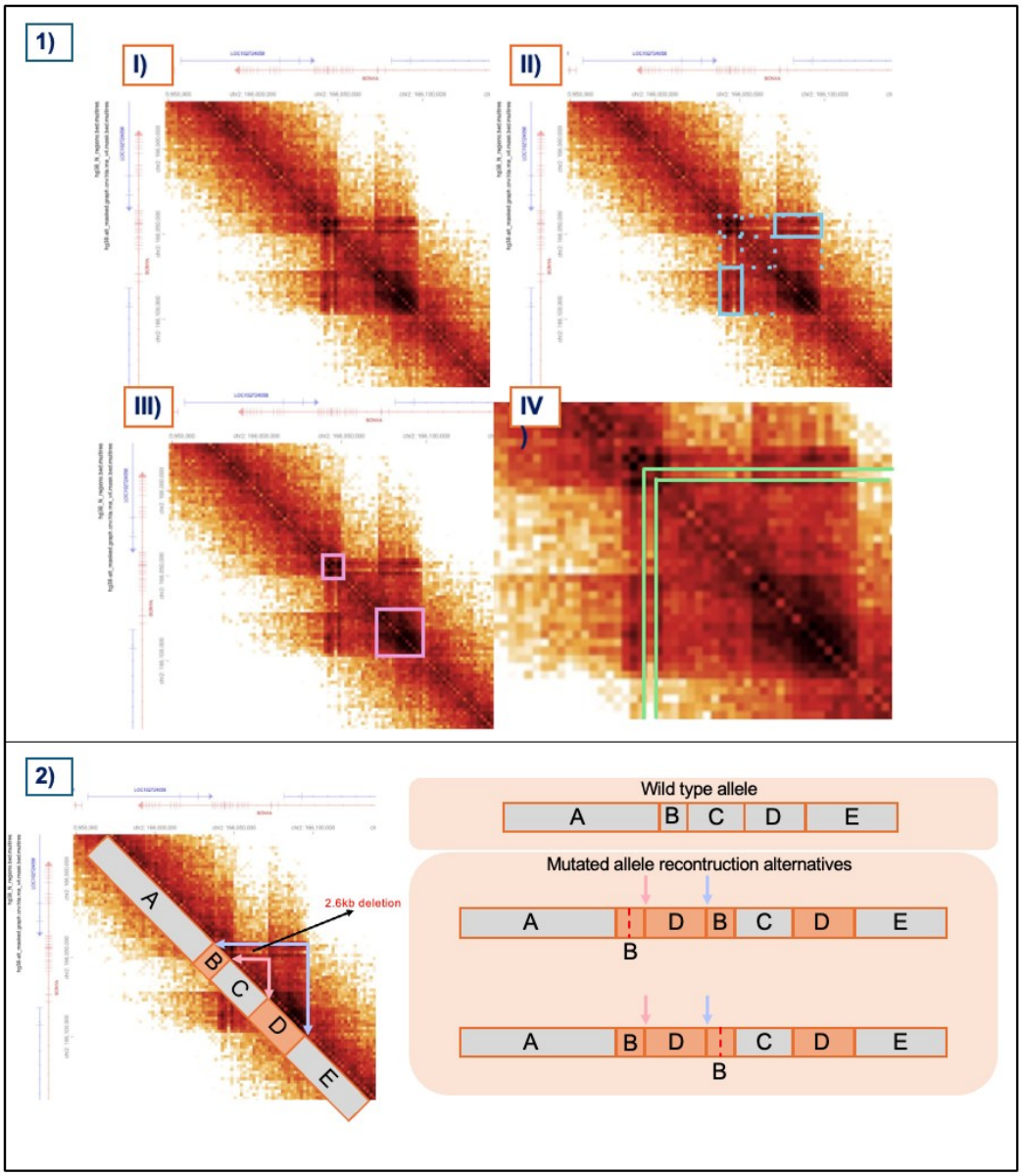

Figure 4.29: Sample X7670: 1) Four colocation plots of the target SCN1A locus, each emphasizing different aspects of the structural variant. Each matrix value indicates the number of read pairs in

2kb bin from SCN1A target region that are in proximity to the same SCN1A target region. Bins with fewer than 15 counts were excluded to improve the overall signal-to-noise ratio. I) Whole region involving the SCN1A structural rearrangement. III) Two duplicated regions. II) Spatial proximity of the larger duplicated region between the duplicated smaller region. IV) Small deletion within the smallest duplicated region. 2) Possible allele reconstruction inferred from proximity-based data: orange segments denote duplicated regions, while gray areas correspond to copy neutral sequence. The terminal portion of B aligns with the starting sequence of D, and conversely, the end of B connects with the beginning of D, supporting the interaction between non-contiguous genomic regions.

4.4.6.2. Sample X7677 - Multi-step rearrangements between chromosomes 2 and 4

Sample X7677 illustrates a different form of complexity, involving multiple lesions distributed across chromosomes 2 and 4. Previous clinical work had described a large deletion on 4p, insertion of a ~2 Mb fragment from chromosome 2 into chromosome 4, and a balanced reciprocal translocation. Constellation recovered all three components and linked them on a phased, allele-specific backbone.

The SV caller detected the breakpoints at the boundaries of the 2 Mb fragment on chromosome 2 (segment 2B) and at its insertion site between segments 4B and 4C on chromosome 4. Read-depth analysis identified a broad heterozygous loss over a 17.7 Mb interval on 4p, matching the previously reported deletion. In the intrachromosomal contact maps, this deleted region appears as a band of reduced contact density, whereas interchromosomal colocation maps revealed focal interaction blocks connecting the relocated 2B segment to its new position on chromosome 4, as well as a distinct signal corresponding to a balanced translocation between distal regions 2D and 4D (Figure 4.30).

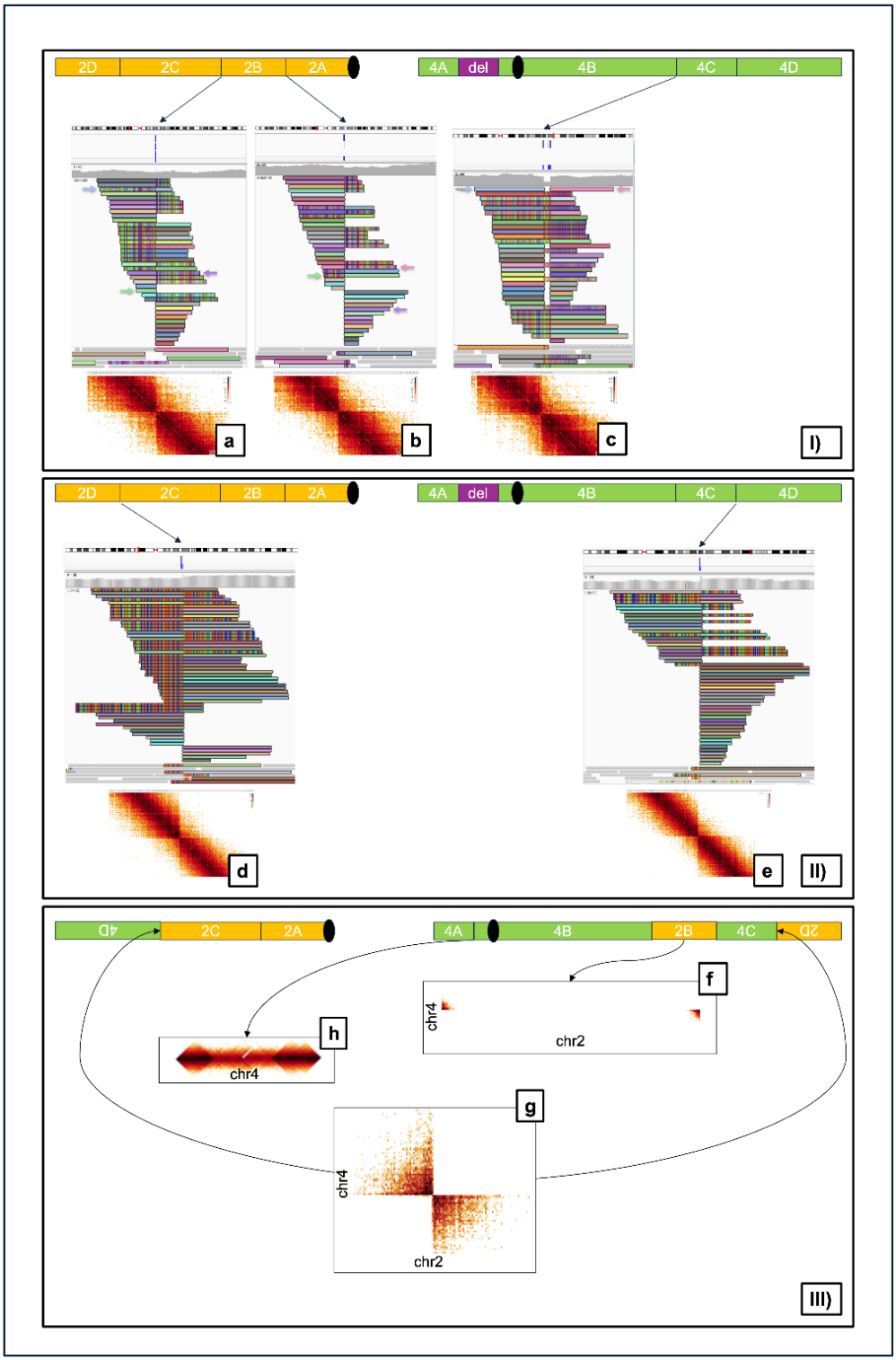

Figure 4.30: Sample X7677: Interchromosomal rearrangement. (I) Breakpoints associated with the excision of segment 2B from chromosome 2 and its translocation between segments 4B and 4C on chromosome 4. In the contact map of chromosome 2, two breakpoints (a, b) mark the boundaries of the excised region, while in the chromosome 4 map, a single breakpoint (c) identifies the translocation site for segment 2B. (II) Breakpoints associated with the balanced translocation

between segments 2D and 4D, corresponding to junctions (d,e). (III) Reconstruction of the composite complex rearrangements connecting chromosomes 2 and 4. Proximity signals (f) show contacts between segment 2B and regions 4B and 4C, while signal (g) marks the balanced translocation connecting chromosomes 2 and 4. Reduction in contact density across bins (h) corresponding to the 17.7 Mb deleted region on chromosome 4.

Haplotagged IGV views clarified the allelic configuration of the deletion on chromosome 4. Across the 17.7 Mb interval, one haplotype shows a pronounced drop in coverage and breakpoint-spanning reads, while the other haplotype retains normal depth. The proximal and distal breakpoints of the deletion are supported by chimeric reads tagged to the same haplotype as the insertion and translocation events, indicating that the excision of the 2 Mb fragment from chromosome 2, its insertion into chromosome 4, and the distal reciprocal translocation all occur on the same derivative chromosome. Thus, in this sample Constellation does not merely detect individual SVs, but reassembles them into a coherent, megabase-scale haplotype-level model (Figure 4.31).

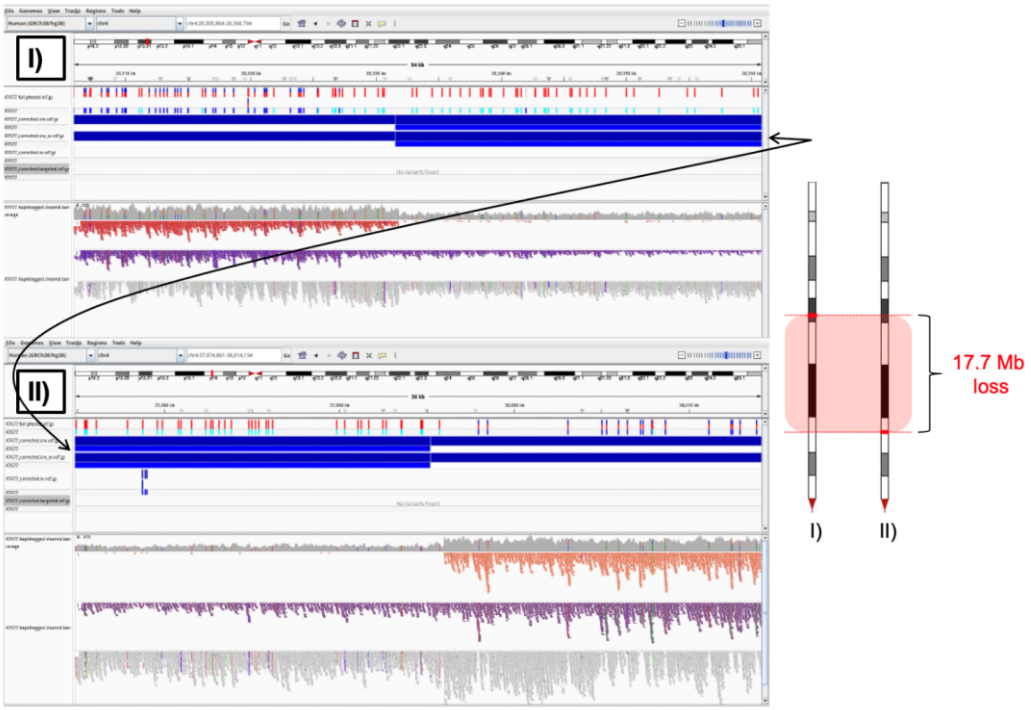

Figure 4.31: Sample X7677: IGV screenshots of Haplotyped BAM alignment. Panels I-II illustrate the proximal and distal breakpoints of the 17.7 Mb deletion on the short arm of chromosome 4. Haplotype 1 shows a copy-number loss, whereas haplotype 2 retains read coverage.

4.4.6.3. Sample X7674 - t(19;22) and duplication on 19q

Sample X7674 carries two lesions: a balanced t(19;22)(q13.4;p11.2) translocation and a large duplication spanning 19q13.33-q13.43. Junction-based SV calling recovered the interchromosomal breakpoints between chromosomes 19 and 22, confirming the presence of the translocation. However, proximity maps centred on the expected breakpoint regions did not display a clean, high-contrast signal. Both intrachromosomal interaction maps and the 19↔22 colocation matrix contained only weak, diffuse contacts, which is compatible with the challenges of interpreting spatial patterns in centromeric and pericentromeric regions of GRCh38.

In contrast, the duplication on 19q produced a robust signature at both the read-depth and proximity levels. The CNV caller reported a series of adjacent copy-number gains covering the q13.33-q13.43 interval on chromosome 19, and the corresponding portion of the colocation map appeared as a distinct block of increased contact density. This local enrichment of contacts, confined to the duplicated interval, reinforces the CNV call and illustrates how long-range information can help to validate broad copy-number changes even when associated ranslocations do not yield an easily interpretable colocation pattern (Figure 4.32).

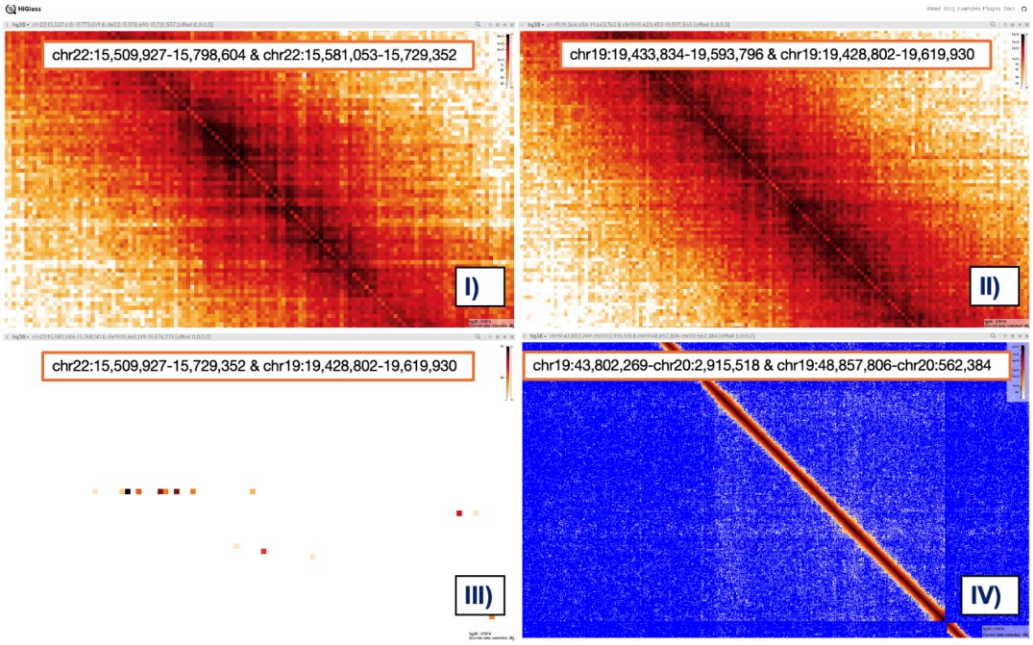

Figure 4.32: Colocation plots of structural variants detected in sample X7674. Panels I-II display

the expected breakpoint regions on chromosomes 22 and 19, respectively, corresponding to the t(19;22)(q13.4;p11.2) translocation. In both cases, no defined breakpoint signal is observed in the proximity interaction maps, suggesting weak or unresolved contact patterns at the translocation junctions. Panel III shows the interchromosomal contact map between chromosomes 19 and 22, in which only low-intensity proximity signals are detected, not matching the expected translocation pattern. Panel IV depicts the duplication on chromosome 19 (19q13.33-q13.43), where the duplicated interval is clearly visible as a localized enrichment in contact density, highlighted within the boxed region. A zero-value color scale (blue) was applied to enhance contrast and emphasize the duplicated signal.

4.4.6.4. Sample X7673, X7675 and X7676 - Smaller events and balanced translocation

The remaining three samples comprise structurally simpler, but clinically relevant, alterations that were also correctly recovered by Constellation.

Sample X7673 harbours a heterozygous ~2 kb deletion involving *STXBP1*. The event falls below the typical resolution of read-depth CNV calling, but was detected by the CNV_SV module through its combined use of split-read and discordant-pair evidence and subsequently confirmed by IGV inspection (Figure 4.33).

Sample X7675 carries a balanced t(3;8)(q21.3;q22.1) translocation. The SV caller identified the reciprocal exchange between chromosomes 3 and 8, and haplotype-aware views showed that the involved segments originate from opposite haplotypes, consistent with a classic balanced translocation between homologous chromosomes. Interchromosomal contact maps displayed the expected cross-shaped interaction pattern at the junction (Figure 4.33).

Finally, sample X7676 presents a tandem duplication on chromosome 18. This lesion was visible both as a local copy-number gain and as a pair of SV breakpoints marking the duplication boundaries. Haplotagged alignments indicated that the duplication is restricted to a single haplotype, and the colocation map showed a characteristic off-diagonal block, again consistent with a tandemly duplicated configuration (Figure 4.33).

Together, these three cases illustrate that Constellation can robustly detect and interpret heterozygous microdeletions, balanced translocations, and focal tandem duplications, complementing the more complex multi-step rearrangements observed in X7670 and X7677.

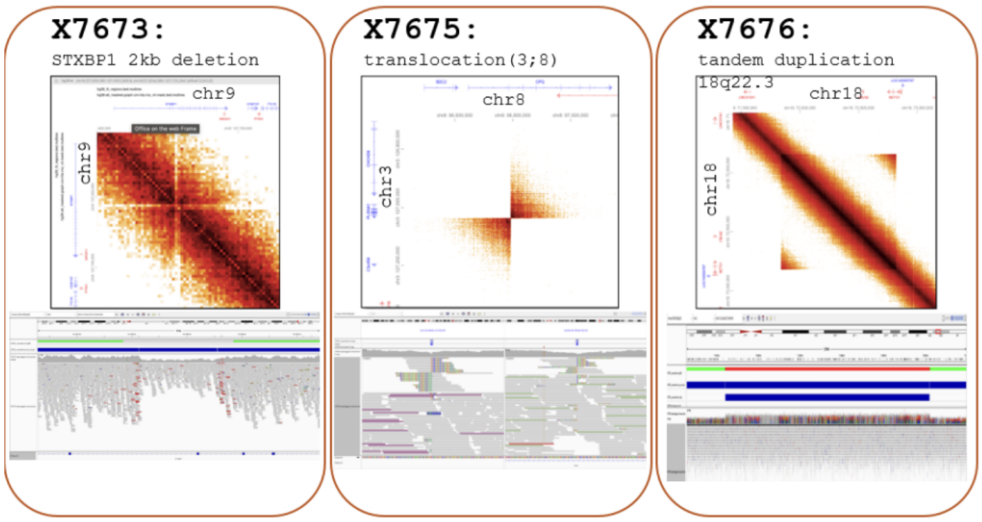

Figure 4.33: Colocation plots and IGV screenshots of structural variants identified across three samples. Sample X7673: heterozygous 2kb deletion of STXBP1 gene. Sample X7675: balanced interchromosomal translocation t(3;8) (q21.3;q22.1), in which a segment from chromosome 3 haplotype 2 is reciprocally exchanged with a region on chromosome 8 haplotype 1. Sample X7676: tandem duplication on chromosome 18, specifically on haplotype 2.

From the perspective of this thesis, these case studies are important for two reasons. First, they show that Constellation can handle the same spectrum of structural complexity that motivated the use of arrays, long reads and optical mapping in the first place, but within a single Illumina run. Second, they illustrate how long-range information and haplotyping are not abstract metrics but concrete tools for disentangling multi-step rearrangements and for assigning copy-number changes to specific alleles, something that is highly relevant when moving from cohort-level benchmarking to patient-level interpretation.

5. Discussion

The work presented in this thesis sits at the intersection of clinical genomics and methodological development. Despite the widespread implementation of whole exome sequencing (WES) and whole genome sequencing (WGS) for patients with suspected genetic disorders, a substantial fraction of cases remain without a molecular diagnosis, even in experienced centers that already apply state of the art pipelines for single nucleotide variants (SNVs) and small insertions and deletions (indels) [1], [5]. In parallel, decades of research have shown that structural variants (SVs) and copy number variants (CNVs) contribute extensively to human phenotypic diversity and disease, often through mechanisms that differ from those of small variants [18], [23], [24], [31]. This gap between current diagnostic practice and the known impact of CNVs and SVs motivates the two main aims of the thesis: stabilizing CNV detection from WES and exploring proximity aware Illumina Constellation mapped read WGS as a richer SV assay.

5.1. Copy number variants from exome sequencing

The first part of the thesis focuses on CNVs derived from WES. Exomes remain a central workhorse in clinical genomics and it is tempting to exploit them as a source of CNV information, but in practice this requires careful modelling of read depth across targets, the construction of suitable panels of normal samples and an explicit awareness of the limitations of targeted assays.

The analysis of Twist exomes illustrates both the potential and the fragility of this strategy. When EXCAVATOR2 was applied to these data, the resulting CNV landscapes were clearly incompatible with biological expectations, with very few duplications and deletions and multi megabase events recurrently observed in both probands and parents, a pattern difficult to reconcile with Mendelian disease genetics and more consistent with systematic artefacts driven by the underlying model (Figure 4.1) [62]. ExomeDepth, which uses a Bayesian framework to compare each sample to a set of reference exomes, corrected many of these extremes and produced more plausible profiles, but generated an inflated number of calls in some libraries and showed strong sensitivity to coverage behavior [63].

A key observation is that this inflation and oscillation tracks GC behavior of the libraries (Figure 4.2). Coverage plots across GC content revealed that some exomes display smooth, symmetric profiles whereas others show pronounced bumps and troughs at specific GC fractions. These coverage waves can mimic the signal of true CNV segments and lead to clusters of calls that are not reproducible across libraries, making exome CNV calling not only a function of the software but also of subtle differences in library preparation and target design that manifest through GC dependent coverage distortions. To move beyond qualitative impressions and obtain a more systematic assessment, the thesis introduces a Snakemake pipeline that integrates three exome CNV callers, ExomeDepth, ClinCNV and gCNV, under harmonized preprocessing and configuration [63], [64], [65], [67]. Two cohorts are analyzed within this framework. The first is CNVPANEL01, a Twist panel with curated CNVs that serves as a benchmark; the second is a heterogeneous clinical cohort from the Burlo Garofalo Hospital, which provides a realistic spectrum of in house exomes but lacks a comprehensive gold standard [52]. On CNVPANEL01, the three tools achieve broadly similar sensitivities for the validated CNVs, especially for deletions, which tend to be more stable than duplications (Table 4.3, Table 4.4), but differ markedly in the size and structure of their call sets: ExomeDepth often produces fine grained segmentations with many small intervals, ClinCNV tends to generate fewer and larger events, and gCNV benefits from joint modelling of all samples and can capture more subtle shifts at the cost of increased computational complexity (Figure 4.9).

These results emphasize that sensitivity on a curated set, although important, represents only one dimension of performance, because for a diagnostic laboratory the effective burden is determined by how many events require manual review per patient. From this point of view, ClinCNV alone represents the best compromise.

Unlike SNVs and small indels, where Genome in a Bottle and similar efforts provide dense, genome wide truth sets, there are very few gold standard resources for CNVs. Available SVs truth sets are typically limited to specific samples and enriched for deletions and simple insertions, while exome based resources such as CNVPANEL01 usually provide one or two curated CNVs per case. As a

consequence, evaluation is largely restricted to estimating sensitivity on a small number of events, with limited power to quantify specificity or false positive rates, and no straightforward way to assess how many truly negative targets are incorrectly called as altered.

Within these constraints, the pipeline has been used to explore how the composition of the reference panel affects performance, reinforcing the notion that CNV calling from exomes is inherently relative. While SNVs and small indels can be called from single samples against the reference genome, exome CNVs rely on contrasts in read depth against a pool of normals that define the expected coverage at each target. In this work, the baseline has been varied from an idealized panel composed solely of CNVPANEL01 samples to a more heterogeneous set that combines CNVPANEL01 and Burlo exomes. ExomeDepth is particularly sensitive to this choice, with heterogeneous and non-optimized baselines leading to explosions in call counts and multiple artefacts, whereas correlation based selection of normals, so that each case is compared to the most covarying reference set, yields more stable and interpretable calls (Figure 4.14). ClinCNV shows greater robustness to baseline heterogeneity at the price of a modest reduction in sensitivity (Figure 4.15), and gCNV performs best when many samples are available and degrades on small tightly defined panels (Figure 4.16).

To translate these insights into something usable in a diagnostic workflow, the pipeline implements a meta caller based on consensus. Rather than accepting all calls from each tool, unions and intersections of call sets are constructed, and the intersection of at least two callers emerges as a pragmatic compromise that retains most validated events at both region and base pair level while substantially reducing the number of candidate CNVs per sample (Table 4.5, Table 4.6). This consensus set is then annotated with AnnotSV and filtered using ACMG categories and population frequency thresholds [82]. On CNVPANEL01, the combination of consensus calling and frequency cutoffs removes benign and common events while preserving likely pathogenic and pathogenic CNVs (Figure 4.17, Figure 4.18). On the Burlo cohort, where no exhaustive truth set is available, formal performance metrics cannot be computed, but the behavior is consistent with expectations:

affected and unaffected individuals harbor a similar number of consensus CNVs, yet likely pathogenic and pathogenic events, particularly duplications, are several times more frequent in affected cases, and moderate frequency filters mainly trim the benign tail and reduce rare duplications in unaffected individuals while preserving clinically relevant events together with a subset of VUS for review (Figure 4.21, Figure 4.22).

Overall, the exome work supports a pragmatic and cautious conclusion. CNV detection from WES is far from a solved problem and the field is split across software that make different assumptions about data structure and about how reference panels should be constructed. Some methods expect small sets of tightly correlated normals, others benefit from large heterogeneous cohorts, and their performance degrades when these conditions are not met. Targeted designs only cover coding regions and a subset of untranslated exons, leaving many regulatory and intergenic CNVs partially or completely invisible, and even within the exome coverage can be highly inhomogeneous, making it difficult to distinguish true copy number changes from persistent under coverage. Different statistical models, from simple depth comparisons to hidden Markov models and joint latent factor approaches, infer CNVs from noisy, design dependent signals that consensus approaches can attenuate but not fully correct. Within this landscape, the analyses in the thesis show that it is possible to obtain clinically useful CNV call sets from exome data by combining an explicitly engineered baseline, multiple callers and a structured annotation and filtering funnel, but also that exome based CNV calling alone cannot close the diagnostic gap that remains after SNV and indel analysis. These limitations set the stage for the evaluation of Constellation mapped read in the second part of the thesis.

5.2.Constellation mapped read as an integrated assay

The second part of the thesis explores Illumina Constellation mapped read as an integrated whole-genome assay in which standard short-read sequencing is augmented with long-range information encoded by the physical proximity of read clusters on the flow cell [51]. In Constellation mapped read, high-molecular-weight DNA is fragmented and captured directly on a modified NovaSeq X flow cell;

clusters that arise from the same long molecule remain spatially close and are grouped into reconstructed “templates” during DRAGEN analysis [7]. These templates provide a linkage signal that extends over hundreds of kilobases and can be exploited during mapping, phasing and SVs calling. The central question in this part of the work is whether this proximity-aware configuration can maintain the strengths of PCR-free WGS for SNVs and INDELs detection while adding phasing and structural resolution that move closer to what is typically obtained with long-read or optical-mapping technologies.

A first observation concerns DNA extraction. Although the sequencing chemistry and secondary analysis are identical across samples, different extraction protocols lead to distinct template-length distributions. In the genomes analyzed here, QIAsymphony and Bionano extractions produced libraries with markedly different long-range properties despite comparable nominal depth. This reinforces the idea that the main bottleneck for Constellation mapped read is not the sequencer but the availability of sufficiently long and intact molecules. In fact, I observed that samples with longer input molecules generate longer reconstructed templates and a stronger proximity signal (Table 4.8).

At the level of SNVs/INDELs calling, Constellation mapped read behaves very similarly to a standard PCR-free WGS library. Coverage, callability and concordance with a matched TruSeq dataset are high, and there is no detectable loss of sensitivity (Table 4.9, Table 4.10). The real benefit appears when phasing and structural rearrangements are considered. Because reads are linked into long templates through their shared spatial origin on the flow cell, variants can be phased into extended haplotype blocks that span large genomic intervals. In this dataset, phasing routinely reaches tens of megabases, whereas standard short-read phasing based on read pairs produces much shorter and highly fragmented blocks.

This long range information is particularly valuable in the setting of complex structural rearrangements. In the clinical cases analyzed here, Constellation enables a richer reconstruction of events such as the multi-step SCN1A duplication in X7670 and the composite rearrangements between chromosomes 2 and 4 in X7677, as well as simpler but clinically relevant lesions in X7673, X7674, X7675 and

X7676. In X7670, Constellation combines depth of coverage, split reads, template based linkage and colocation maps to move from a list of CNVs and breakpoints to an allele level model of the SCN1A locus. Copy number callers identify two gains separated by a copy neutral segment, the SV module refines the tandem duplication and the internal deletion, and haplotype aware views show that all junctions map to the same phased chromosome. Proximity maps add an additional layer by revealing strong contacts between the duplicated segments, supporting a B-D-B configuration with an internal deletion confined to one of the duplicated copies (Figure 4.29). In X7677, the same combination of signals is used to disentangle a 17.7 Mb deletion on 4p, the excision and insertion of a 2 Mb segment from chromosome 2 and a distal balanced translocation between chromosomes 2 and 4. Read depth profiles, junction based calls and colocation matrices together show how these lesions are chained on a single derivative chromosome, so that Constellation does not merely rediscover individual SVs but reconstructs a coherent, phased model of the rearranged genome (Figure 4.30). The remaining genomes illustrate that this is not restricted to the most complex cases. In X7673, a heterozygous STXBP1 microdeletion below the usual resolution of pure depth based CNV callers is recovered by the integrated CNV_SV module and confirmed by local inspection. In X7675, a balanced t(3;8) translocation generates the expected cross shaped pattern in interchromosomal contact maps and haplotype aware views show that the exchanged segments originate from opposite homologues. In X7676, a tandem duplication on chromosome 18 appears simultaneously as a focal copy number gain, a pair of SV breakpoints and a characteristic off diagonal block in the colocation matrix, again restricted to a single haplotype (Figure 4.33). Even in X7674, where the balanced t(19;22) falls in centromeric pericentromeric regions that yield only weak proximity signal, the large 19q duplication produces a clear block of increased contact density that reinforces the CNV call (Figure 4.32). Taken together, these examples show that, when high molecular weight DNA is available, the combination of depth, split reads, template based linkage and colocation matrices makes it possible to infer allele level configurations that would be extremely difficult to resolve from standard short read WGS alone.

In practical terms, an assay that remains compatible with existing Illumina infrastructure approaches the interpretive power that currently often requires a combination of long read sequencing, array based SV analysis and optical mapping, while still delivering conventional SNV and indel calls from the same dataset.

From an operational perspective, the current commercial implementation of Constellation mapped read technology, now marketed as Illumina TruPath Genome, should also be considered in terms of cost, throughput and ease of use [83], [84]. Standard Illumina WGS on the NovaSeq X Series remains the highest-throughput short-read option: Illumina reports a $200 USD genome on the 25B flow cell, although this estimate refers to sequencing consumables only and assumes 100 Gb per genome rather than a complete end-to-end clinical workflow [85]. For standard WGS throughput calculations, Illumina separately assumes >120 Gb of data per sample to achieve 30× genome coverage, and reports that NovaSeq X Plus can sequence up to 128 human genomes at 30× coverage per dual 25B flow-cell run [85], [86]. In contrast, TruPath Genome is positioned as a lower-throughput but higher-information whole-genome workflow, with a public list price of $395 USD per genome including consumables and DRAGEN Germline analysis at an industry-standard depth of at least 30× coverage, and with a maximum throughput of up to 16 genomes per run [83], [84]. Therefore, TruPath Genome should not be interpreted simply as a cheaper replacement for standard short-read WGS. Rather, it represents a trade-off in which lower maximum throughput and higher per-genome list price are balanced by workflow simplification, on-flow-cell library preparation, proximity-mapped-read information, ultra-long phasing and improved interpretability of complex structural variation [84].

6. Conclusions

Next generation sequencing has reshaped rare disease diagnostics by enabling systematic detection of SNVs and small indels. In many centers, analytical pipelines for these variant classes are now mature, standardized and highly reproducible. Yet a substantial fraction of patients remain undiagnosed, and converging evidence suggests that CNVs and other structural rearrangements account for a significant part of this missing yield [4]. The work presented in this thesis addresses this gap from two complementary angles. On the one hand, it pushes exome based CNV calling to its practical limits through careful benchmarking and pipeline engineering. On the other hand, it explores Illumina Constellation mapped read as a proximity aware whole genome assay that enriches short read sequencing with long range information [51].

In the exome domain, the starting point was the observation that CNV landscapes produced by standard tools on Twist exomes were often incompatible with biological expectations, either because of very sparse sets of multi megabase events or because of explosions of kilobase scale calls in selected libraries. Systematic analyses showed that a major driver of this behavior is GC dependent coverage distortion, which can mimic CNV signal and is differently absorbed by individual callers. By embedding ExomeDepth, ClinCNV and gCNV in a Snakemake workflow, running them on both the CNVPANEL01 reference panel with validated CNVs and on a real world clinical exome cohort, and explicitly experimenting with baseline composition, the thesis disentangles caller specific properties from dataset specific artefacts [52], [63], [64], [65], [67]. The results demonstrate that, when a sufficiently large WES cohort is available (at least 30-50 samples), ClinCNV offers the best compromise between sensitivity and overall call count, while gCNV is preferable in settings where maximizing sensitivity is more important than limiting the number of events. ExomeDepth, in contrast, already achieves high sensitivity with as few as 5-15 reference samples, provided they are homogeneous in coverage profile, although the resulting number of calls per sample remains relatively high. Taken together, these observations indicate that no single tool provides a universally stable solution, and that a consensus of callers, followed by structured

annotation and frequency based filtering, yields CNV call sets that are both accurate and more manageable in size for routine diagnostics.

At the same time, the exome work makes clear that certain limitations cannot be removed by better software alone. Targeted designs provide only a sparse view of the genome and often have uneven coverage across genes and exons. CNV detection relies on relative read depth against panels of normal samples, which must be constructed, maintained and periodically re optimized, especially in laboratories where capture protocols and sequencing conditions evolve over time. Many pathogenic CNVs extend beyond coding regions or involve complex breakpoints that are only partially visible in exome data. Even with an optimized multi caller pipeline and a carefully curated baseline, some validated CNVs remain difficult to recover, and specificity can only be improved at the cost of losing a fraction of true events.

The second part of the thesis turns to Constellation mapped read, which extend short read WGS with a proximity signal derived from the physical clustering of reads originating from the same long DNA molecules. By comparing Constellation mapped read with matched TruSeq PCR free WGS, the work shows that SNVs and indels calling performance is essentially preserved, while phasing and SVs detection benefit substantially from the added information. Constellation mapped read reconstructs long templates and haplotype blocks that span tens of megabases and supports the detection of CNVs and SVs with concordant depth and junction evidence. In a set of clinically characterized genomes, this translates into concrete interpretive gains. Complex duplications at SCN1A, balanced translocations and focal deletions can be resolved at the level of individual alleles by integrating colocation matrices, haplotagged reads and conventional coverage views, within a single Illumina based workflow.

These findings highlight both the promise and the current constraints of proximity aware short read sequencing. On the positive side, Constellation mapped read suggests that it is possible to approach the structural resolution of long read or optical mapping technologies while retaining the throughput, cost structure and infrastructure compatibility of existing short read platforms. On the limiting side,

the quality of the long range signal is tightly coupled to DNA extraction. The relatively small number of cases and the heterogeneity of extraction protocols in this study call for larger evaluations.

Taken together, the exome and Constellation mapped read results can be viewed as steps along a continuum that leads from today's practice, in which diagnoses are dominated by SNVs and indels, toward integrated genome first diagnostics where structural variation is analyzed and interpreted as a first class component of every case. In the exome setting, the thesis provides a concrete blueprint for how to stabilize CNV calling through multi caller consensus and annotation driven prioritization. In the whole genome setting, it offers early evidence that proximity informed sequencing can deliver long range phasing and structural insight without abandoning short read technology. The overarching conclusion is that closing the diagnostic gap will not come from a single breakthrough, but from the careful alignment of wet lab quality, reference data, and computational pipelines. Within this evolving landscape, the approaches developed here show that it is already possible to extend short read sequencing toward richer SV detection, and they outline a realistic path toward future assays in which SNVs, indels, CNVs, SVs and phasing are routinely considered together in a unified clinical workflow.

7. Limitations and future perspectives

Several limitations of this work need to be acknowledged. The exome CNV benchmarks rely on CNVPANEL01, which provides a valuable but finite set of validated CNVs that not cover the full spectrum of structural variation encountered in clinical practice, especially for very small or very complex events. The clinical exome cohort from Burlo adds realism but lacks a gold standard, so conclusions about enrichment and performance are necessarily indirect.

The Constellation study is based on a limited number of cases and on heterogeneous DNA extraction methods, which makes it difficult to generalize the results across all possible clinical scenarios.

Despite these constraints, the work supports the following narrative: standard NGS has already transformed clinical genetics through SNV and small indel calling; the question is no longer whether these variants can be detected, but how to integrate them with SVs in a way that increases diagnostic yield without overwhelming laboratories with complexity. Exome based CNV calling can provide valuable information, especially when driven by carefully curated reference panels and consensus across multiple tools, but it may remain sensitive to capture design and library specific artefacts. Constellation mapped read points toward a future in which short read sequencing is enriched with long range information, permitting joint analysis of SNVs, CNVs, SVs and phasing in a single, integrated assay.

In this perspective, the work presented in the thesis can be seen as a set of methodological steps along a continuum. At one end, exome CNV pipelines are pushed to their practical limits and equipped with consensus and annotation strategies that make them more suitable for clinical laboratories. At the other end, Constellation mapped read demonstrates that it is possible to preserve the strengths of PCR free WGS while adding phasing and improved structural resolution, provided that DNA quality and library preparation are carefully controlled.

Future developments will likely involve both methodological and organizational advances. Methodologically, there is room to extend the exome pipeline to additional CNV callers, to integrate machine learning for prioritization of structural events, and to leverage long read and optical data as higher order gold standards.

Ultimately, the goal is to move from an NGS practice where most diagnoses are based on SNVs and small indels, to an integrated genome first diagnostics where SV is jointly analyzed and interpreted as a first class citizen. The results of this thesis suggest that this transition is feasible, but that it requires careful engineering of pipelines, thoughtful use of reference data, and, in the case of emerging assays such as Constellation mapped read, sustained attention to the quality of the starting material.

References

[1] D. R. Adams and C. M. Eng, “Next-Generation Sequencing to Diagnose Suspected Genetic Disorders,” *New England Journal of Medicine*, vol. 379, no. 14, pp. 1353–1362, Oct. 2018, doi: 10.1056/NEJMra1711801.

[2] C. F. Wright, D. R. FitzPatrick, and H. V. Firth, “Paediatric genomics: diagnosing rare disease in children,” *Nat. Rev. Genet.*, vol. 19, no. 5, pp. 253–268, May 2018, doi: 10.1038/nrg.2017.116.

[3] A. Auton *et al.*, “A global reference for human genetic variation,” *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015, doi: 10.1038/nature15393.

[4] L. J. Ewans *et al.*, “Whole exome and genome sequencing in mendelian disorders: a diagnostic and health economic analysis,” *European Journal of Human Genetics*, vol. 30, no. 10, pp. 1121–1131, Oct. 2022, doi: 10.1038/s41431-022-01162-2.

[5] C. C. Y. Chung, S. P. Y. Hue, N. Y. T. Ng, P. H. L. Doong, A. T. W. Chu, and B. H. Y. Chung, “Meta-analysis of the diagnostic and clinical utility of exome and genome sequencing in pediatric and adult patients with rare diseases across diverse populations,” *Genetics in Medicine*, vol. 25, no. 9, p. 100896, Sep. 2023, doi: 10.1016/j.gim.2023.100896.

[6] C. F. Wright, D. R. FitzPatrick, and H. V. Firth, “Paediatric genomics: diagnosing rare disease in children,” *Nat. Rev. Genet.*, vol. 19, no. 5, pp. 253–268, May 2018, doi: 10.1038/nrg.2017.116.

[7] S. Behera *et al.*, “Comprehensive genome analysis and variant detection at scale using DRAGEN,” *Nat. Biotechnol.*, vol. 43, no. 7, pp. 1177–1191, Jul. 2025, doi: 10.1038/s41587-024-02382-1.

[8] G. A. Van der Auwera *et al.*, “From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline,” *Curr. Protoc. Bioinformatics*, vol. 43, no. 1, Oct. 2013, doi: 10.1002/0471250953.bi1110s43.

[9] J. M. Zook *et al.*, “An open resource for accurately benchmarking small variant and reference calls,” *Nat. Biotechnol.*, vol. 37, no. 5, pp. 561–566, May 2019, doi: 10.1038/s41587-019-0074-6.

[10] “SNVs/INDELs picture,” https://www.xcode.life/genetics/what-is-the-ultimate-source-of-genetic-variation/.

[11] M. A. Schaub, A. P. Boyle, A. Kundaje, S. Batzoglou, and M. Snyder, “Linking disease associations with regulatory information in the human genome,” *Genome Res.*, vol. 22, no. 9, pp. 1748–1759, Sep. 2012, doi: 10.1101/gr.136127.111.

[12] M. T. Maurano *et al.*, “Systematic Localization of Common Disease-Associated Variation in Regulatory DNA,” *Science (1979).*, vol. 337, no. 6099, pp. 1190–1195, Sep. 2012, doi: 10.1126/science.1222794.

[13] F. Zhang and J. R. Lupski, “Non-coding genetic variants in human disease: Figure 1.,” *Hum. Mol. Genet.*, vol. 24, no. R1, pp. R102–R110, Oct. 2015, doi: 10.1093/hmg/ddv259.

[14] I. A. Adzhubei *et al.*, “A method and server for predicting damaging missense mutations,” *Nat. Methods*, vol. 7, no. 4, pp. 248–249, Apr. 2010, doi: 10.1038/nmeth0410-248.
[15] M. Kircher, D. M. Witten, P. Jain, B. J. O’Roak, G. M. Cooper, and J. Shendure, “A general framework for estimating the relative pathogenicity of human genetic variants,” *Nat. Genet.*, vol. 46, no. 3, pp. 310–315, Mar. 2014, doi: 10.1038/ng.2892.
[16] S. Richards *et al.*, “Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology,” *Genetics in Medicine*, vol. 17, no. 5, pp. 405–424, May 2015, doi: 10.1038/gim.2015.30.
[17] I. A. E. M. van Belzen, A. Schönhuth, P. Kemmeren, and J. Y. Hehir-Kwa, “Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology,” *NPJ Precis. Oncol.*, vol. 5, no. 1, p. 15, Mar. 2021, doi: 10.1038/s41698-021-00155-6.
[18] L. Feuk, A. R. Carson, and S. W. Scherer, “Structural variation in the human genome,” *Nat. Rev. Genet.*, vol. 7, no. 2, pp. 85–97, Feb. 2006, doi: 10.1038/nrg1767.
[19] M. Kato *et al.*, “Population-genetic nature of copy number variations in the human genome,” *Hum. Mol. Genet.*, vol. 19, no. 5, pp. 761–773, Mar. 2010, doi: 10.1093/hmg/ddp541.
[20] P. H. Sudmant *et al.*, “Global diversity, population stratification, and selection of human copy-number variation,” *Science (1979).*, vol. 349, no. 6253, Sep. 2015, doi: 10.1126/science.aab3761.
[21] K. Inoue and J. R. Lupski, “M <scp>OLECULAR</scp> M <scp>ECHANISMS FOR</scp> G <scp>ENOMIC</scp> D <scp>ISORDERS</scp>,” *Annu. Rev. Genomics Hum. Genet.*, vol. 3, no. 1, pp. 199–242, Sep. 2002, doi: 10.1146/annurev.genom.3.032802.120023.
[22] W. Gu, F. Zhang, and J. R. Lupski, “Mechanisms for human genomic rearrangements,” *Pathogenetics*, vol. 1, no. 1, p. 4, Dec. 2008, doi: 10.1186/1755-8417-1-4.
[23] F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski, “Copy Number Variation in Human Health, Disease, and Evolution,” *Annu. Rev. Genomics Hum. Genet.*, vol. 10, no. 1, pp. 451–481, Sep. 2009, doi: 10.1146/annurev.genom.9.081307.164217.
[24] P. Stankiewicz and J. R. Lupski, “Structural Variation in the Human Genome and its Role in Disease,” *Annu. Rev. Med.*, vol. 61, no. 1, pp. 437–455, Feb. 2010, doi: 10.1146/annurev-med-100708-204735.
[25] C. M. B. Carvalho and J. R. Lupski, “Mechanisms underlying structural variant formation in genomic disorders,” *Nat. Rev. Genet.*, vol. 17, no. 4, pp. 224–238, Apr. 2016, doi: 10.1038/nrg.2015.25.
[26] M. Mahmoud, N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz, and F. J. Sedlazeck, “Structural variant calling: the long and the short of it,”

Genome Biol., vol. 20, no. 1, p. 246, Dec. 2019, doi: 10.1186/s13059-019-1828-7.

[27] M. Gabrielaite *et al.*, "A Comparison of Tools for Copy-Number Variation Detection in Germline Whole Exome and Whole Genome Sequencing Data," *Cancers (Basel).*, vol. 13, no. 24, p. 6283, Dec. 2021, doi: 10.3390/cancers13246283.

[28] S. Kosugi, Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani, "Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing," *Genome Biol.*, vol. 20, no. 1, p. 117, Dec. 2019, doi: 10.1186/s13059-019-1720-5.

[29] Z. Liu, Z. Xie, and M. Li, "Comprehensive and deep evaluation of structural variation detection pipelines with third-generation sequencing data," *Genome Biol.*, vol. 25, no. 1, p. 188, Jul. 2024, doi: 10.1186/s13059-024-03324-5.

[30] I. E. Sønderby *et al.*, "Effects of copy number variations on brain structure and risk for psychiatric illness: Large-scale studies from the <scp>ENIGMA</scp> working groups on <scp>CNVs</scp>," *Hum. Brain Mapp.*, vol. 43, no. 1, pp. 300–328, Jan. 2022, doi: 10.1002/hbm.25354.

[31] R. Redon *et al.*, "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, pp. 444–454, Nov. 2006, doi: 10.1038/nature05329.

[32] S. Pande, M. Dawood, and C. M. Grochowski, "Structural Variants: Mechanisms, Mapping, and Interpretation in Human Genetics," *Genes (Basel).*, vol. 16, no. 8, p. 905, Jul. 2025, doi: 10.3390/genes16080905.

[33] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer, "The Database of Genomic Variants: a curated collection of structural variation in the human genome," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D986–D992, Jan. 2014, doi: 10.1093/nar/gkt958.

[34] H. V. Firth *et al.*, "DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources," *The American Journal of Human Genetics*, vol. 84, no. 4, pp. 524–533, Apr. 2009, doi: 10.1016/j.ajhg.2009.03.010.

[35] H. L. Rehm *et al.*, "ClinGen — The Clinical Genome Resource," *New England Journal of Medicine*, vol. 372, no. 23, pp. 2235–2242, Jun. 2015, doi: 10.1056/NEJMsr1406261.

[36] D. T. Miller *et al.*, "Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies," *The American Journal of Human Genetics*, vol. 86, no. 5, pp. 749–764, May 2010, doi: 10.1016/j.ajhg.2010.04.006.

[37] M. Wayhelova *et al.*, "The clinical benefit of array-based comparative genomic hybridization for detection of copy number variants in Czech children with intellectual disability and developmental delay," *BMC Med. Genomics*, vol. 12, no. 1, p. 111, Dec. 2019, doi: 10.1186/s12920-019-0559-7.

[38] J. Wiszniewska *et al.*, “Combined array CGH plus SNP genome analyses in a single assay for optimized clinical testing,” *European Journal of Human Genetics*, vol. 22, no. 1, pp. 79–87, Jan. 2014, doi: 10.1038/ejhg.2013.77.
[39] R. L. Collins *et al.*, “A structural variation reference for medical and population genetics,” *Nature*, vol. 581, no. 7809, pp. 444–451, May 2020, doi: 10.1038/s41586-020-2287-8.
[40] L. Kadalayil *et al.*, “Exome sequence read depth methods for identifying copy number changes,” *Brief. Bioinform.*, vol. 16, no. 3, pp. 380–392, May 2015, doi: 10.1093/bib/bbu027.
[41] R. Tan *et al.*, “An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data,” *Hum. Mutat.*, vol. 35, no. 7, pp. 899–907, Jul. 2014, doi: 10.1002/humu.22537.
[42] F. Zare, M. Dow, N. Monteleone, A. Hosny, and S. Nabavi, “An evaluation of copy number variation detection tools for cancer using whole exome sequencing data,” *BMC Bioinformatics*, vol. 18, no. 1, p. 286, Dec. 2017, doi: 10.1186/s12859-017-1705-x.
[43] V. Gordeeva, E. Sharova, K. Babalyan, R. Sultanov, V. M. Govorun, and G. Arapidi, “Benchmarking germline CNV calling tools from exome sequencing data,” *Sci. Rep.*, vol. 11, no. 1, p. 14416, Jul. 2021, doi: 10.1038/s41598-021-93878-2.
[44] S. Miyatake *et al.*, “Detecting copy-number variations in whole-exome sequencing data using the eXome Hidden Markov Model: an ‘exome-first’ approach,” *J. Hum. Genet.*, vol. 60, no. 4, pp. 175–182, Apr. 2015, doi: 10.1038/jhg.2014.124.
[45] G. A. Logsdon, M. R. Vollger, and E. E. Eichler, “Long-read human genome sequencing and its applications,” *Nat. Rev. Genet.*, vol. 21, no. 10, pp. 597–614, Oct. 2020, doi: 10.1038/s41576-020-0236-x.
[46] P. E. Warburton and R. P. Sebra, “Long-Read DNA Sequencing: Recent Advances and Remaining Challenges,” *Annu. Rev. Genomics Hum. Genet.*, vol. 24, no. 1, pp. 109–132, Aug. 2023, doi: 10.1146/annurev-genom-101722-103045.
[47] S. Sinha *et al.*, “Long read sequencing enhances pathogenic and novel variation discovery in patients with rare diseases,” *Nat. Commun.*, vol. 16, no. 1, p. 2500, Mar. 2025, doi: 10.1038/s41467-025-57695-9.
[48] B. Levy, R. D. Burnside, and Y. Akkari, “Optical Genome Mapping: A New Tool for Cytogenomic Analysis,” *Genes (Basel).*, vol. 16, no. 8, p. 924, Jul. 2025, doi: 10.3390/genes16080924.
[49] R. Elyanow, H.-T. Wu, and B. J. Raphael, “Identifying structural variants using linked-read sequencing data,” *Bioinformatics*, vol. 34, no. 2, pp. 353–360, Jan. 2018, doi: 10.1093/bioinformatics/btx712.
[50] L. Harewood *et al.*, “Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours,” *Genome Biol.*, vol. 18, no. 1, p. 125, Dec. 2017, doi: 10.1186/s13059-017-1253-8.

[51] “Constellation Mapped Read Technology,” https://acpa-achropuce.com/wp-content/uploads/2025/05/Introducing-constellation-mapped-read-technology.pdf.
[52] “CNVPANEL01 from Coriell Institute,” https://www.coriell.org/0/Sections/Search/Panel_Detail.aspx?Ref=CNVPANEL01&Product=HDP.
[53] “Twist Bioscience,” https://www.twistbioscience.com/.
[54] “bcl2fastq,” https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html.
[55] S. Chen, Y. Zhou, Y. Chen, and J. Gu, “fastp: an ultra-fast all-in-one FASTQ preprocessor,” *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, Sep. 2018, doi: 10.1093/bioinformatics/bty560.
[56] Md. Vasimuddin, S. Misra, H. Li, and S. Aluru, “Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems,” in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, IEEE, May 2019, pp. 314–324. doi: 10.1109/IPDPS.2019.00041.
[57] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.
[58] “Picard,” https://broadinstitute.github.io/picard/.
[59] G. A. Van der Auwera *et al.*, “From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline,” *Curr. Protoc. Bioinformatics*, vol. 43, no. 1, Oct. 2013, doi: 10.1002/0471250953.bi1110s43.
[60] G. Jun, M. K. Wing, G. R. Abecasis, and H. M. Kang, “An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data,” *Genome Res.*, vol. 25, no. 6, pp. 918–925, Jun. 2015, doi: 10.1101/gr.176552.114.
[61] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010, doi: 10.1093/bioinformatics/btq033.
[62] R. D’Aurizio, T. Pippucci, L. Tattini, B. Giusti, M. Pellegrini, and A. Magi, “Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2,” *Nucleic Acids Res.*, p. gkw695, Aug. 2016, doi: 10.1093/nar/gkw695.
[63] V. Plagnol *et al.*, “A robust model for read count data in exome sequencing experiments and implications for copy number variant calling,” *Bioinformatics*, vol. 28, no. 21, pp. 2747–2754, Nov. 2012, doi: 10.1093/bioinformatics/bts526.
[64] G. Demidov, M. Sturm, and S. Ossowski, “ClinCNV: multi-sample germline CNV detection in NGS data,” Jun. 13, 2022. doi: 10.1101/2022.06.10.495642.

[65] M. Babadi *et al.*, “GATK-gCNV: A Rare Copy Number Variant Discovery Algorithm and Its Application to Exome Sequencing in the UK Biobank,” Aug. 26, 2022. doi: 10.1101/2022.08.25.504851.
[66] J. Köster and S. Rahmann, “Snakemake—a scalable bioinformatics workflow engine,” *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, Oct. 2012, doi: 10.1093/bioinformatics/bts480.
[67] F. Mölder *et al.*, “Sustainable data analysis with Snakemake.,” *F1000Res.*, vol. 10, p. 33, 2021, doi: 10.12688/f1000research.29032.3.
[68] K. Warton, L.-J. Graham, N. Yuwono, and G. Samimi, “Comparison of 4 commercial kits for the extraction of circulating DNA from plasma,” *Cancer Genet.*, vol. 228–229, pp. 143–150, Dec. 2018, doi: 10.1016/j.cancergen.2018.02.004.
[69] H. A. Dahn *et al.*, “Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing,” *Gigascience*, vol. 11, Aug. 2022, doi: 10.1093/gigascience/giac068.
[70] L. Caceres, “Qubit dsDNA HS Assay v2,” Sep. 26, 2023. doi: 10.17504/protocols.io.kxygx3zrwg8j/v2.
[71] “Agilent Genomic DNA TapeStation,” chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.agilent.com/cs/library/applications/5991-1797EN.pdf?srsltid=AfmBOoqn3eT2d_wYCrDROQXGyJnt8oOrmTwAuB0VGTsYZ13kfWyjCGMy.
[72] “Basespace,” https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub.html?
[73] “Illumina Connected Analytics (ICA),” https://developer.illumina.com/news-updates/illumina-connected-analytics-streamlined-sequencer-integration-and-automation-of-data-processing-workflows?
[74] “Constellation Mapped Read manual,” https://illumina.gitbook.io/constellation/UDy3oYQowTUmKldPiq34.
[75] M. Martin *et al.*, “WhatsHap: fast and accurate read-based phasing,” Nov. 02, 2016. doi: 10.1101/085050.
[76] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.
[77] J. T. Robinson *et al.*, “Integrative genomics viewer,” *Nat. Biotechnol.*, vol. 29, no. 1, pp. 24–26, Jan. 2011, doi: 10.1038/nbt.1754.
[78] P. Kerpedjiev *et al.*, “HiGlass: web-based visual exploration and analysis of genome interaction maps,” *Genome Biol.*, vol. 19, no. 1, p. 125, Dec. 2018, doi: 10.1186/s13059-018-1486-1.
[79] V. Plagnol *et al.*, “A robust model for read count data in exome sequencing experiments and implications for copy number variant calling,” *Bioinformatics*, vol. 28, no. 21, pp. 2747–2754, Nov. 2012, doi: 10.1093/bioinformatics/bts526.

[80] H. Kim *et al.*, “Copy-number analysis by base-level normalization: An intuitive visualization tool for evaluating copy number variations,” *Clin. Genet.*, vol. 103, no. 1, pp. 35–44, Jan. 2023, doi: 10.1111/cge.14236.
[81] F.-N. Tilemis *et al.*, “Germline CNV Detection through Whole-Exome Sequencing (WES) Data Analysis Enhances Resolution of Rare Genetic Diseases,” *Genes (Basel).*, vol. 14, no. 7, p. 1490, Jul. 2023, doi: 10.3390/genes14071490.
[82] V. Geoffroy *et al.*, “The AnnotSV webserver in 2023: updated visualization and ranking,” *Nucleic Acids Res.*, vol. 51, no. W1, pp. W39–W45, Jul. 2023, doi: 10.1093/nar/gkad426.
[83] “Illumina launches TruPath Genome, setting a new standard in genomic insight.”
[84] “TruPath Genome Solution. Product page.”
[85] “NovaSeq X Series Enables Broader, Deeper Sequencing.”
[86] “NovaSeq X Series specifications.”

Supplementary

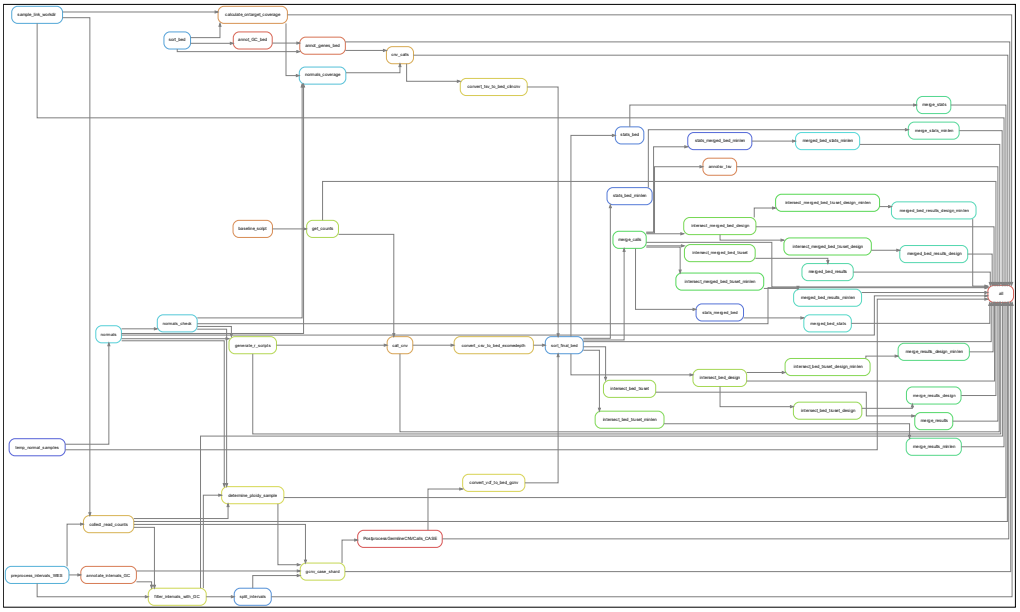

Figure S. 1: Overview of the Snakemake workflow used in this study.

						CNV count mean/median		
			CNVPANEL01			**CNVPANEL01+Burlo**		
	No selection	**No selection**	**5 nn**	**10 nn**	**15 nn**	**20 nn**	**25 nn**	**30 nn**
ExomeDepth	340/335	688/691	349/343	343/341	342/339	340/338	341/343	340/337
clinCNV	79/70	62/55	53/53	65/61	68/65	71/79	74/68	76/69

gCNV	202/190	162/148	450/450	367/354	305/293	273/265	252/241	236/224

Table S. 1: CNV counts by selecting X nearest neighbors (nn) as normals.

	TP covered > 75% of bases / TP total							
	CNVPANEL01	**CNVPANEL01+Burlo**						
	No selection	**No selection**	**5 nn**	**10 nn**	**15 nn**	**20 nn**	**25 nn**	**30 nn**
ExomeDepth	44/55	25/55	45/55	45/55	45/55	44/55	44/55	44/55
clinCNV	43/55	44/55	39/55	40/55	39/55	39/55	41/55	41/55
gCNV	43/55	43/55	39/55	41/55	42/55	43/55	43/55	43/55

Table S. 2: True positives identified by selecting X nearest neighbors (nn) as normals.

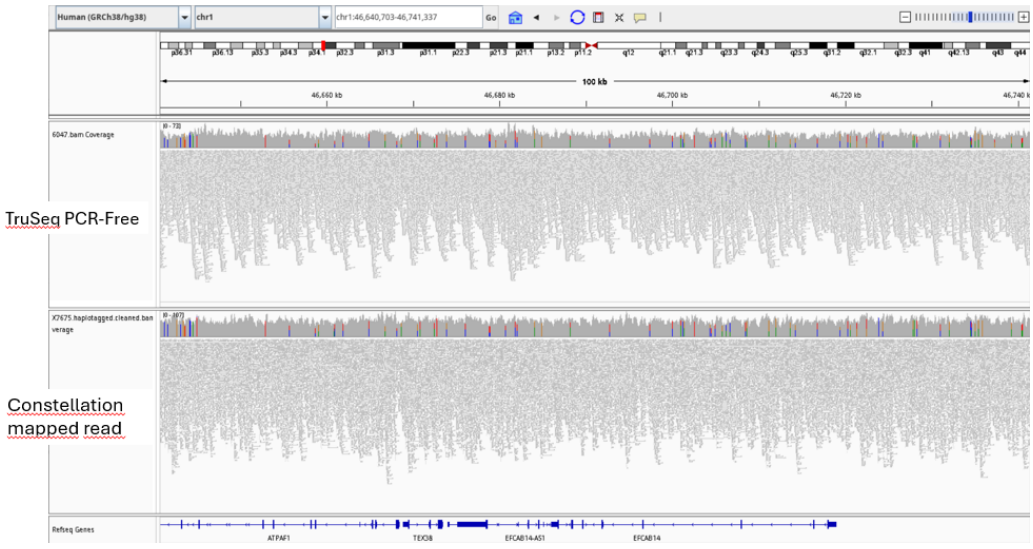

Figure S. 2: IGV (Integrated Genome Browser) screenshot of a 100kb window displaying the alignments of X7675 sample sequenced with Illumina TruSeq PCR-free (top) and Constellation mapped read (bottom).