

REVIEW

Open Access



Computational strategies for copy number variation detection, disease association, and beyond

Amir Hossein Saeidian^{1,2†}, Hani Sabaie^{3†}, Mahdi Akbarzadeh^{4†}, Hassan Vahidnezhad^{2,5,6,7}, Leila Youssefian^{2,8}, Toktam Saadattalab⁹, Haowei Du¹, Xi Luo¹, Nichole Marie Owen¹, Xiaonan Zhao¹, Joseph Glessner^{2,5,6*†} and Hakon Hakonarson^{2,5,6,10,11*†}

[†]Amir Hossein Saeidian, Hani Sabaie, and Mahdi Akbarzadeh contributed equally to this study as the first authors.

[†]Hakon Hakonarson and Joseph Glessner contributed equally to this study as corresponding authors.

*Correspondence:
glessner@chop.edu;
hakonarson@chop.edu

² Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, USA
Full list of author information is available at the end of the article

Abstract

Copy number variations (CNVs) are key structural variations that contribute to human genetic diversity, evolution, and disease susceptibility. Advances in sequencing technologies and computational methods have improved CNV detection, yet association studies remain challenged by methodological limitations and a lack of standardisation. This review provides an overview of computational strategies for germline CNV detection and disease association. We highlight the value of CNV analysis for uncovering genetic contributions to complex traits and disease risk and outline an analysis workflow including key benchmarking methods. We also discuss current challenges and future directions for advancing CNV detection and association analysis.

Introduction

Structural variations (SVs) are crucial human genetic variations, typically defined by alterations involving 50 or more DNA nucleotides [1]. These SVs include a broad spectrum of genomic changes, including copy number variations (CNVs), insertions (≥ 50 bp), including mobile element insertions (MEIs), inversions, and chromosomal abnormalities [2, 3] (Fig. 1A). Investigating SVs is crucial for understanding the diversity of human genetics, their role in evolution, and their functional effects (reviewed in Collins and Talkowski [2]).

CNVs, a class of SVs characterised by changes in the number of copies of a particular segment of DNA, are among the most frequently examined and prevalent SVs. Previous studies have shown that insertions and CNVs collectively account for the majority of SVs in the human genome, often exceeding 90% of all detected SVs [2]. However, long-read sequencing (LR-seq) studies have demonstrated that the relative distribution of SV classes can vary across datasets, with insertions and deletions often identified



This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

at comparable frequencies. The observed spectrum of SVs is therefore influenced by sequencing technology, analytical methods, and genomic context [5]. CNVs involve deletions or duplications within the genome, typically defined as spanning 50 base pairs or more and often exceeding 1 kb. These variations are fairly prevalent, with approximately 10,000 occurrences per genome, and approximately 800 of these are larger than 10 kilobases [2]. These structural alterations account for a significant portion of genetic diversity, encompassing an estimated 5–10% of the human genome [6, 7].

CNVs at a locus can arise as recurrent or non-recurrent events. Recurrent CNVs are typically mediated by genomic features, such as low-copy repeats, resulting in similar breakpoints across individuals, whereas non-recurrent CNVs exhibit variable breakpoints and structures (Fig. 1B). Beyond structural features, CNVs exert diverse functional effects that are important for downstream interpretation but are not directly part of computational detection workflows. As summarised in Fig. 1C, a primary mechanism is gene dosage alteration, where deletions may result in haploinsufficiency and duplications in triplosensitivity, particularly in dosage-sensitive genes [2, 8]. CNVs can also disrupt gene integrity and alter regulatory landscapes, for example, by perturbing topologically associating domains (TADs) and modifying enhancer-gene interactions [3, 8, 9]. In addition, CNVs can have substantial effects on gene expression and may act as expression quantitative trait loci (eQTLs) [10, 11], or contribute to the formation of gene fusions in the context of complex rearrangements [12]. Deletions may also unmask recessive variants on the homologous chromosome, thereby revealing pathogenic alleles that would otherwise remain silent [8].

Genomic disorders (GDs) arising from CNVs or genomic rearrangements provide definitive evidence of CNV pathogenicity [13]. Many well-characterised GDs,

(See figure on next page.)

Fig. 1 Structural and functional features of copy number variations (CNVs). **A** Structural variations (SVs) are genomic alterations, including deletions (DEL), duplications (DUP), insertions (INS; ≥ 50 bp), including mobile element insertions (MEIs), inversions (INV), complex rearrangements (CPX), and chromosomal abnormalities (CA), including reciprocal translocations and aneuploidy. These include balanced (BA) and unbalanced (UB) types. BA SVs maintain DNA dosage but may disrupt gene structure, whereas UB SVs cause DNA gain or loss. CNVs are SVs that are typically divided into DEL, which reduces the copy number, and DUP, which increases it. Moreover, complex CNV architectures like DUP-triplication (TRP)/INV-DUP configurations can occur. **B** Recurrent and non-recurrent CNV architectures and breakpoint patterns. The reference genome (REF) is divided into segments, with dosage-sensitive genes in critical regions. Low-copy repeats (LCRs) mediate non-allelic homologous recombination (NAHR), causing recurrent CNVs. Tandem paralogous repeats (TPR) can also mediate these rearrangements. CNVs are shown as coloured rectangles (red for DEL, blue for DUP, and orange for inverted TRP structures). Segment 1 contains the critical region gene flanked by LCRs, predisposing it to recurrent CNVs of the same size and content across individuals. Segment 2 shows that recurrent rearrangements are possible within TPRs and alter the copy number of a dosage-sensitive gene located inside the repeat. Segment 3 illustrates non-recurrent DUP-TRP/INV-DUP structures. Segment 4 represents non-recurrent rearrangements that display breakpoint grouping within paralogous repeats, in contrast to the tight clustering seen in recurrent cases. Segment 5 demonstrates non-recurrent rearrangements without any clustering or grouping of breakpoints, resulting in entirely unique CNVs across individuals. The smallest region of overlap (SRO) refers to the minimal genomic segment that is consistently altered among different individuals with overlapping CNVs. Panel B was inspired by the schematic representation of recurrent and non-recurrent CNVs described in [4]. **C** Functional consequences of CNVs. CNVs can affect gene and protein function through gene dosage alterations, disruption of gene structure, changes in regulatory architecture (including topologically associating domains), effects on gene expression, and formation of gene fusions or unmasking of recessive variants. These functional consequences are essential for interpreting the phenotypic impact of CNVs. Created with BioRender.com

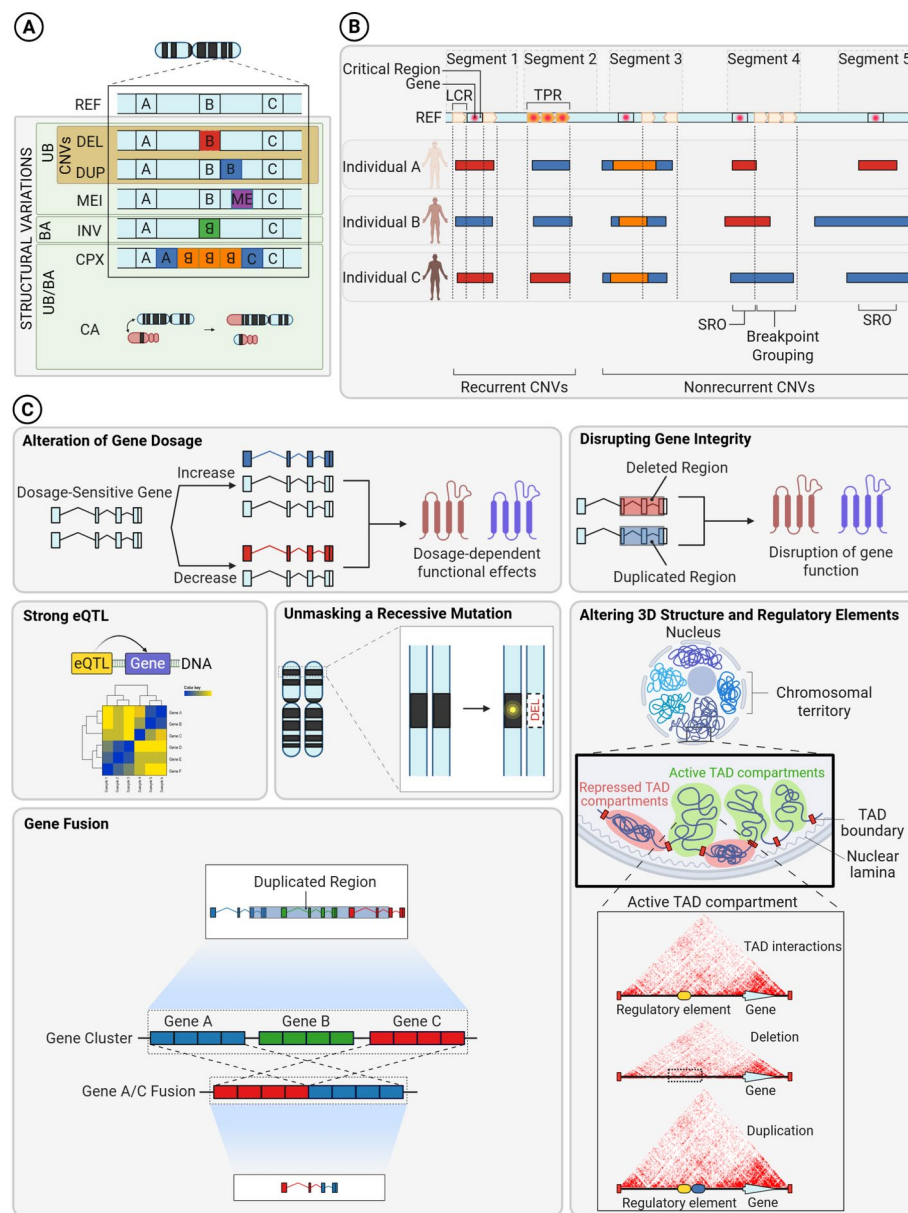


Fig. 1 (See legend on previous page.)

particularly those mediated by recurrent large CNVs, involve megabase-sized regions encompassing multiple genes [14]. While some of these disorders have a single dominant genetic driver, others follow oligogenic and polygenic models, underscoring the complexity of the genetic architecture and the influence of modifier effects [15–17]. Several recurrent GD CNVs found in the general population can influence physiological traits, such as height, even in the absence of overt clinical disease [18–20]. Beyond classical GDs, CNVs are associated with a broad spectrum of neurodevelopmental and psychiatric conditions, including schizophrenia, autism, and bipolar disorder [14, 21, 22] (Fig. 2). Taken together, CNVs are widely recognised as important contributors to disease susceptibility, and in some cases, they can exhibit large effect

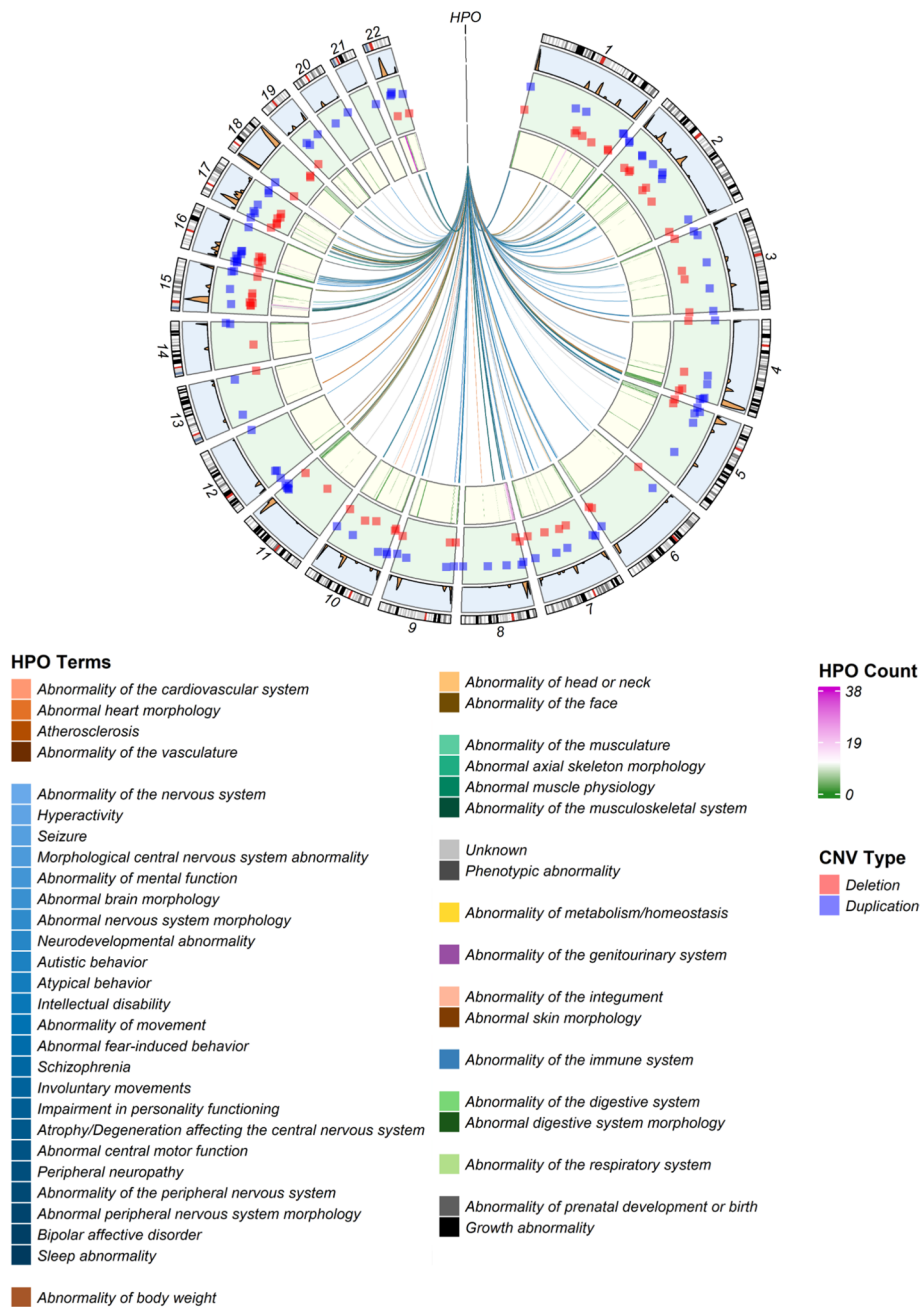


Fig. 2 Genome-wide distribution of disease-associated copy number variations (CNVs) and their associated phenotypes. To illustrate the broad impact of CNVs, we visualised data from Collins et al. [14], which showed that large rare CNVs are significantly associated with multiple human diseases and traits across the genome. The study performed a large-scale meta-analysis across multiple cohorts, identifying 163 rare, large CNV loci significantly associated with diverse human diseases and traits, particularly neurodevelopmental and psychiatric disorders. Circos plot displaying chromosomes 1–22 with outer ideograms. The second track shows the density plots of CNV counts along each chromosome. The third track depicts individual CNV loci, with red squares representing deletions and blue squares representing duplication. The fourth track is a heatmap indicating the number of Human Phenotype Ontology (HPO) terms mapped to each locus (low to high intensity). The innermost links connect genomic loci to HPO categories, which are colour-coded as indicated in the figure legend. The figure was generated using the circlize R package [23]

sizes approaching those observed in Mendelian disorders. For example, variation at the 22q11.2 locus illustrates this impact, with deletions associated with an approximately 60-fold increase in schizophrenia risk, whereas duplications may reduce risk to approximately 15% of baseline levels [8, 21]. At the same time, many CNVs are incompletely penetrant and show variable expressivity, reflecting the broader complexity of their genetic architecture [2].

Large-scale biobanks are essential for advancing this field. By linking genome-wide CNV discovery with large-scale and broadly characterised phenotypic data, resources such as the UK Biobank enable systematic CNV association analysis across tens of thousands of individuals. These analyses refine penetrance estimates, identify pleiotropic effects across traits, and reveal modifier influences from local and genome-wide genetic backgrounds [8, 18, 24–27]. Advances in sequencing technologies and computational methods have enabled higher-resolution and more sensitive CNV detection at a larger scale, including major contributions from large-scale efforts such as the Human Genome Structural Variation Consortium, which has leveraged LR-seq and haplotype-resolved assemblies to improve the discovery and characterisation of SV [28–32]. However, CNV association studies face several gaps in best practices, with challenges ranging from methodological issues to a lack of standardisation and a robust data infrastructure. These limitations hinder accurate and reproducible CNV detection and association analysis. A primary obstacle is the variability and low consensus in CNV calls across different algorithms and platforms [33–36]. Tool performance is highly context-dependent, with strengths and limitations varying across sequencing technologies, data types, and analytical strategies. As a result, no single algorithm consistently outperforms others across all platforms and use cases [33, 34, 37–41]. With numerous tools available, each yielding different outcomes, selecting the appropriate tool is overwhelming. Most current tool comparisons rely on benchmarks provided by tool developers, potentially introducing bias. Without independent benchmarking, researchers and clinicians risk making sub-optimal choices that may compromise the analysis and clinical interpretation [33, 37–40, 42–45]. Furthermore, unlike single nucleotide polymorphism (SNP)-based association analysis, which has standardised tools such as PLINK [46, 47] and REGENIE [48], there is currently no universally adopted or community-standard tool for CNV-based association analysis [8, 49], although ongoing community efforts are actively developing and benchmarking methods in this field. Existing tools often lack user-friendliness, essential reporting features, or fail to integrate necessary analytical steps [49–52].

In this review, we focus on germline CNVs, as their detection, interpretation, and association analyses rely on distinct study designs and computational frameworks compared to somatic CNVs, which are typically investigated in cancer genomics and mosaic contexts. While somatic CNVs have well-established roles in both cancer and non-cancer conditions [40, 53–65], their analysis involves fundamentally different methodological considerations and is therefore beyond the scope of this review. Accordingly, we concentrated on computational approaches for the detection and association analysis of germline CNVs.

We first highlight the value of CNV association analysis in revealing genetic contributions to complex traits that are often missed by SNP-based analysis. We then provide a general workflow for CNV association analysis. We then present the platforms and

methods for CNV detection and outline the benchmarking frameworks and challenges of comparing CNV detection tools. Next, we review CNV calling and association analysis tools, focusing on independent benchmarking studies for unbiased performance evaluation. In the absence of independent evidence, we supplemented the discussion with a critical overview of community-recommended tools to provide practical, evidence-based, and context-specific recommendations for researchers and clinicians seeking to select appropriate CNV detection and association approaches. Finally, we conclude by addressing the current challenges and future perspectives of CNV detection and association analyses.

Unique contributions of copy number variation association studies

CNV association analyses provide insights that go beyond the capabilities of SNP-based methodologies, offering unique opportunities and expanding our understanding of the genetic architecture of complex traits [8, 20, 26, 27, 66]. In some cases, traits such as birth weight, total cholesterol, low-density lipoprotein (LDL) cholesterol, and apolipoprotein B have been shown to be associated with CNV burden, indicating an additional layer of genetic architecture shaped by rare, high-impact CNVs or more common variants with mild effects [27]. There is a growing body of literature on the use of CNV modelling for trait association testing (Additional file 1: Fig. S1, Additional file 2: Table S1), although the majority still rely on SNP arrays, which are limited by their low resolution [8]. Traditional SNP arrays lack the resolution to reliably detect small CNVs and are generally insensitive to the full spectrum of CNVs present across the genome, particularly those that are smaller but potentially impactful [67–69]. Moreover, the substantial overlap between genomic CNV loci and genome-wide association study (GWAS) signals, as well as their enrichment in recombination hotspots, suggests that CNVs may explain a portion of the missing heritability that cannot be accounted for by SNPs alone [70].

SNP and CNV signals can influence the measurement and interpretation of each other. In array-based genotyping, SNP-calling algorithms typically assume a fixed diploid copy number state and have limited ability to capture underlying CNV variation. Consequently, unmodelled CNVs can lead to incorrect genotype assignments and apparent deviations from the Hardy–Weinberg equilibrium. Conversely, ignoring SNP information during CNV calling may overlook allele-specific gains and losses and limit the ability to leverage the linkage disequilibrium (LD) between CNVs and nearby SNPs [71]. Many CNVs, particularly rare or structurally complex ones, are poorly tagged by SNPs, which means that standard approaches are unable to detect their effects [8, 72–78]. One major reason is that CNVs, particularly recurrent ones, can arise repeatedly at specific genomic loci due to elevated local mutation rates, resulting in multiple independent events with similar functional consequences. The *de novo* mutation rate for CNVs has been estimated at approximately 0.2 events per individual, reflecting the propensity of certain genomic regions to undergo recurrent structural changes. Consequently, CNVs are often not consistently associated with specific haplotypes, making their tagging by SNPs challenging [72–75, 79]. Consequently, the LD between CNVs and nearby SNPs is often weak or variable, particularly in repetitive sequence contexts, where elevated CNV mutation rates and increased genotyping errors further destabilise LD. Furthermore, multiallelic CNVs (mCNVs), which have a wide range of copy numbers beyond

the diploid expectation, can break typical LD patterns and generate population-specific runaway haplotypes that cannot be tagged by short variants. These mCNVs also tend to have a lower average observed LD with nearby SNPs [2, 72, 78, 80, 81]. The rarity and structural diversity of certain CNVs also limit the statistical power of standard SNP-based association methods to detect or accurately model them, while complex CNV rearrangements pose additional challenges for accurate interpretation and tagging [47, 52, 66]. According to the Wellcome Trust Case Control Consortium, 79% of common CNVs (minor allele frequency (MAF) > 10%) can be tagged by nearby SNPs with high LD. However, only approximately 22% of less common CNVs (those with MAF below 5%) are well-tagged by SNPs [77], and this tagging is shown to be even less effective in populations of African or African American descent due to lower LD between SNPs and CNVs [78]. Collectively, these factors make strong SNP-CNV tagging uncommon when considering the full spectrum of CNVs, particularly rare, complex, and multiallelic variants across the genome.

Consequently, CNV-specific association studies play a crucial role in uncovering genetic signals that are undetectable in SNP-based association studies. For instance, in a recent study analysing fine-mapped CNV association regions, 17% (23 out of 133 regions) were classified as CNV-only, highlighting the unique contribution of CNV analysis to genetic discovery [66]. CNV-only associations refer to genetic signals that can be detected exclusively through CNV GWAS but not by SNP GWAS. Examples of CNV-only associations illustrate the added value of CNV GWAS. For instance, CNVs affecting *SPDYE1*, *SPDYE6*, and *POLR*-related genes have been associated with chronotype [66]. Notably, these genes had not previously been linked to chronotype by SNP-GWAS in the UK Biobank, suggesting that CNVs in these regions may influence an individual's sleep patterns in a dose-dependent manner [66]. Additional CNV-only associations have been found, such as a region at the *ZDHHC11B* gene for the forced expiratory volume (FEV)/forced vital capacity (FVC) ratio and another at the *CDK11A* gene associated with standing height [66].

The mechanisms by which CNVs exert their effects can also differ fundamentally from those of SNPs, as CNVs can act through changes in dosage that may have opposite or identical phenotypic consequences that are not fully captured by SNP-based models [20, 26]. CNV analysis also informs our understanding of gene function and dosage sensitivity, generating or corroborating hypotheses about how specific genes influence biological processes and disease risk [20, 26, 27]. By studying pathogenic CNVs in the general population rather than solely in clinical cohorts, researchers can better characterise the full range of phenotypic outcomes, from severe early onset conditions to mild or asymptomatic manifestations. This aligns with a model of variable expressivity and incomplete penetrance, providing a more nuanced view of genotype–phenotype relationships [26, 27, 82–84].

From a clinical perspective, CNV association studies generate morbidity maps that help clinicians anticipate and monitor potential outcomes for carriers of specific CNVs [26, 85, 86]. While loss-of-function variants in genes such as *BRCA1* and *LDLR* are known to increase disease risk irrespective of variant type, CNV association studies

provide complementary insights by characterising the population-level impact of these alterations. In particular, they enable the assessment of penetrance, variable expressivity, comorbidities, and overall CNV burden in large cohorts, thereby refining our understanding of disease risk and clinical outcomes [26]. These insights improve risk stratification and support clinical decision-making, particularly for recurrent or multi-gene CNVs.

General workflow for copy number variation association analysis

The workflow for CNV association analysis identifies CNVs that are significantly associated with human traits and diseases and interprets their biological and clinical relevance (Fig. 3).

The key signal representations used for CNV detection across platforms and strategies for defining CNV regions (CNVRs) are illustrated in Fig. 4.

Data acquisition and initial processing

The workflow begins with data acquisition from array- or sequencing-based platforms. A detailed overview of the platforms and methodologies for CNV detection is provided in Additional file 3: Note S1. For array-based datasets, raw signals are processed to obtain the log R ratio (LRR), which represents the relative signal intensity, and the B-allele frequency (BAF), which shows the B allele proportion at each locus. For sequencing-based datasets, read depth profiles and alignment signatures are extracted, with split reads or discordant read pairs potentially included based on the CNV calling algorithm. This stage includes normalisation and QC to reduce technical artefacts before CNV detection [5, 19, 20, 24, 26, 66, 86, 134–141].

(See figure on next page.)

Fig. 3 Workflow of copy number variation (CNV) association analysis. The process begins with data acquisition from multiple platforms, including SNP arrays, short-read exome sequencing (ES), short-read genome sequencing (GS), and long-read sequencing (LR-seq). ES is represented here by read depth signals, which constitute its primary input for CNV detection. Following initial normalisation and quality control (QC), CNV calling is conducted using platform-specific CNV callers or ensemble/merging strategies. Some callers listed in this figure, such as DRAGEN, represent comprehensive end-to-end analysis platforms applicable across multiple sequencing modalities, including both ES and GS, rather than standalone CNV detection algorithms. Notably, there are two exceptions: probe-level analyses, which utilise normalised signal intensities from SNP arrays directly, and exon-level analyses, which test the normalised read depth per exon in ES. These exceptions bypass CNV calling by directly engaging with the quantitative signals. After CNV calling (or signal-level approaches in these exceptions), post-calling QC is implemented to ensure reliability by filtering out low-quality events and regions with noise. Detected CNVs are then categorised at various resolutions for association testing, including probe-level (signal or proxy-state), tiling-window, exon-level, region-level (CNVRs), CNV-level, or gene-level. Although other grouping schemes could be theoretically envisioned, these levels represent the approaches most frequently adopted in practice. Statistical association testing is conducted using standard regression models (linear, logistic, mixed models), burden tests, or kernel-based approaches, supported by an expanding array of dedicated CNV association tools, such as PLINK [46, 47], ParseCNV2 [49], CNVtools [87], CNVRanger [88], and CNest [66]. Effect modelling frameworks (dosage-based, categorical, probabilistic, or type-specific) and post-association analysis steps (fine-mapping, replication, orthogonal validation, integration with SNP-GWAS, pathway analysis, and clinical annotation) further refine biological and clinical inferences. Created with BioRender.com

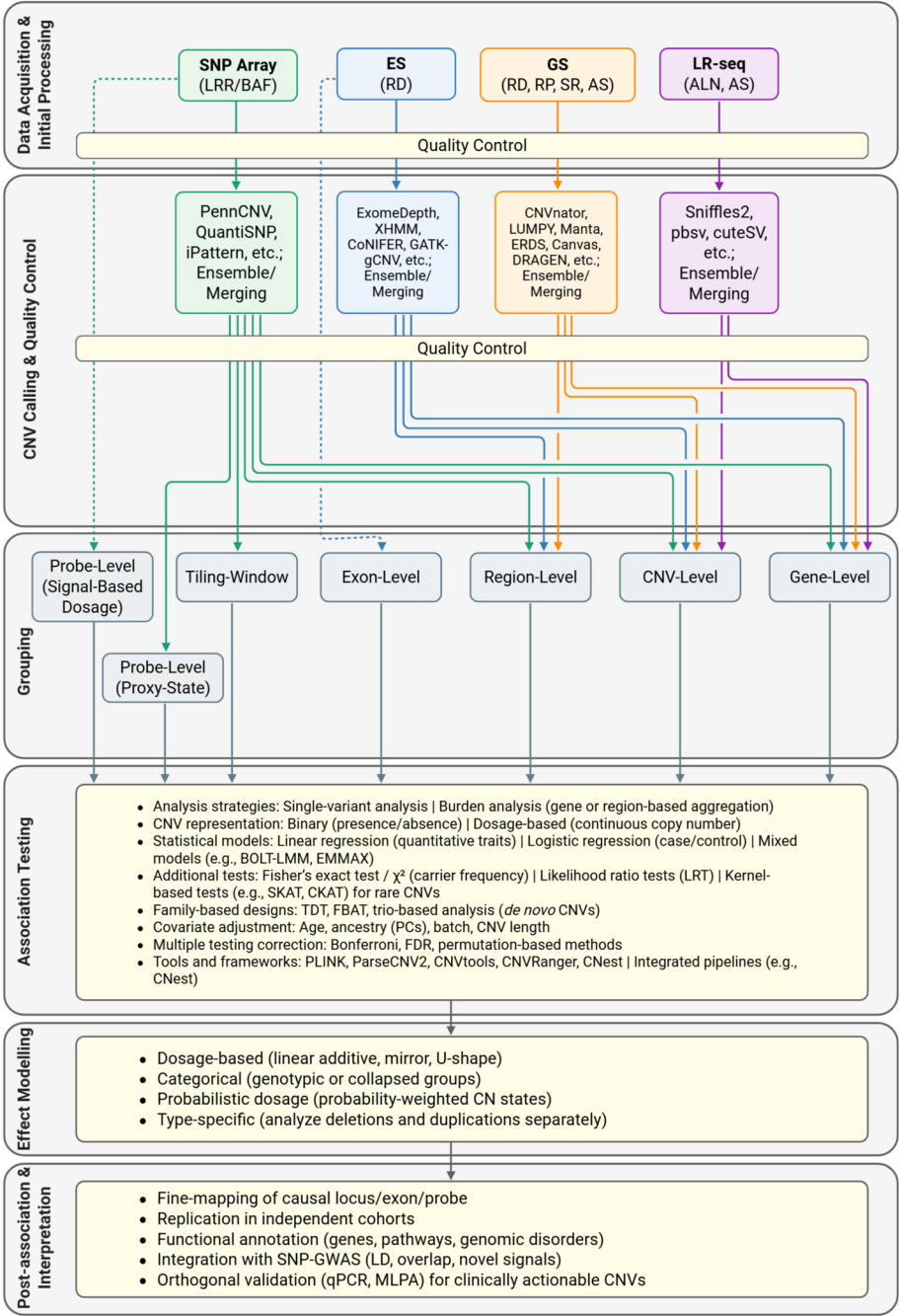


Fig. 3 (See legend on previous page.)

CNV calling

After preprocessing, CNV calling algorithms are applied to infer individual copy number states (see Additional file 3: Note S2 and Additional file 3: Note S3 [103, 113, 120, 142–156] for details on CNV detection strategies and segmentation approaches). CNV detection performance depends strongly on the underlying platform, with key differences in resolution, breakpoint accuracy, and sensitivity to CNV size. Different tools employ statistical models to integrate various signal features to

call CNVs (Additional file 1: Table S2 [66, 105–107, 109–114, 119–122, 125–127, 129, 131, 142, 144, 145, 147, 149–151, 157–207] and Additional file 1: Figs. S2–S8 [208]). To increase reliability, it is common to use multiple algorithms in parallel [28].

Approaches and tools for copy number variation detection via ensemble and merging strategies

Ensemble and merging approaches for CNV detection are designed to overcome the limitations of individual calling tools, such as high false-positive rates, low concordance, and incomplete detection [28]. Because no single CNV detection method consistently delivers the best results across various contexts, combining the strengths of multiple tools has emerged as a promising strategy for CNV detection.

These approaches can be broadly categorised based on their primary strategies for combining and refining the variant calls. These ensemble and merging strategies have been developed across different platforms, including SNP arrays, exome sequencing (ES), genome sequencing (GS), and LR-seq, and their applicability is often platform specific. One category is simple consensus or overlap-based approaches, which rely on the direct overlap of variant calls or require a minimum number of callers to agree on a variant to define high-confidence calls. Although straightforward, these methods may not always achieve the highest precision or recall. For example, HugeSeq, developed for GS data, identifies high-confidence SVs and CNVs as those detected by two or more algorithms with at least 50% reciprocal overlap, integrating outputs from tools such as BreakDancer, Pindel, CNVnator, and BreakSeq using BEDtools [183]. NextSV, designed for SV detection from low-coverage LR-seq data, integrates the results from multiple SV callers to improve robustness. It produces two call sets: a sensitive set (union of all calls) that maximises recall, and a stringent set (intersection of calls) that prioritises precision. For instance, in the case of deletions, calls are combined based on reciprocal overlap thresholds, which are typically 50% or higher [173].

(See figure on next page.)

Fig. 4 Copy number variation (CNV) signal representations and CNV region (CNVR) definition strategies. **A** Illustrations of CNV signal representation across various platforms are provided. For SNP arrays, normalised metrics, such as B-allele frequency (BAF) and log R ratio (LRR), are presented alongside copy number states. In short-read sequencing-based methodologies, pertinent CNV evidence encompasses read depth (RD), discordant read pairs (RP), split reads (SR), assembly (AS), and long-read sequencing (LR-seq), using either alignment- or assembly-based approaches, where deletions and duplications are identified through alignment gaps or supplementary alignments. These signals provide complementary evidence of CNVs, with RD reflecting dosage changes, RP indicating abnormal insert sizes or orientations, and SR enabling breakpoint resolution; these same signatures are also used during manual inspection (e.g., in IGV) to distinguish true variants from technical artefacts. Additional technical details on detection methodologies and signal processing are available in Additional file 3: Note S1 [2, 28, 29, 56–61, 89–103] and Additional file 3: Note S2 [28, 94, 103–132]. **B** Schematic representation of CNVR definitions and common CNVR patterns observed in case–control analyses. Overlapping CNV calls across individuals can be consolidated into CNVRs using various methods, including trimming, reciprocal overlap (RO), and fragment-based techniques [50]. Given the variability in CNV size and breakpoints, distinct CNVR patterns may emerge in case–control data. Well-behaved CNVRs show highly concordant CNV boundaries across samples. Random boundary variance reflects minor stochastic variation in breakpoints. Multiple significant CNVRs represent distinct clusters of overlapping CNVs. A CNV peninsula denotes a small case-specific extension with limited probe support adjacent to a shared region. Central consensus with variable extension indicates a shared core CNV with variable flanking boundaries. Control encroachment occurs when CNVs in controls overlap regions defined primarily by cases [52]. Parts of the schematic were generated using the R package karyoploteR [133]. Created with BioRender.com

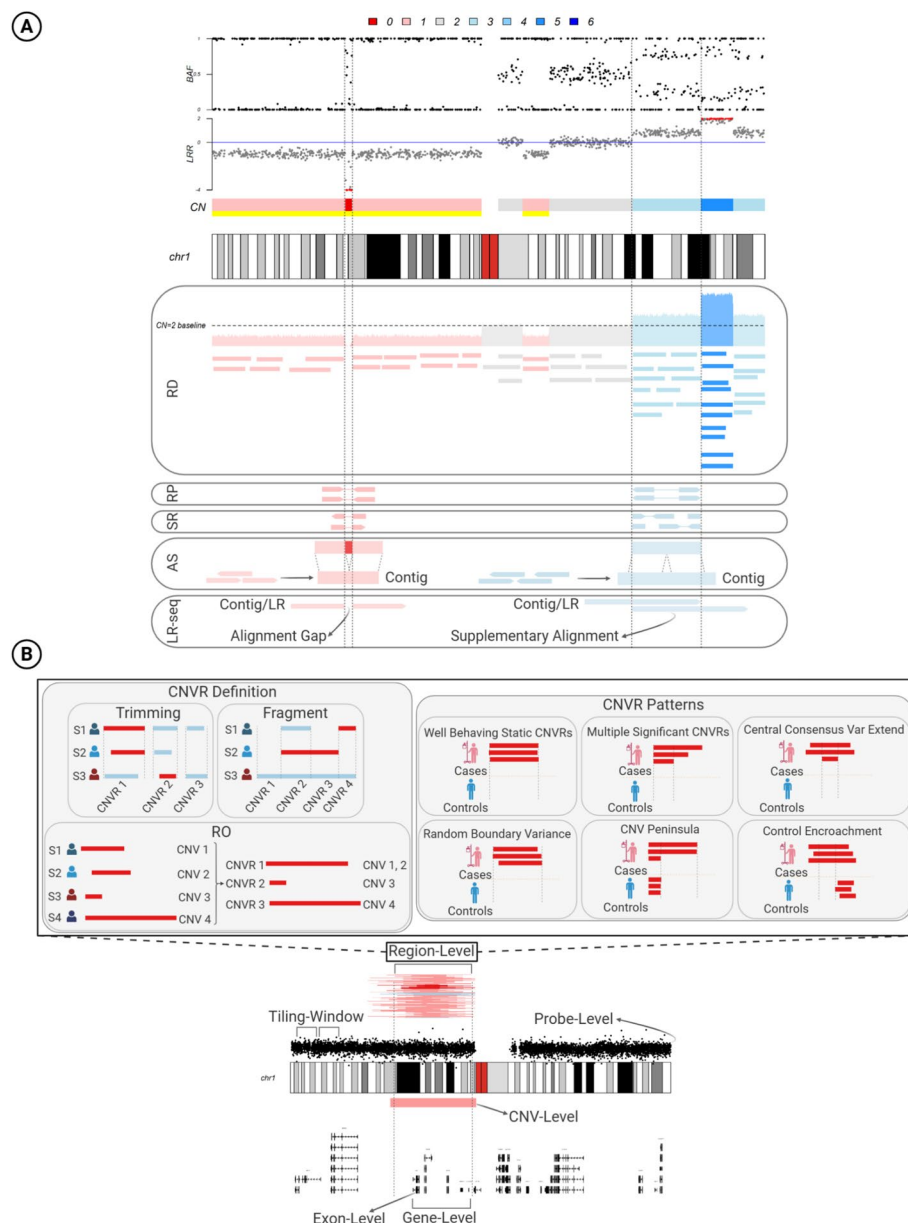


Fig. 4 (See legend on previous page.)

SURVIVOR is another widely used tool that filters and merges SVs based on user-defined criteria, such as coordinate distance, SV type, and chromosome, and supports tools such as Delly, LUMPY, Pindel, and cn.MOPs [176].

Although combiSV also combines outputs from multiple callers, it is not a simple overlap-based method. It is a recent ensemble tool specifically designed to improve SV detection from LR-seq data by combining the outputs of multiple SV detectors. It takes variant call format (VCF) outputs from up to six different tools, namely cuteSV, pbsv, Sniffles, NanoVar, NanoSV, and SVIM, and integrates their calls to produce a consensus SV call set with increased recall and precision. Unlike simple overlap-based methods, combiSV uses caller-specific prioritisation to select the most accurate information for each SV parameter (e.g., position, length, and genotype) based on benchmarking

results, placing it in the category of advanced ensemble methods. It also allows users to set thresholds for the minimum number of supporting callers and read the coverage for each variant [171].

A more sophisticated category involves machine learning and statistical integration approaches that use data-driven models to assign weights or integrate features across variant calls, leading to better accuracy and a reduced number of false positives. CN-Learn, designed for ES data, is a random forest-based framework that integrates CNV calls from multiple algorithms, such as CANOES, CODEX, XHMM, and CLAMMS, using features such as GC content and mappability, trained on a subset of validated CNVs [193]. FusorSV, which targets GS data, is part of the Structural Variation Engine and uses a data mining approach to assess tool performance against a truth set. It combines calls using a mutual exclusion strategy that minimises false positives while maximising sensitivity [161].

Another group of tools falls under advanced merging with refinement or validation pipelines, which not only combine calls but also refine breakpoints, re-genotype variants, and validate them using additional evidence. For instance, EnsembleCNV detects and genotypes CNVs from SNP array data using a two-phase pipeline involving initial detection and a re-genotyping step that uses local likelihood models to refine the boundaries [131]. MetaSV, which is optimised for GS data, merges calls from multiple SV detection tools using intra- and inter-tool merging strategies, integrates local assembly via SPAdes, refines breakpoints, performs genotyping, and annotates variants [189]. Parliament2, designed for the scalable analysis of GS data, integrates multiple SV callers, uses SURVIVOR for merging, SVTyper for validation, and assigns a quality score to each SV based on supporting evidence [205]. SVMerge, also developed for GS data, takes a modular approach by integrating SV calls from various tools and refining them using local de novo assembly, making it extensible to future SV calling tools [201].

Ensemble and merging algorithms improve CNV detection but face key limitations, including a lack of standardisation, variable sequencing coverage, dependency on specific algorithm combinations, inconsistent benchmarking, and immature standalone tools. For an overview of ensemble algorithms for SV detection, readers should refer to the detailed review by Ho et al. [28].

Overview of benchmarking

Benchmarking SVs, including CNVs, is a crucial process in genomic research that is designed to evaluate and validate the accuracy of variant calling methods. The constant progress in genomic sequencing technologies has led to enhanced detection of SVs, particularly with the advent of LR-seq, which allows for the refined characterisation of these variants [42].

Benchmarking consists of four core components. The first is the establishment of truth data [42], a gold standard or ground truth dataset (such as those produced by the Genome in a Bottle Consortium, GIAB [209]), to which the analytical results can be compared. These truth sets, also referred to as high-confidence or ground truth call sets, are typically generated through the integration of multiple sequencing technologies, variant calling methods, and extensive manual curation to minimise platform-specific biases [42]. However, widely used resources such as those

from the GIAB consortium are derived from a limited number of individuals and may not fully capture the diversity or spectrum of clinically relevant CNVs across populations [42]. Multiple benchmark datasets are available for assessing SVs, including CNVs (reviewed by Majidian et al. [42]). An important complement to GIAB-derived truth sets is the Platinum Pedigree resource [210], which leverages Mendelian inheritance across a multi-generation pedigree (CEPH-1463) sequenced with PacBio HiFi, Illumina, and Oxford Nanopore Technologies. By tracking haplotype transmission within the family, this resource validates variant calls genome-wide, including single nucleotide variants, indels, tandem repeats, and SVs, across difficult genomic regions often excluded from conventional truth sets, prioritising sensitivity and completeness over specificity. The second component involves specific tools and algorithms used to align and compare analytical outputs with truth data. Third, performance metrics are employed to comprehensively assess concordance and diagnostic accuracy using standard statistical measures. These include counts such as true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), from which key indicators are derived. Sensitivity (or recall) measures the proportion of actual positives correctly identified ($\frac{TP}{TP+FN}$), whereas specificity measures the proportion of actual negatives correctly identified ($\frac{TN}{TN+FP}$). Precision, also referred to as the positive predictive value (PPV), indicates the proportion of predicted positives that are truly positive ($\frac{TP}{TP+FP}$). Conversely, the negative predictive value (NPV) shows the proportion of predicted negatives that are truly negative ($\frac{TN}{TN+FN}$). To balance precision and recall, the F1 score, calculated as the harmonic mean of these two metrics ($\frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$), is used as a single metric reflecting overall performance [42]. Finally, effective benchmarking necessitates a clear representation of results, which is commonly achieved through visualisations and comprehensive reports designed to convey findings in an interpretable manner [211].

One of the key challenges in SV/CNV benchmarking is that the same genomic alteration can be reported differently by various tools, often with varying breakpoint coordinates, reference alleles, and event descriptions, complicating direct cross-method comparisons [42]. Furthermore, the selection of comparison parameters critically impacts benchmarking outcomes. Overly permissive SV-matching thresholds may cause over-merging of distinct variants, artificially inflate performance metrics, introduce incorrect annotations, falsely classify unique SVs as shared, reduce observed allelic diversity, and overestimate allele frequencies. In contrast, overly stringent thresholds can underestimate performance, miss valid annotations, and reduce the statistical power of downstream analyses [212, 213]. CNV detection is particularly challenging in repetitive genomic regions, where variants are frequently mediated by repetitive elements that complicate alignment and variant resolution [4]. Although advances in LR-seq have improved the characterisation of these regions, substantial challenges remain in establishing reliable truth sets and comparing variant calls, primarily due to breakpoint degeneracy and the existence of multiple equally valid breakpoint placements [29, 42, 214]. Consequently, many benchmark datasets deliberately exclude highly repetitive regions [209, 215, 216]. Despite these limitations, truth sets have been highly effective in several contexts, including standardising performance evaluation across tools using metrics such as precision and recall,

benchmarking variants in structurally complex but medically relevant regions, and enabling the development of assembly-based benchmarking approaches (e.g. TT-Mars), which improve evaluation in regions where coordinate-based comparisons are unreliable [42, 214]. Finally, the heterogeneity of output formats produced by different variant calling tools necessitates rigorous standardisation and normalisation prior to comparative analysis [42].

Benchmarking tools are essential frameworks in genomics that enable researchers to objectively evaluate and compare the performance of SV and CNV detection methods using standardised reference datasets and curated truth sets.

Truvari is widely recognised for benchmarking SV callsets by computing the precision, recall, F1-score, and genotype concordance through rigorous comparisons of test callsets with validated benchmark sets. It supports flexible matching criteria, including breakpoint proximity, size similarity, and sequence similarity, making it suitable for both high-resolution and less precise SV calls [213].

SVanalyzer SVbenchmark employs advanced sequence-based matching algorithms, enabling a robust assessment of SV calls with differing representations, particularly in complex or repetitive genomic regions, and improves accuracy even for variants with imprecise breakpoints or those within tandem repeats [209].

TT-Mars introduces a distinct benchmarking paradigm by evaluating SV calls against high-quality haplotype-resolved genome assemblies rather than solely against conventional truth sets. Instead of matching a predefined set of calls, TT-Mars assesses whether the sequence content implied by each SV call aligns with the actual sequence in the assembled haplotypes, making it especially valuable in regions lacking curated truth sets or where variant representation is ambiguous owing to genomic complexity [214].

CNVbenchmarker [217] and its enhanced successor, CNVbenchmarker2 [38], are comprehensive R-based frameworks for systematically benchmarking CNV calling tools using targeted next-generation sequencing gene panel data. CNVbenchmarker automates the execution and evaluation of multiple CNV callers, including DECoN, CoNVaDING, panelcn.MOPS, ExomeDepth, and CODEX2, using validated datasets with gold-standard confirmation via multiplex ligation-dependent probe amplification (MLPA) or array-based comparative genomic hybridisation (aCGH). It calculates an extensive range of performance metrics, including sensitivity, specificity, PPV, NPV, and F1-score, and supports dataset-specific parameter optimisation [217]. CNVbenchmarker2 extends these capabilities to benchmark a broader set of CNV tools, such as Atlas-CNV, ClinCNV, CNVkit, and GATK-gCNV, and allows the systematic exploration of over 100 tool parameters, generation of detailed graphical reports, and evaluation of meta-caller strategies that combine the results from multiple tools [38]. Both frameworks are designed to be user-friendly and adaptable, enabling research and clinical laboratories to benchmark CNV detection tools directly using their datasets.

witty.er (<https://github.com/Illumina/witty.er>) is a dedicated benchmarking tool for large variants that is explicitly designed to assess the performance of SV and CNV callers. Developed by Illumina, it compares query VCFs to truth VCFs for multiple SV types, applies flexible positional and genotype matching criteria, and produces detailed benchmarking statistics with annotated VCFs. Functionally, it is analogous to hap.py for

small-variant benchmarking, making it a central resource for performance assessment in large variant calling.

svclassify serves as a foundational resource by enabling the creation of high-confidence benchmark SV call sets. Rather than acting as a benchmarking evaluator, it integrates multi-platform sequencing evidence and applies machine learning algorithms to classify candidate SVs as likely true- or false-positive. The resulting high-confidence SV sets serve as truth sets for downstream benchmarking, ensuring reliable performance evaluation and supporting the reproducibility of SV benchmarking studies [218].

Supporting tools such as SURVIVOR play complementary roles in benchmarking workflows. SURVIVOR can simulate SVs for benchmarking, merge and filter SV callsets, and generate consensus or simulated truth datasets for evaluation, although it does not produce standardised benchmarking reports [176].

By offering standardised workflows for metric calculation, parameter optimisation, and cross-tool comparison, and by integrating with auxiliary tools for callset preparation, simulation, and validation, these benchmarking solutions serve as foundational elements for the development of CNV tools, method selection, and validation of clinical pipelines in both research and diagnostic settings.

Insights from independent benchmarking studies of copy number variation callers

Bioinformatic algorithms are essential for CNV detection across sequencing and micro-array platforms. Because benchmarking results vary substantially depending on the data type, CNV size, and study design, findings are best interpreted in a platform-aware context. With a wide range of algorithms available, this review synthesises findings from independent benchmarking studies to provide an unbiased overview of tool performance across different contexts. Detailed platform-specific benchmarking results and tool comparisons are provided in Additional file 3: Note S4 [33–35, 37–39, 41, 43–45, 124, 162, 197, 217, 219–229].

Overall, systematic benchmarking of independent CNV callers across various technologies has revealed that no single tool or approach is universally optimal for every experimental context, CNV size, or type. Rather than employing a simplistic overlap strategy, the iterative merging of top-performing tool pairs can enhance both accuracy and sensitivity. However, not all combinations are advantageous, underscoring the need for systematic benchmarking to identify synergistic combinations. Therefore, achieving robust and accurate CNV detection requires a multi-layered strategy involving careful selection and parameter tuning of tools, often necessitating the intelligent combination of multiple callers and rigorous post-processing steps. Key steps include optimising parameters, such as window size, probe selection, caller-aligner pairing, and reference set composition, as well as removing recurrent artefacts through custom filtering and applying quality control thresholds. Comprehensive post-processing and quality checks, such as depth of coverage analysis, uniformity assessment, and outlier detection, are critical for minimising false positives and improving confidence in detected CNVs [28, 35–39, 41, 43, 44, 197, 217, 226–229].

Crucially, computational predictions alone are insufficient for high-confidence CNV calls, particularly for rare, small, and clinically relevant variants. All such findings should be validated using orthogonal methods, ranging from visual inspection and Binary

Alignment Map (BAM) review to quantitative PCR (qPCR), MLPA, or array-based approaches before being reported or interpreted clinically. Incorporating post-calling steps, such as genotyping refinement, confidence scoring (potentially using machine learning or Bayesian models), and integration of haplotype or alternative reference information, can further enhance the reliability of CNV calls. Finally, practical considerations, such as workflow reproducibility (e.g., through containerisation and workflow managers), user-friendliness, integration with existing clinical or diagnostic pipelines, and laboratory bioinformatics capacity, are equally important for successful CNV analysis [33, 34, 36–39, 43–45, 197, 217, 220, 222, 225, 226, 229]. While these approaches are effective for detecting CNV events, it is important to distinguish between CNV discovery and genotyping. Association studies, particularly those involving common or multiallelic CNVs, require accurate estimation of copy number states across cohorts, as implemented in cohort-based and mixture-model approaches, such as CNest and GenomeSTRiP [66, 230].

Quality control and pre-processing

QC is a critical step, both before and after CNV calling, to ensure sample- and variant-level reliability prior to downstream analysis. At the pre-calling stage, individuals with low genotyping rates, outlier LRR or BAF metrics, sex discordance, or abnormal coverage profiles (for sequencing data) are excluded [21, 66, 134, 136–141, 231–235]. Additional corrections, such as GC content adjustment to mitigate hybridisation bias and principal component analysis (PCA) to remove batch effects, are also applied at this stage [20, 66, 86, 134–141, 231]. At the post-calling stage (or equivalently, during variant-level QC for signal-based analyses), low-quality loci are removed. For CNV calls, this includes excluding samples with excessive CNV calls (suggestive of poor DNA quality), filtering events in noisy or polymorphic loci, and removing calls that fail the minimum thresholds for probe density, read depth, or length. For signal-based analyses (e.g., probe- or exon-level association), this includes excluding units with poor coverage, extreme variance, or low mappability. Adjacent CNV calls in the same individual may be merged to prevent over-fragmentation, and problematic genomic regions prone to spurious calls or somatic variation, such as centromeres, telomeres, immunoglobulin loci, T-cell receptor regions, segmental duplications, and low-copy repeats, are routinely excluded [20, 21, 76, 86, 134–141, 231, 232, 236–238]. As discussed in the previous section, high-confidence CNVs are typically defined by concordance across multiple detection methods, probabilistic quality scores, or independent experimental validation [19, 20, 26, 76, 134, 236, 238]. In addition, visual inspection is a critical quality control step used to validate CNV calls and filter out potential artefacts. Researchers typically examine several key characteristics during this process, including signal intensity patterns and allele frequency distributions in array-based data (e.g., LRR and BAF), as well as read depth profiles, split-read and discordant read pair signals, and local mapping quality in sequencing data, often visualised using genome browsers such as Integrative Genomics Viewer (IGV). In addition, genomic context is evaluated using resources such as UCSC Genome Browser, for example, to assess proximity to repetitive elements, segmental duplications, or low-mappability regions, which may indicate potential artefacts [20, 24, 26, 27, 86, 135, 139, 232, 238, 239].

Copy number variation grouping, statistical association testing, and effect modelling

Approaches to CNV association analysis can be organised hierarchically based on three major components: (i) strategies for grouping CNVs, (ii) statistical frameworks for association testing, and (iii) effect modelling schemes for representing CNVs within a statistical model. The choice at each level should be guided by the study design, sample size, variant frequency, expected effect size, and underlying biological hypotheses.

At the grouping stage, CNVs can be analysed at different levels of granularity. In probe-level analysis, individual probes on genotyping arrays are used as proxies for copy number, with the estimated copy number at each probe typically entered into regression models as a continuous (linear dosage) variable. This eliminates the need for complex CNV overlap criteria, making it efficient for large-scale array-based studies [19, 20, 27]. Gene-level analysis aggregates rare CNVs based on the genes they overlap, especially those predicted to have a functional impact (e.g., loss-of-function), thereby increasing the statistical power [18, 20, 138, 240]. Importantly, the criteria used to assign CNVs to genes vary widely across studies and can substantially influence the association results. These range from simple intersection-based definitions, such as any overlap (≥ 1 bp) [21, 26, 135], to functionally informed approaches requiring overlap with specific elements (e.g., coding exons, splice sites, or UTRs) [27, 241], with exon-centric definitions commonly used for genic burden analyses [21, 136, 238, 239]. Some approaches extend gene boundaries by incorporating flanking regions (e.g., 10–50 kb) to capture potential regulatory effects [18, 76]. More stringent strategies include proportional or reciprocal overlap thresholds (e.g., $\geq 50\%$), minimum overlap with predefined critical regions, or requiring that a large fraction of the CNV lies within the target region [20, 26, 66, 85, 136, 231]. When deletions span multiple genes, they are typically assigned to each affected gene and included independently in gene-level analyses [18]. In such cases, additional prioritisation strategies may be applied, for example, by focusing on variants predicted to disrupt coding regions (pLoF events) or by stratifying analyses based on gene-level intolerance metrics, such as the probability of loss-of-function intolerance (pLI) or the loss-of-function observed/expected upper bound fraction (LOEUF), which can help identify biologically relevant signals and potential driver genes within multi-gene events [26, 86, 242]. Region-level analysis merges overlapping CNV calls into CNVRs and tests these merged regions for association with the phenotype [26, 27, 134, 136, 243]. In sequencing-based data, particularly GS, CNV grouping requires additional consideration because of the higher breakpoint resolution and variability across detection algorithms [5, 18]. Although breakpoints can be defined at near base pair resolution, different callers may produce slightly discordant or fragmented calls for the same event [244]. Consequently, clustering strategies, such as reciprocal overlap thresholds or distance-based merging, are commonly applied to harmonise CNVs across samples and tools while preserving breakpoint precision [244]. In tiling-window analysis, the genome is divided into fixed-size, potentially overlapping windows (e.g. 10 kb), and CNVs within each window are aggregated for unbiased genome-wide scans [76, 231]. At the CNV level, individual CNVs, common or rare, are tested separately, which is most appropriate for well-characterised variants with sufficient frequency and expected effect sizes [18, 20].

At the association testing stage, different analytical paradigms can be applied to grouped data. Single-variant analysis tests each probe, CNV, or region independently, whereas burden-style analysis aggregates multiple CNVs (e.g. all rare CNVs overlapping a gene) to evaluate their collective effect [18, 19, 21, 26, 27, 66, 76, 134, 139, 236–239, 245]. Standard regression methods, such as logistic regression for binary traits and linear regression for quantitative traits, are widely used [18, 19, 26, 27, 66]. In sequencing-based analyses, CNVs can be modelled using continuous dosage estimates derived from the read depth or allele balance, rather than binary presence/absence. This approach is analogous to SNP dosage models used in GWAS and enables more precise effect estimation compared to array-based approaches [66]. Fisher's exact test and the chi-square test are commonly used to compare the frequency of a CNV between groups; in small samples or when expected cell counts are low, Fisher's exact test is preferred because it calculates the exact probability [26, 134, 139, 231, 237, 239, 246]. Likelihood ratio tests (LRTs) can be used to assess the significance of CNV terms within regression models [136]. Other methods include kernel-based tests (e.g., SKAT, CKAT, CONCUR, MCKAT, SMCKAT) for rare-variant contexts with heterogeneous effects [158, 247–252], Cox proportional hazards models for time-to-event outcomes such as age at disease onset [26], mixed models (e.g., EMMAX [253], BOLT-LMM [254, 255]) to account for relatedness and population structure [20, 76], permutation-based tests for empirical significance estimation [134, 139, 239, 245], and family-based association tests in related cohorts [237, 256, 257]. Family-based association analyses, including family-based association tests (FBATs) and transmission disequilibrium-based approaches, are particularly important in CNV studies, as they leverage pedigree information (e.g. trios or multiplex families) to evaluate transmission patterns of variants [236, 237, 241]. Trio-based designs are particularly valuable for detecting de novo CNVs, assessing Mendelian consistency, and investigating their contribution to disease risk, especially in neurodevelopmental disorders such as autism spectrum disorder [5, 70, 78, 236, 237, 241]. The transmission disequilibrium test (TDT) evaluates whether alleles or CNVs are transmitted from heterozygous parents to affected offspring more frequently than expected by chance, providing a robust framework that is resistant to population stratification [258]. In addition, conditional logistic regression stratified by family enables within-family comparisons by treating relatives as matched sets, thereby controlling for shared genetic background and environmental factors [237, 259]. In addition, some studies have focused specifically on de novo CNVs and assessed their association using regression-based frameworks (e.g. logistic regression), particularly in trio-based designs [241]. These designs are inherently robust to population stratification and control-related biases, which can confound case–control analyses [237]. Furthermore, these approaches facilitate the assessment of variant segregation within families, helping to clarify the pathogenic relevance of CNVs and identify variants that may contribute to disease susceptibility [241].

In addition to autosomal CNVs, the analysis of sex chromosomes presents distinct methodological challenges owing to differences in baseline ploidy between sexes, which violate standard diploid assumptions and require sex-specific normalisation or modelling strategies (Additional file 3: Note S5 [21, 25, 27, 47, 49, 50, 52, 66, 86, 134, 138, 158, 197, 225, 238, 260–262]). Despite these complexities, accurate detection of sex chromosome CNVs remains important due to their established roles in developmental disorders and sex-specific disease risk [263].

Furthermore, to reduce confounding and improve the accuracy of effect estimation, CNV association analyses routinely adjust for relevant covariates, such as age, ancestry (commonly represented via principal components), genotyping array or batch, assessment centre, and CNV length [18–21, 26, 66, 136, 238]. The inclusion of these covariates helps to control for population stratification and technical artefacts. Importantly, CNV patterns and their phenotypic associations can differ substantially across ancestral populations due to variation in demographic history, LD structure, and reference-related biases [8, 78]. CNV frequencies are often population-specific, which can influence statistical power, effect size estimation, and the generalisability of association findings [19, 78]. In particular, reduced LD between CNVs and SNPs in certain populations limits tagging efficiency, whereas the reliance on reference genomes derived predominantly from European populations may introduce alignment bias and affect CNV detection accuracy [8, 27, 78]. These factors can result in population-specific signals and reduced replication across studies if not properly accounted for, highlighting the importance of ancestry-aware analytical strategies in CNV association analysis.

Given the large number of statistical tests performed in genome-wide CNV studies, appropriate multiple test corrections are essential to control false positives. Common approaches include the Bonferroni correction, false discovery rate (FDR) control (e.g. the Benjamini–Hochberg procedure), and permutation-based methods [18, 24, 26, 85, 134, 231, 236, 237, 239, 245]. In some cases, the “effective number of tests” approach is applied to set genome-wide significance thresholds, where the effective count is estimated from the correlation structure among tests arising from overlapping or co-occurring CNVRs rather than assuming full independence [19, 26].

In the effect modelling stage, the representation of CNV status within the statistical model is determined. Dosage-based models treat copy numbers as continuous variables. The basic linear or additive dosage model assumes a proportional change in effect with each unit copy number change [66]. Specific forms include the mirror-effect model, where deletions and duplications have equal magnitude but opposite effects, and the U-shape model, where both deletions and duplications affect the phenotype in the same direction [19, 26, 27]. Probabilistic dosage models incorporate uncertainty in CNV calls by weighting the effects according to the probability of each copy number state [19]. Alternatively, categorical models treat each copy number state as a discrete category (genotypic model) or collapse the states into broader groups (dominant/recessive coding) [76, 243, 264]. Finally, type-specific models analyse deletions and duplications separately, allowing for the identification of distinct biological effects [20, 26, 27].

Copy number variation association tools

Specialised tools are essential for CNV association studies because CNVs differ fundamentally from SNPs in terms of their biological nature, measurement properties, and analytical complexities. Unlike SNPs, which are typically biallelic, CNVs can encompass a wide range of integer copy numbers and may manifest as deletions or duplications spanning entire genomic regions, resulting in more than three possible genotypes and increased allelic diversity [71, 88, 265]. CNV measurement is inherently noisier than SNP genotyping because of technical variations in intensity signals and systematic biases between cases and controls, which, if unaddressed, can inflate false-positive rates [71,

87]. Defining and merging CNVRs across individuals adds further complexity, as CNVs vary considerably in size and breakpoint locations. Accurate boundary refinement and region merging require advanced algorithms to handle overlapping calls, fragmented CNVs, and complex rearrangements [49, 50, 52, 266]. Importantly, the choice of merging strategy can substantially influence downstream analyses. Overly permissive merging may combine distinct variants and inflate false-positive signals, whereas overly stringent criteria may fragment shared CNVs, reduce statistical power, and obscure true associations [50, 52]. Breakpoint heterogeneity across individuals further complicates this process and may dilute association signals if not appropriately modelled [52]. In addition, different merging strategies (e.g., reciprocal overlap, density-based trimming, or fragment approaches) can lead to markedly different CNVR definitions, affecting both the localisation and interpretability of association signals [50, 66, 88]. Moreover, CNV data require specialised QC pipelines and effect models tailored to CNV biology. Standard SNP-based GWAS tools lack native support for these complexities, often missing crucial QC steps and statistical models that account for CNV-specific uncertainties and the more complex, often non-linear relationships between copy number and phenotypes [46, 47, 49, 52]. Consequently, the unique challenges posed by CNV data in terms of calling, representation, uncertainty, and association analysis necessitate the development of dedicated tools specifically tailored to the complexities of CNV association studies. Although various methods have been devised for calling CNVs, there is a scarcity of strategies for CNV association analysis (Additional file 1: Table S3 [46, 47, 49, 50, 52, 66, 71, 87, 88, 122, 181, 261, 265–268] and Additional file 1: Figs. S2–S8). Dedicated tools have nonetheless been developed, ranging from early extensions of GWAS software (for example, PLINK) to integrated pipelines that couple detection and association (for example, Birdsuite, R-GADA, ParseCNV/2), likelihood-based statistical frameworks (for example, CNVtools, CNVassoc), CNVR-defining approaches (for example, CNVRuler and CNVRanger), more recent scalable solutions for next generation sequencing (NGS) data (for example, CNest), and emerging frameworks tailored to LR-seq that integrate discovery, genotyping, phasing, and imputation for population-scale CNV association. For details and examples of their applications, see Additional file 3: Note S6 [5, 21, 26, 46, 47, 49, 50, 52, 66, 70, 71, 77, 87, 88, 122, 125, 181, 230, 261, 265–297].

Post-association analysis and interpretation

Post-association analysis and interpretation refine and validate CNV findings and place them in a broader biological context. Fine-mapping approaches, including stepwise conditional analyses or identification of the most strongly associated probe or exon within a locus, help pinpoint likely causal CNVs or sub-regions within that locus [20, 27, 66, 231]. Replication in independent cohorts and orthogonal experimental validation methods, such as qPCR, MLPA, or GS, are essential for establishing the robustness of these associations. Integration with SNP GWAS data allows the assessment of overlap, LD relationships, and the potential for CNVs to provide novel association signals beyond those captured by SNPs [20, 26, 27, 66, 76]. Functional annotation and pathway analyses

identify enriched biological processes, pathways, or protein interaction networks among associated loci [26, 76, 134, 236–238], whereas clinical relevance can be evaluated by mapping CNVs to known genomic disorders, syndromes, or by assessing pleiotropic effects [20, 26, 70, 232]. Additional integrative analyses, such as colocalisation with expression or epigenetic QTLs and phenome-wide association studies (PheWAS), can provide deeper insights into the molecular mechanisms and phenotypic spectrum of CNV effects [18, 27, 240]. Finally, sharing results, particularly summary statistics, through public repositories such as the GWAS Catalogue facilitates reproducibility, secondary analyses, and translation into clinical or public health applications [18, 20, 21, 27, 66].

Conclusion and future perspectives

CNVs are a major class of SVs that shape human genetic diversity and influence susceptibility to a wide range of diseases. Substantial progress in sequencing technologies and computational frameworks has greatly enhanced CNV detection; however, important challenges persist in translating CNV discoveries into reliable association studies. A central limitation is that no single CNV calling tool consistently outperforms the others across all experimental settings; thus, the combined use of multiple algorithms remains the most robust strategy for CNV detection. Beyond detection, rigorous QC, careful CNV grouping, and appropriate statistical frameworks for association testing and effect modelling are essential to ensure valid and interpretable results. Together, these considerations underscore the complexity of CNV analysis and highlight the need for continued methodological innovation to fully elucidate the contribution of CNVs to health and diseases.

The landscape of CNV analysis is anticipated to experience profound changes driven by advances in reference genome assemblies, integrative methodologies, benchmarking standards, and computational tools. The development of comprehensive reference genomes, particularly the telomere-to-telomere (T2T) CHM13 assembly, has provided a gap-free human genome sequence that corrects structural errors and adds approximately 200 Mbp of previously unresolved sequences, many of which are located in centromeric, acrocentric, and telomeric regions, and includes additional protein-coding genes [298]. Building on this, the creation of human reference pangenomes that represent diverse populations through genome graphs offers a powerful means to better characterise repetitive regions and complex SVs [299–302]. The Human Pangenome Reference Consortium has produced a draft comprising phased diploid assemblies with substantial euchromatic sequences and SVs that are absent from GRCh38 [303]. Tools such as GraphTyper [304, 305], which leverage pangenome graphs for population-scale genotyping of SV, exemplify how the integration of genome graph representations into large-scale studies can overcome reference bias, enhance read alignment, and facilitate accurate SV genotyping across diverse populations. However, transitioning existing annotations and benchmarks to graph-based references presents challenges, particularly in adapting gene definitions and regulatory annotations to this new framework [2].

In parallel, LR-seq has emerged as a transformative technology for SV discovery and CNV analysis. A population-scale LR-seq study by Beyter et al. [5] identified over 22,000 SVs per individual, three to five times more than those found using short-read

sequencing, highlighting the substantial gains in sensitivity, especially for tandem repeat-associated SVs. Moreover, the integration of LR-seq-derived SVs with imputation into large cohorts has enabled functional insights, such as the identification of a rare deletion in *PCSK9* associated with reduced LDL cholesterol levels and a multiallelic repeat in *ACAN* linked to height [5]. More recently, Bai et al. [306] constructed a reference panel using 482 haplotype-resolved long-read assemblies. They developed an online imputation tool, ImputeSV, designed to impute SVs and tandem repeat variants from SNP data. This tool was applied to impute 54,578 common SVs in 456,643 participants from the UK Biobank. The study identified 17,335 SV-trait associations across 2,624 traits and estimated that SVs contributed to at least 4.7% of the common genetic variance associated with complex traits. These findings underscore the importance of LR-seq as a critical complement to pangenome approaches, offering comprehensive SV catalogues that better capture population diversity and disease-relevant variations.

Future directions also emphasise the integration of CNV and SNP association studies within a unified analytical framework. Many CNVs are poorly tagged by SNPs or occur on distinct haplotypes, underscoring the need for CNV-specific analysis. Approaches such as CNest have facilitated the classification of CNV associations based on their overlap with SNP signals, distinguishing between CNV-only, CNV-allele, SNP-CNV near, and SNP-CNV far associations. Joint modelling of SNP and CNV associations is anticipated to reveal shared genetic mechanisms, enhance the prioritisation of causal genes, and refine polygenic risk score (PGS) predictions, contingent upon the establishment of LD maps between SNPs and CNVs [27, 66, 307].

Future CNV research will increasingly adopt multimodal integration strategies that combine CNV data with other molecular layers, such as transcriptomic, epigenomic, and chromatin topology, to better understand the functional impact of SVs/CNVs. Recent studies have highlighted how rare germline SVs can disrupt highly expressed and mutationally constrained genes, alter regulatory domains such as topologically associating domains (TADs), and dysregulate gene expression in tissue-specific contexts [3, 28]. For example, the integration of CNV data with RNA sequencing and epigenetic profiles from disease-relevant tissues has demonstrated that singleton gene-disruptive germline CNVs preferentially impact genes expressed in the tissue of origin, potentially leading to downstream transcriptional dysregulation observable in paediatric tumours. Furthermore, non-coding CNVs that overlap with tissue-specific chromatin boundaries are associated with disruptions in three-dimensional genome architecture, linking CNVs to alterations in the regulatory networks [308].

Benchmarking and evaluation of CNV detection remain challenging, particularly in repetitive genomic regions. Emerging tools, such as TT-Mars, which compare SV calls to haplotype-resolved assemblies, are helping to address benchmarking challenges in repetitive regions [214]. Additionally, the nf-core/variantbenchmarking pipeline streamlines evaluation by integrating diverse truth sets and supporting comparisons across assemblies, including CNV-specific benchmarks [309].

Beyond benchmarking, there is a critical need to standardise reporting guidelines and metadata definitions to ensure reproducibility and interoperability, ultimately facilitating the inclusion of CNV findings in public resources, such as the GWAS Catalogue. In this context, causal inference methods, such as transcriptome-wide Mendelian

randomisation, are particularly valuable, as they provide robust frameworks for linking CNVs to functional gene expression changes and phenotypic effects, thereby strengthening the evidence required for CNV association results to be formally integrated into catalogues and downstream analyses, such as PGS and drug target prioritisation [27, 82, 310].

Innovative workflows and haplotype-aware methods have transformed the large-scale discovery and analysis of CNVs. For example, the CNest workflow enables high-resolution CNV detection from NGS read depth across very large cohorts and has already uncovered hundreds of novel associations in the UK Biobank [66]. Crucially, CNest adheres to the Global Alliance for Genomics and Health (GA4GH) standards, ensuring interoperability and standardised data exchange in population-scale studies [311]. Complementing this, haplotype-informed approaches, such as HI-CNV, substantially boost detection sensitivity by leveraging shared haplotypes in biobank-scale datasets, identifying more than six times as many CNVs per individual as earlier methods [20]. Finally, advances in statistical phasing tools, such as SHAPEIT5, provide the haplotype scaffolding necessary for CNV analyses, enabling the exploration of phase-dependent allelic series and their downstream phenotypic consequences [312].

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-026-04189-6>.

Additional file 1: Figs. S1-S8 and Tables S2-S3. Document containing eight figures and two tables. Fig. S1 shows trends in CNV association studies over time. Figs. S2-S7 present citation landscapes for array-based, exome sequencing-based, genome sequencing-based, targeted panel-based, long-read sequencing-based, and ensemble-based CNV detection tools, respectively, while Fig. S8 presents the citation landscape for CNV association tools. Table S2 summarises common CNV callers, including their data inputs, calling approach, and associated repositories. Table S3 summarises common CNV association tools, including their input data types, study designs, statistical methods, strengths, and limitations.

Additional file 2: Table S1. An Excel spreadsheet listing 131 published copy number variation association studies, including authors, publication year, study title, sequencing/array platform used, journal, and citation count.

Additional file 3: Notes S1-S6. The document provides a detailed background on CNV detection platforms and methodologies (Note S1), an overview of detection strategies across array- and sequencing-based platforms (Note S2), segmentation algorithms used in CNV calling (Note S3), platform-specific benchmarking comparisons (Note S4), challenges in sex chromosome CNV analysis (Note S5), and a tool-by-tool review of CNV association software (Note S6).

Acknowledgements

Not applicable.

Peer review information

Alison Monroe was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions

A. H. S., H.S., and M.A. contributed to the conceptualisation, literature review, and initial drafting of the manuscript. H.V., L.Y., T.S., and H.D. contributed to the critical review of the literature and writing. X.L., N.M.O., and X.Z. edited and refined the manuscript. J.G. and H.H. supervised the overall study. All authors have read and approved the final version of the manuscript.

Funding

This work was funded in part by Institutional Development Funds from The Children's Hospital of Philadelphia to the Center for Applied Genomics and by the Children's Hospital of Philadelphia Endowed Chair in Genomic Research to H.H.

Data availability

This review did not include newly generated primary data. Fig. 2 was produced using publicly available data from Collins et al. [14, 313], specifically the annotated list of disease-associated CNV loci provided in Table S3 of that study. Additional file 1: Fig. S1 was generated using a published compilation of CNV GWAS studies provided in Supplementary Information 2 of Harris

et al. [8, 314] and supplemented with additional studies curated by the authors. The combined list of CNV association studies is shown in Additional file 1: Fig. S1, with detailed study characteristics provided in Additional file 2: Table S1.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. ²Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, USA. ³Department of Medical Genetics, Faculty of Medicine, Tabriz University of Medical Sciences, Tabriz, Iran. ⁴Cellular and Molecular Endocrine Research Center, Research Institute for Endocrine Molecular Biology, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. ⁵Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ⁶Division of Genetic and Genomic Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, USA. ⁷Department of Dermatology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ⁸Department of Pathology, Cytogenetics Laboratory, City of Hope National Medical Center, Duarte, CA, USA. ⁹Department of Medicinal Chemistry, Faculty of Pharmacy, Tehran University of Medical Sciences, Tehran, Iran. ¹⁰Division of Pulmonary Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, USA. ¹¹Faculty of Medicine, University of Iceland, Reykjavik, Iceland.

Received: 27 November 2025 Accepted: 2 July 2026

Published online: 10 July 2026

References

- Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74.
- Collins RL, Talkowski ME. Diversity and consequences of structural variation in the human genome. *Nat Rev Genet*. 2025;26(7):443–62. <https://doi.org/10.1038/s41576-024-00808-9>.
- Spielmann M, Lupiáñez DG, Mundlos S. Structural variation in the 3D genome. *Nat Rev Genet*. 2018;19(7):453–67.
- Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet*. 2016;17(4):224–38.
- Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Björnsson E, Jonsson H, et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet*. 2021;53(6):779–86.
- Zarrei M, MacDonald J, Merico D, Scherer S. A copy number variation map of the human genome. *Nat Rev Genet*. 2015;16:172–83.
- Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature*. 2020;583(7814):83–9.
- Harris L, McDonagh EM, Zhang X, Fawcett K, Foreman A, Daneck P, et al. Genome-wide association testing beyond SNPs. *Nat Rev Genet*. 2025;26(3):156–70.
- Hurles ME, Dermitzakis ET, Tyler-Smith C. The functional impact of structural variation in humans. *Trends Genet*. 2008;24(5):238–45.
- Jakubosky D, D'Antonio M, Bonder MJ, Smail C, Donovan MKR, Young Greenwald WW, et al. Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat Commun*. 2020;11(1):2927.
- Scott AJ, Chiang C, Hall IM. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res*. 2021;31(12):2249–57.
- Gunning AC, Strucinska K, Muñoz Oreja M, Parrish A, Caswell R, Stals KL, et al. Recurrent *de novo* NAHR reciprocal duplications in the *ATAD3* gene cluster cause a neurogenetic trait with perturbed cholesterol and mitochondrial metabolism. *Am J Hum Genet*. 2020;106(2):272–9.
- Harel T, Lupski JR. Genomic disorders 20 years on—mechanisms for clinical manifestations. *Clin Genet*. 2018;93(3):439–49.
- Collins RL, Glessner JT, Porcu E, Lepamets M, Brandon R, Lauricella C, et al. A cross-disorder dosage sensitivity map of the human genome. *Cell*. 2022;185(16):3041–55.e25.
- Smolen C, Girirajan S. The gene dose makes the disease. *Cell*. 2022;185(16):2850–2.
- Girirajan S, Rosenfeld JA, Coe BP, Parikh S, Friedman N, Goldstein A, et al. Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *N Engl J Med*. 2012;367(14):1321–31.
- Albers CA, Paul DS, Schulze H, Freson K, Stephens JC, Smethurst PA, et al. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nat Genet*. 2012;44(4):435–9, s1–2.
- Aguirre M, Rivas MA, Priest J. Phenome-wide burden of copy-number variation in the UK biobank. *Am J Hum Genet*. 2019;105(2):373–83.

19. Macé A, Tuke MA, Deelen P, Kristiansson K, Mattsson H, Nöukas M, et al. CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. *Nat Commun.* 2017;8(1):744.
20. Hujoel MLA, Sherman MA, Barton AR, Mukamel RE, Sankaran VG, Terao C, et al. Influences of rare copy-number variation on human complex traits. *Cell.* 2022;185(22):4233–48.e27.
21. Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet.* 2017;49(1):27–35.
22. Trost B, Thiruvahindrapuram B, Chan AJS, Engchuan W, Higginbotham EJ, Howe JL, et al. Genomic architecture of autism from comprehensive whole-genome sequence annotation. *Cell.* 2022;185(23):4409–27.e18.
23. Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. *Bioinformatics.* 2014;30(19):2811–2.
24. Fawcett KA, Demidov G, Shrine N, Paynton ML, Ossowski S, Sayers I, et al. Exome-wide analysis of copy number variation shows association of the human leukocyte antigen region with asthma in UK Biobank. *BMC Med Genom.* 2022;15(1):119.
25. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562(7726):203–9.
26. Auwerx C, Jöeloo M, Sadler MC, Tesio N, Ojavee S, Clark CJ, et al. Rare copy-number variants as modulators of common disease susceptibility. *Genome Med.* 2024;16(1):5.
27. Auwerx C, Lepamets M, Sadler MC, Patxot M, Stojanov M, Baud D, et al. The individual and global impact of copy-number variants on complex human traits. *Am J Hum Genet.* 2022;109(4):647–68.
28. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet.* 2020;21(3):171–89.
29. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. *Genome Biol.* 2019;20(1):246.
30. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the major structural variant alleles of the human genome. *Cell.* 2019;176(3):663–75.e19.
31. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science.* 2021;372(6537):eabf7117.
32. Logsdon GA, Ebert P, Audano PA, Loftus M, Porubsky D, Ebler J, et al. Complex genetic variation in nearly complete human genomes. *Nature.* 2025;644(8076):430–41.
33. Gabriellaite M, Torp MH, Rasmussen MS, Andreu-Sánchez S, Vieira FG, Pedersen CB, et al. A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers (Basel).* 2021;13(24):6283.
34. van Baardwijk MN, Heijnen L, Zhao H, Baudis M, Stubbs AP. A systematic benchmark of copy number variation detection tools for high density SNP genotyping arrays. *Genomics.* 2024;116(6):110962.
35. Whitford W, Lehnert K, Snell RG, Jacobsen JC. Evaluation of the performance of copy number variant prediction tools for the detection of deletions from whole genome sequencing data. *J Biomed Inform.* 2019;94:103174.
36. Yao R, Zhang C, Yu T, Li N, Hu X, Wang X, et al. Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Mol Cytogenet.* 2017;10:30.
37. De La Vega FM, Irvine SA, Anur P, Potts K, Kraft L, Torres R, et al. Benchmarking of germline copy number variant callers from whole genome sequencing data for clinical applications. *Bioinformatics Adv.* 2025;5(1):vbaf071.
38. Munté E, Roca C, Del Valle J, Feliubadaló L, Pineda M, Gel B, et al. Detection of germline CNVs from gene panel data: benchmarking the state of the art. *Brief Bioinform.* 2024;26(1):bbae645.
39. Yuan N, Jia P. Comprehensive assessment of long-read sequencing platforms and calling algorithms for detection of copy number variation. *Brief Bioinform.* 2024;25(5):bbae441.
40. Song M, Ma S, Wang G, Wang Y, Yang Z, Xie B, et al. Benchmarking copy number aberrations inference tools using single-cell multi-omics datasets. *Brief Bioinform.* 2025;26(2):bbaf076.
41. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* 2019;20(1):117.
42. Majidian S, Agostinho DP, Chin CS, Sedlazeck FJ, Mahmoud M. Genomic variant benchmark: if you cannot measure it, you cannot improve it. *Genome Biol.* 2023;24(1):221.
43. Coutelier M, Holtgrewe M, Jäger M, Flöttman R, Mensah MA, Spielmann M, et al. Combining callers improves the detection of copy number variants from whole-genome sequencing. *Eur J Hum Genet.* 2022;30(2):178–86.
44. Trost B, Walker S, Wang Z, Thiruvahindrapuram B, MacDonald JR, Sung WWL, et al. A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. *Am J Hum Genet.* 2018;102(1):142–55.
45. Gordeeva V, Sharova E, Babalyan K, Sultanov R, Govorun VM, Arapidi G. Benchmarking germline CNV calling tools from exome sequencing data. *Sci Rep.* 2021;11(1):14416.
46. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.
47. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559–75.
48. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet.* 2021;53(7):1097–103.
49. Glessner JT, Li J, Liu Y, Khan M, Chang X, Sleiman PMA, et al. ParseCNV2: efficient sequencing tool for copy number variation genome-wide association studies. *Eur J Hum Genet.* 2023;31(3):304–12.
50. Kim JH, Hu HJ, Yim SH, Bae JS, Kim SY, Chung YJ. CNVRuler: a copy number variation-based case-control association analysis tool. *Bioinformatics.* 2012;28(13):1790–2.
51. Forer L, Schönherr S, Weissensteiner H, Haider F, Kluckner T, Gieger C, et al. CONAN: copy number variation analysis software for genome-wide association studies. *BMC Bioinformatics.* 2010;11(1):318.
52. Glessner JT, Li J, Hakonarson H. ParseCNV integrative copy number variation association software with quality tracking. *Nucleic Acids Res.* 2013;41(5):e64.

53. Chen X, Fang LT, Chen Z, et al. A benchmarking study of copy number variation inference methods using single-cell RNA-sequencing data. *Precis Clin Med*. 2025;8(2):pbaf011.
54. Mallory XF, Edrisi M, Navin N, Nakhleh L. Assessing the performance of methods for copy number aberration detection from single-cell DNA sequencing data. *PLoS Comput Biol*. 2020;16(7):e1008012.
55. Schmid KT, Symeonidi A, Hlushchenko D, Richter ML, Colomé-Tatché M. Benchmarking scRNA-seq copy number variation callers. *bioRxiv*. 2024:2024.12.18.629083.
56. De Falco A, Caruso F, Su X-D, Iavarone A, Ceccarelli M. A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. *Nat Commun*. 2023;14(1):1074.
57. Fan J, Lee HO, Lee S, Ryu DE, Lee S, Xue C, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res*. 2018;28(8):1217–27.
58. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat Biotechnol*. 2021;39(5):599–608.
59. Gao T, Soldatov R, Sarkar H, Kurkiewicz A, Biederstedt E, Loh PR, et al. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. *Nat Biotechnol*. 2023;41(3):417–26.
60. Harmanci AS, Harmanci A, Zhou X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nat Commun*. 2020;11:89.
61. Müller S, Cho A, Liu SJ, Lim DA, Diaz A. CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. *Bioinformatics*. 2018;34(18):3217–9.
62. Shao X, Lv N, Liao J, Long J, Xue R, Ai N, et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Med Genet*. 2019;20(1):175.
63. Yi K, Ju YS. Patterns and mechanisms of structural variations in human cancer. *Exp Mol Med*. 2018;50(8):1–11.
64. Maury EA, Sherman MA, Genovese G, Gilgenast TG, Kamath T, Burris SJ, et al. Schizophrenia-associated somatic copy-number variants from 12,834 cases reveal recurrent NRXN1 and ABCB11 disruptions. *Cell Genom*. 2023;3(8):100356.
65. McConnell MJ, Lindberg MR, Brennand KJ, Piper JC, Voet T, Cowing-Zitron C, et al. Mosaic copy number variation in human neurons. *Science*. 2013;342(6158):632–7.
66. Fitzgerald T, Birney E. CNest: a novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank. *Cell Genom*. 2022;2(8):100167.
67. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol*. 2011;29(6):512–20.
68. Wiszniewska J, Bi W, Shaw C, Stankiewicz P, Kang SHL, Pursley AN, et al. Combined array CGH plus SNP genome analyses in a single assay for optimized clinical testing. *Eur J Hum Genet*. 2014;22(1):79–87.
69. Verlouw JAM, Clemens E, de Vries JH, Zolk O, Verkerk AJMH, am Zehnhoff-Dinnesen A, et al. A comparison of genotyping arrays. *Eur J Hum Genet*. 2021;29(11):1611–24.
70. Li YR, Glessner JT, Coe BP, Li J, Mohebbnasab M, Chang X, et al. Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. *Nat Commun*. 2020;11(1):255.
71. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*. 2008;40(10):1253–60.
72. Zhang F, Gu W, Hurler ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Ann Rev Genom Hum Genet*. 2009;10(Volume 10, 2009):451–81.
73. Fu W, Zhang F, Wang Y, Gu X, Jin L. Identification of copy number variation hotspots in human populations. *Am J Hum Genet*. 2010;87(4):494–504.
74. Brandler WM, Antaki D, Gujral M, Noor A, Rosanio G, Chapman TR, et al. Frequency and complexity of de novo structural mutation in autism. *Am J Hum Genet*. 2016;98(4):667–79.
75. Belyeu JR, Brand H, Wang H, Zhao X, Pedersen BS, Feusier J, et al. De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am J Hum Genet*. 2021;108(4):597–607.
76. Null M, Yilmaz F, Astling D, Yu HC, Cole JB, Hallgrímsson B, et al. Genome-wide analysis of copy number variants and normal facial variation in a large cohort of Bantu Africans. *HGG Adv*. 2022;3(1):100082.
77. Craddock N, Hurler ME, Cardin N, Pearson RD, Plagnol V, Robson S, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*. 2010;464(7289):713–20.
78. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. *Nature*. 2020;581(7809):444–51.
79. Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. *Trends Genet*. 2013;29(10):575–84.
80. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large multiallelic copy number variations in humans. *Nat Genet*. 2015;47(3):296–303.
81. Jakubosky D, Smith EN, D'Antonio M, Jan Bonder M, Young Greenwald WW, D'Antonio-Chronowska A, et al. Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats. *Nat Commun*. 2020;11(1):2928.
82. Zamariolli M, Auwerx C, Sadler MC, van der Graaf A, Lepik K, Schoeler T, et al. The impact of 22q11.2 copy-number variants on human traits in the general population. *Am J Hum Genet*. 2023;110(2):300–13.
83. Hanssen R, Auwerx C, Jöeloo M, Sadler MC, Henning E, Keogh J, et al. Chromosomal deletions on 16p11.2 encompassing SH2B1 are associated with accelerated metabolic disease. *Cell Rep Med*. 2023;4(8):101155.
84. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet*. 2013;132(10):1077–130.
85. Crawford K, Bracher-Smith M, Owen D, Kendall KM, Rees E, Pardiñas AF, et al. Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *J Med Genet*. 2019;56(3):131–8.
86. Rodríguez-López J, Flórez G, Blanco V, Pereiro C, Fernández JM, Fariñas E, et al. Genome wide analysis of rare copy number variations in alcohol abuse or dependence. *J Psychiatr Res*. 2018;103:212–8.
87. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, et al. A robust statistical method for case-control association testing with copy number variation. *Nat Genet*. 2008;40(10):1245–52.

88. da Silva V, Ramos M, Groenen M, Crooijmans R, Johansson A, Regitano L, et al. CNVRanger: association analysis of CNVs with gene expression and quantitative phenotypes. *Bioinformatics*. 2020;36(3):972–3.
89. Saudi Mendeliome Group. Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases. *Genome Biol*. 2015;16(1):134.
90. Ashley EA. Towards precision medicine. *Nat Rev Genet*. 2016;17(9):507–22.
91. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*. 2015;517(7536):608–11.
92. Goldfeder RL, Priest JR, Zook JM, Grove ME, Waggott D, Wheeler MT, et al. Medical implications of technical accuracy in genome sequencing. *Genome Med*. 2016;8(1):24.
93. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333–51.
94. Gordeeva V, Sharova E, Arapidi G. Progress in methods for copy number variation profiling. *Int J Mol Sci*. 2022;23(4):2143.
95. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res*. 2017;27(5):677–85.
96. lafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. *Nat Genet*. 2004;36(9):949–51.
97. Klein CJ, Foroud TM. Neurology individualized medicine: when to use next-generation sequencing panels. *Mayo Clin Proc*. 2017;92(2):292–305.
98. Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med*. 2018;20(1):159–63.
99. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet*. 2010;86(5):749–64.
100. Schaefer GB, Mendelsohn NJ. Clinical genetics evaluation in identifying the etiology of autism spectrum disorders: 2013 guideline revisions. *Genet Med*. 2013;15(5):399–407.
101. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. *Science*. 2004;305(5683):525–8.
102. Song JHT, Lowe CB, Kingsley DM. Characterization of a human-specific tandem repeat associated with bipolar disorder and schizophrenia. *Am J Hum Genet*. 2018;103(3):421–30.
103. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinform*. 2013;14(11):S1.
104. Ahsan MU, Liu Q, Perdomo JE, Fang L, Wang K. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nat Methods*. 2023;20(8):1143–58.
105. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, et al. QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*. 2007;35(6):2013–25.
106. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun*. 2017;8(1):1326.
107. Darvishi K. Application of Nexus copy number software for CNV detection and analysis. *Curr Protoc Hum Genet*. 2010;Chapter 4:Unit 4.14.1–28.
108. Ding H, Luo J. MAMnet: detecting and genotyping deletions and insertions based on long reads and a deep learning approach. *Brief Bioinform*. 2022;23(5):bbac195.
109. English AC, Salerno WJ, Reid JG. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*. 2014;15(1):180.
110. Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol*. 2019;20(1):277.
111. Gong L, Wong CH, Cheng WC, Tjong H, Menghi F, Ngan CY, et al. Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nat Methods*. 2018;15(6):455–60.
112. Heller D, Vingron M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics*. 2020;36(22–23):5519–21.
113. Jeng XJ, Cai TT, Li H. Optimal sparse segment identification with application in copy number variation analysis. *J Am Stat Assoc*. 2010;105(491):1156–66.
114. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol*. 2020;21(1):189.
115. Killick R, Fearnhead P, Eckley IA. Optimal detection of changepoints with a linear computational cost. *J Am Stat Assoc*. 2012;107(500):1590–8.
116. Lin J, Wang S, Audano PA, Meng D, Flores JI, Kusters W, et al. SVision: a deep learning approach to resolve complex structural variants. *Nat Methods*. 2022;19(10):1230–3.
117. Luo J, Ding H, Shen J, Zhai H, Wu Z, Yan C, et al. BreakNet: detecting deletions using long reads and a deep learning approach. *BMC Bioinform*. 2021;22(1):577.
118. Magi A, Tattini L, Pippucci T, Torricelli F, Benelli M. Read count approach for DNA copy number variants detection. *Bioinformatics*. 2012;28(4):470–8.
119. Niu YS, Zhang H. The screening and ranking algorithm to detect dna copy number variations. *Ann Appl Stat*. 2012;6(3):1306–26.
120. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5(4):557–72.
121. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*. 2010;466(7304):368–72.
122. Pique-Regi R, Cáceres A, González JR. R-Gada: a fast and flexible pipeline for copy number analysis in association studies. *BMC Bioinform*. 2010;11:380.

123. Pirooznia M, Goes FS, Zandi PP. Whole-genome CNV analysis: advances in computational approaches. *Front Genet.* 2015;6:138.
124. Roy S, Motsinger RA. Evaluation of calling algorithms for array-CGH. *Front Genet.* 2013;4:217.
125. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods.* 2018;15(6):461–8.
126. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol.* 2024;42(10):1571–80.
127. Tham CY, Tirado-Magallanes R, Goh Y, Fullwood MJ, Koh BTH, Wang W, et al. NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome Biol.* 2020;21(1):56.
128. Tibshirani R, Wang P. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics.* 2008;9(1):18–29.
129. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* 2007;17(11):1665–74.
130. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol.* 2019;37(10):1155–62.
131. Zhang Z, Cheng H, Hong X, Di Narzo AF, Franzen O, Peng S, et al. EnsembleCNV: an ensemble machine learning algorithm to identify and genotype copy number variation using SNP array data. *Nucleic Acids Res.* 2019;47(7):e39.
132. Zhang Z, Lange K, Ophoff R, Sabatti C. Reconstructing dna copy number by penalized estimation and imputation. *Ann Appl Stat.* 2010;4(4):1749–73.
133. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics.* 2017;33(19):3088–90.
134. Lo Faro V, Ten Brink JB, Snieder H, Jansonius NM, Bergen AA. Genome-wide CNV investigation suggests a role for cadherin, Wnt, and p53 pathways in primary open-angle glaucoma. *BMC Genom.* 2021;22(1):590.
135. Hakkaart C, Pearson JF, Marquart L, Dennis J, Wiggins GAR, Barnes DR, et al. Copy number variants as modifiers of breast cancer risk for BRCA1/BRCA2 pathogenic variant carriers. *Commun Biol.* 2022;5(1):1061.
136. Dennis J, Tyrer JP, Walker LC, Michailidou K, Dorling L, Bolla MK, et al. Rare germline copy number variants (CNVs) and breast cancer risk. *Commun Biol.* 2022;5(1):65.
137. Kendall KM, Rees E, Escott-Price V, Einon M, Thomas R, Hewitt J, et al. Cognitive performance among carriers of pathogenic copy number variants: analysis of 152,000 UK biobank subjects. *Biol Psychiatry.* 2017;82(2):103–10.
138. Kikuchi M, Kobayashi K, Nishida N, Sawai H, Sugiyama M, Mizokami M, et al. Genome-wide copy number variation analysis of hepatitis B infection in a Japanese population. *Hum Genome Var.* 2021;8(1):22.
139. Li Z, Chen J, Xu Y, Yi Q, Ji W, Wang P, et al. Genome-wide analysis of the role of copy number variation in schizophrenia risk in Chinese. *Biol Psychiatry.* 2016;80(4):331–7.
140. Tansey KE, Rees E, Linden DE, Ripke S, Chambert KD, Moran JL, et al. Common alleles contribute to schizophrenia in CNV carriers. *Mol Psychiatry.* 2016;21(8):1085–9.
141. Yuan J, Hu J, Li Z, Zhang F, Zhou D, Jin C. A replication study of schizophrenia-related rare copy number variations in a Han Southern Chinese population. *Hereditas.* 2017;154(1):2.
142. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 2011;21(6):974–84.
143. Anjum S, Morganella S, D'Angelo F, Iavarone A, Ceccarelli M. VEGAWES: variational segmentation on whole exome sequencing for copy number detection. *BMC Bioinform.* 2015;16:315.
144. Cabello-Aguilar S, Vendrell JA, Van Goethem C, Brousse M, Gozè C, Frantz L, et al. IfCNV: a novel isolation-forest-based package to detect copy-number variations from various targeted NGS datasets. *Mol Ther Nucleic Acids.* 2022;30:174–83.
145. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet.* 2012;91(4):597–607.
146. Magi A, Benelli M, Marseglia G, Nannetti G, Scordo MR, Torricelli F. A shifting level model algorithm that identifies aberrations in array-CGH data. *Biostatistics.* 2010;11(2):265–80.
147. Magi A, Benelli M, Yoon S, Roviello F, Torricelli F. Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic Acids Res.* 2011;39(10):e65.
148. Magi A, Pippucci T, Sidore C. XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments. *BMC Genomics.* 2017;18(1):747.
149. Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, et al. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol.* 2013;14(10):R120.
150. Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS ONE.* 2011;6(1):e16327.
151. Morganella S, Cerulo L, Viglietto G, Ceccarelli M. VEGA: variational segmentation for copy number detection. *Bioinformatics.* 2010;26(24):3020–7.
152. Mumford D, Shah J. Optimal approximations by piecewise smooth functions and associated variational problems. *Commun Pure Appl Math.* 1989;42(5):577–685.
153. Vardhanabhuti S, Jeng XJ, Wu Y, Li H. Parametric modeling of whole-genome sequencing data for CNV identification. *Biostatistics.* 2014;15(3):427–41.
154. Wang LY, Abyzov A, Korb J, Snyder M, Gerstein M. MSB: a mean-shift-based approach for the analysis of structural variation in the genome. *Genome Res.* 2009;19(1):106–17.
155. Yuan X, Yu J, Xi J, Yang L, Shang J, Li Z, et al. CNV_JFTV: an isolation forest and total variation-based detection of CNVs from short-read sequencing data. *IEEE ACM Trans Comput Biol Bioinform.* 2021;18(2):539–49.
156. Zhang Y, Liu W, Duan J. On the core segmentation algorithms of copy number variation detection tools. *Brief Bioinform.* 2024;25(2):bbae022.

157. Abyzov A, Li S, Kim DR, Mohiyuddin M, Stütz AM, Parrish NF, et al. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat Commun.* 2015;6(1):7256.
158. Babadi M, Fu JM, Lee SK, Smirnov AN, Gauthier LD, Walker M, et al. GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nat Genet.* 2023;55(9):1589–97.
159. Backenroth D, Homsy J, Murillo LR, Glessner J, Lin E, Brueckner M, et al. CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res.* 2014;42(12):e97.
160. Bartenhagen C, Dugas M. Robust and exact structural variation detection with paired-end and soft-clipped alignments: SoftSV compared with eight algorithms. *Brief Bioinform.* 2015;17(1):51–62.
161. Becker T, Lee W-P, Leone J, Zhu Q, Zhang C, Liu S, et al. FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biol.* 2018;19(1):38.
162. Behera S, Catreux S, Rossi M, Truong S, Huang Z, Ruehle M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol.* 2025;43(7):1177–91.
163. Boeva V, Popova T, Bleakley K, Chiche P, Cappel J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics.* 2011;28(3):423–5.
164. Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res.* 2017;27(12):2050–60.
165. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods.* 2009;6(9):677–81.
166. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics.* 2015;32(8):1220–2.
167. Chiang T, Liu X, Wu TJ, Hu J, Sedlazeck FJ, White S, et al. Atlas-CNV: a validated approach to call single-exon CNVs in the eMERGESeq gene panel. *Genet Med.* 2019;21(9):2135–44.
168. D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A. Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. *Nucleic Acids Res.* 2016;44(20):e154.
169. Demidov G, Sturm M, Ossowski S. ClinCNV: multi-sample germline CNV detection in NGS data. *bioRxiv.* 2022:2022.06.10.495642.
170. Dharanipragada P, Voleti S, Parekh N. iCopyDAV: Integrated platform for copy number variations-Detection, annotation and visualization. *PLoS ONE.* 2018;13(4):e0195334.
171. Dierckxsens N, Li T, Vermeesch JR, Xie Z. A benchmark of structural variation detection by long reads through a realistic simulated model. *Genome Biol.* 2021;22(1):342.
172. Eisfeldt J, Vezzi F, Olason P, Nilsson D, Lindstrand A. TIDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. *F1000Res.* 2017;6:664.
173. Fang L, Hu J, Wang D, Wang K. NextSV: a meta-caller for structural variants from low-coverage long-read sequencing data. *BMC Bioinformatics.* 2018;19(1):180.
174. Fowler A. DECoN: a detection and visualization tool for exonic copy number variants. *Methods Mol Biol.* 2022;2493:77–88.
175. Gambin T, Akdemir ZC, Yuan B, Gu S, Chiang T, Carvalho CMB, et al. Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. *Nucleic Acids Res.* 2017;45(4):1633–48.
176. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun.* 2017;8:14061.
177. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res.* 2015;43(6):e39.
178. Jiang Y, Wang R, Urrutia E, Anastopoulos IN, Nathanson KL, Zhang NR. CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biol.* 2018;19(1):202.
179. Johansson LF, van Dijk F, de Boer EN, van Dijk-Bos KK, Jongbloed JD, van der Hout AH, et al. CoNVaDING: single exon variation detection in targeted NGS data. *Hum Mutat.* 2016;37(5):457–64.
180. Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, et al. Cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* 2012;40(9):e69.
181. Kronenberg ZN, Osborne EJ, Cone KR, Kennedy BJ, Domyan ET, Shapiro MD, et al. Wham: identifying structural variants of biological consequence. *PLoS Comput Biol.* 2015;11(12):e1004572.
182. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, et al. Copy number variation detection and genotyping from exome sequence data. *Genome Res.* 2012;22(8):1525–32.
183. Lam HY, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, et al. Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat Biotechnol.* 2012;30(3):226–9.
184. Lam HYK, Mu XJ, Stütz AM, Tanzer A, Cayting PD, Snyder M, et al. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol.* 2010;28(1):47–55.
185. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* 2014;15(6):R84.
186. Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, et al. Contra: copy number analysis for targeted resequencing. *Bioinformatics.* 2012;28(10):1307–13.
187. Love MI, Myšičková A, Sun R, Kalscheuer V, Vingron M, Haas SA. Modeling read counts for CNV detection in exome sequencing data. *Stat Appl Genet Mol Biol.* 2011;10(1):Article 52.
188. Michaelson JJ, Sebat J. forestSV: structural variant discovery through statistical learning. *Nat Methods.* 2012;9(8):819–21.
189. Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, et al. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics.* 2015;31(16):2741–4.
190. Packer JS, Maxwell EK, O'Dushlaine C, Lopez AE, Dewey FE, Chernomorsky R, et al. CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics.* 2016;32(1):133–5.

191. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*. 2012;28(21):2747–54.
192. Popic V, Rohlicek C, Cunial F, Hajirasouliha I, Meleshko D, Garimella K, et al. Cue: a deep-learning framework for structural variant discovery and genotyping. *Nat Methods*. 2023;20(4):559–68.
193. Pounraja VK, Jayakar G, Jensen M, Kelkar N, Girirajan S. A machine-learning approach for accurate detection of copy number variants from exome sequencing. *Genome Res*. 2019;29(7):1134–43.
194. Povysil G, Tzika A, Vogt J, Haunschmid V, Messiaen L, Zschocke J, et al. Panelcn.MOPS: copy-number detection in targeted NGS panel data for clinical diagnostics. *Hum Mutat*. 2017;38(7):889–97.
195. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;28(18):i333–9.
196. Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics*. 2016;32(15):2375–7.
197. Samarakoon PS, Sorte HS, Kristiansen BE, Skodje T, Sheng Y, Tjønnfjord GE, et al. Identification of copy number variants from exome sequence data. *BMC Genomics*. 2014;15(1):661.
198. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*. 2011;27(19):2648–54.
199. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput Biol*. 2016;12(4):e1004873.
200. Wang C, Evans JM, Bhagwate AV, Prodduturi N, Sarangi V, Middha M, et al. PatternCNV: a versatile tool for detecting copy number changes from exome sequencing data. *Bioinformatics*. 2014;30(18):2678–80.
201. Wong K, Keane TM, Stalker J, Adams DJ. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol*. 2010;11(12):R128.
202. Xi R, Lee S, Xia Y, Kim TM, Park PJ. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res*. 2016;44(13):6274–86.
203. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;25(21):2865–71.
204. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res*. 2009;19(9):1586–92.
205. Zarate S, Carroll A, Mahmoud M, Krashenina O, Jun G, Salerno WJ, et al. Parliament2: accurate structural variant calling at scale. *Gigascience*. 2020;9(12):giaa145.
206. Zhang J, Wang J, Wu Y. An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. *BMC Bioinform*. 2012;13 Suppl 6(Suppl 6):S6.
207. Zhu M, Need AC, Han Y, Ge D, Maia JM, Zhu Q, et al. Using ERDS to infer copy-number variants in high-coverage genomes. *Am J Hum Genet*. 2012;91(3):408–21.
208. Litmaps. Litmaps (Version 2025–01–16) [Visualization purposes]. 2024.
209. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, et al. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol*. 2020;38(11):1347–55.
210. Kronenberg Z, Nolan C, Porubsky D, Mokveld T, Rowell WJ, Lee S, et al. The Platinum Pedigree: a long-read benchmark for genetic variants. *Nat Methods*. 2025;22(8):1669–76.
211. Narci K, Vannieuwerkerke N, Garcia MU, Bot NC, EladH, Kirk J, et al. NF-core/variantbenchmarking: 1.2.0 doubtful Adams. Zenodo. 2025.
212. Hukku A, Pividori M, Luca F, Pique-Regi R, Im HK, Wen X. Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *Am J Hum Genet*. 2021;108(1):25–35.
213. English AC, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol*. 2022;23(1):271.
214. Yang J, Chaisson MJ. TT-Mars: structural variants assessment based on haplotype-resolved assemblies. *Genome Biol*. 2022;23(1):110.
215. Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol*. 2022;40(5):672–80.
216. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, et al. PrecisionFDA truth challenge V2: calling variants from short and long reads in difficult-to-map regions. *Cell Genom*. 2022;2(5):S2666-979X(22)00058-1.
217. Moreno-Cabrera JM, Del Valle J, Castellanos E, Feliubadaló L, Pineda M, Brunet J, et al. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *Eur J Hum Genet*. 2020;28(12):1645–55.
218. Parikh H, Mohiyuddin M, Lam HY, Iyer H, Chen D, Pratt M, et al. svclassify: a method to establish benchmark structural variant calls. *BMC Genom*. 2016;17:64.
219. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet*. 2007;39(7 Suppl):S16–21.
220. Lepkes L, Kayali M, Blümcke B, Weber J, Suszynska M, Schmidt S, et al. Performance of in silico prediction tools for the detection of germline copy number variations in cancer predisposition genes in 4208 female index patients with familial breast and ovarian cancer. *Cancers (Basel)*. 2021;13(1):118.
221. Nutsua ME, Fischer A, Nebel A, Hofmann S, Schreiber S, Krawczak M, et al. Family-based benchmarking of copy number variation detection software. *PLoS ONE*. 2015;10(7):e0133465.
222. Roca I, González-Castro L, Fernández H, Couce ML, Fernández-Marmiesse A. Free-access copy-number variant detection tools for targeted next-generation sequencing data. *Mutat Res Rev Mutat Res*. 2019;779:114–25.
223. Royer-Bertrand B, Cisarova K, Niel-Butschi F, Mittaz-Crettol L, Fodstad H, Superti-Furga A. CNV detection from exome sequencing data in routine diagnostics of rare genetic disorders: opportunities and limitations. *Genes (Basel)*. 2021;12(9):1427.

224. Silva C, Ferrão J, Marques B, Pedro S, Correia H, Valente A, et al. Comparative analysis of hybrid-SNP microarray and nanopore sequencing for detection of large-sized copy number variants in the human genome. *Mol Cytogenet.* 2025;18(1):18.
225. Smolander J, Khan S, Singaravelu K, Kauko L, Lund RJ, Laiho A, et al. Evaluation of tools for identifying large copy number variations from ultra-low-coverage whole-genome sequencing data. *BMC Genom.* 2021;22(1):357.
226. Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum Mutat.* 2014;35(7):899–907.
227. Zhang L, Bai W, Yuan N, Du Z. Comprehensively benchmarking applications for detecting copy number variation. *PLoS Comput Biol.* 2019;15(5):e1007069.
228. Zhao S, Xu D, Cai J, Shen Q, He M, Pan X, et al. Benchmarking strategies for CNV calling from whole genome bisulfite data in humans. *Comput Struct Biotechnol J.* 2025;27:912–9.
229. Zhou A, Lin T, Xing J. Evaluating nanopore sequencing data processing pipelines for structural variation identification. *Genome Biol.* 2019;20(1):237.
230. Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet.* 2011;43(3):269–76.
231. Montanucci L, Lewis-Smith D, Collins RL, Niestroj LM, Parthasarathy S, Xian J, et al. Genome-wide identification and phenotypic characterization of seizure-associated copy number variations in 741,075 individuals. *Nat Commun.* 2023;14(1):4392.
232. Saarentaus EC, Havulinna AS, Mars N, Ahola-Olli A, Kiiskinen TTJ, Partanen J, et al. Polygenic burden has broader impact on health, cognition, and socioeconomic outcomes than most rare and high-risk copy number variants. *Mol Psychiatry.* 2021;26(9):4884–95.
233. Warland A, Kendall KM, Rees E, Kirov G, Caseras X. Schizophrenia-associated genomic copy number variants and subcortical brain volumes in the UK Biobank. *Mol Psychiatry.* 2020;25(4):854–62.
234. Kendall KM, Bracher-Smith M, Fitzpatrick H, Lynham A, Rees E, Escott-Price V, et al. Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: analysis of the UK Biobank. *Br J Psychiatry.* 2019;214(5):297–304.
235. Lal D, Ruppert AK, Trucks H, Schulz H, de Kovel CG, Kasteleijn-Nolst Trenité D, et al. Burden analysis of rare micro-deletions suggests a strong impact of neurodevelopmental genes in genetic generalised epilepsies. *PLoS Genet.* 2015;11(5):e1005226.
236. Simpson NH, Ceroni F, Reader RH, Covill LE, Knight JC, Hennessy ER, et al. Genome-wide analysis identifies a role for common copy number variants in specific language impairment. *Eur J Hum Genet.* 2015;23(10):1370–7.
237. Sokolowski M, Wasserman J, Wasserman D. Rare CNVs in suicide attempt include schizophrenia-associated loci and neurodevelopmental genes: a pilot genome-wide and family-based study. *PLoS ONE.* 2016;11(12):e0168531.
238. Vevera J, Zarrei M, Hartmannová H, Jedličková I, Mušálková D, Přistoupilová A, et al. Rare copy number variation in extremely impulsively violent males. *Genes Brain Behav.* 2019;18(6):e12536.
239. Green EK, Rees E, Walters JT, Smith KG, Forty L, Grozeva D, et al. Copy number variation in bipolar disorder. *Mol Psychiatry.* 2016;21(1):89–93.
240. Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C, Aguirre M, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet.* 2021;53(2):185–94.
241. Leppa VM, Kravitz SN, Martin CL, Andrieux J, Le Caignec C, Martin-Coignard D, et al. Rare inherited and de novo CNVs reveal complex contributions to ASD risk in multiplex families. *Am J Hum Genet.* 2016;99(3):540–54.
242. Monlong J, Girard SL, Meloche C, Cadieux-Dion M, Andrade DM, Lafreniere RG, et al. Global characterization of copy number variants in epilepsy patients from whole genome sequencing. *PLoS Genet.* 2018;14(4):e1007285.
243. Chen J, Calhoun VD, Perrone-Bizzozero NI, Pearson GD, Sui J, Du Y, et al. A pilot study on commonality and specificity of copy number variants in schizophrenia and bipolar disorder. *Transl Psychiatry.* 2016;6(5):e824.
244. Chen L, Abel HJ, Das I, Larson DE, Ganel L, Kanchi KL, et al. Association of structural variation with cardiometabolic traits in Finns. *Am J Hum Genet.* 2021;108(4):583–96.
245. de Jesús A-M, Pinto D, Parra EJ, Valladares-Salgado A, Cruz M, Scherer SW. Characterization of large copy number variation in Mexican type 2 diabetes subjects. *Sci Rep.* 2017;7(1):17105.
246. Yin CL, Chen HI, Li LH, Chien YL, Liao HM, Chou MC, et al. Genome-wide analysis of copy number variations identifies PARK2 as a candidate gene for autism spectrum disorder. *Mol Autism.* 2016;7:23.
247. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011;89(1):82–93.
248. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet.* 2012;91(2):224–37.
249. Zhan X, Girirajan S, Zhao N, Wu MC, Ghosh D. A novel copy number variants kernel association test with application to autism spectrum disorders studies. *Bioinformatics.* 2016;32(23):3603–10.
250. Brucker A, Lu W, Marceau West R, Yu QY, Hsiao CK, Hsiao TH, et al. Association test using Copy Number Profile Curves (CONCUR) enhances power in rare copy number variant analysis. *PLoS Comput Biol.* 2020;16(5):e1007797.
251. Maus Esfahani N, Catchpoole D, Khan J, Kennedy PJ. MCKAT: a multi-dimensional copy number variant kernel association test. *BMC Bioinformatics.* 2021;22(1):588.
252. Maus Esfahani N, Catchpoole D, Kennedy PJ. SMCKAT, a sequential multi-dimensional CNV kernel-based association test. *Life (Basel).* 2021;11(12):1302.
253. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet.* 2010;42(4):348–54.
254. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet.* 2015;47(3):284–90.
255. Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. *Nat Genet.* 2018;50(7):906–8.

256. Ionita-Laza I, Perry GH, Raby BA, Klanderma B, Lee C, Laird NM, et al. On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. *Genet Epidemiol*. 2008;32(3):273–84.
257. Liu M, Moon S, Wang L, Kim S, Kim YJ, Hwang MY, et al. On the association analysis of CNV data: a fast and robust family-based association method. *BMC Bioinform*. 2017;18(1):217.
258. Wu X, Huai C, Shen L, Li M, Yang C, Zhang J, et al. Genome-wide study of copy number variation implicates multiple novel loci for schizophrenia risk in Han Chinese family trios. *iScience*. 2021;24(8):102894.
259. Mendes M, Chen DZ, Engchuan W, Leal TP, Thiruvahindrapuram B, Trost B, et al. Chromosome X-wide common variant association study in autism spectrum disorder. *Am J Hum Genet*. 2025;112(1):135–53.
260. Han J, Walters JT, Kirov G, Pocklington A, Escott-Price V, Owen MJ, et al. Gender differences in CNV burden do not confound schizophrenia CNV associations. *Sci Rep*. 2016;6:25986.
261. Subirana I, Diaz-Uriarte R, Lucas G, Gonzalez JR. CNVassoc: association analysis of CNV data using R. *BMC Med Genom*. 2011;4:47.
262. Wittig M, Helbig I, Schreiber S, Franke A. CNVneta: a data mining tool for large case-control copy number variation datasets. *Bioinformatics*. 2010;26(17):2208–9.
263. Prakash SK, Bondy CA, Maslen CL, Silberbach M, Lin AE, Perrone L, et al. Autosomal and X chromosome structural variants are associated with congenital heart defects in Turner syndrome: the NHLBI GenTAC registry. *Am J Med Genet A*. 2016;170(12):3157–64.
264. Wang H, Dombroski BA, Cheng PL, Tucci A, Si YQ, Farrell JJ, et al. Structural variation detection and association analysis of whole-genome-sequence data from 16,543 Alzheimer's disease sequencing project subjects. *Alzheimers Dement*. 2025;21(6):e70277.
265. Jeng J, Wu Q, Li H. A statistical method for identifying trait-associated copy number variants. *Hum Hered*. 2015;79(3–4):147–56.
266. Labani M, Afrasiabi A, Beheshti A, Lovell NH, Alinejad-Rokny H. PeakCNV: a multi-feature ranking algorithm-based tool for genome-wide copy number variation-association study. *Comput Struct Biotechnol J*. 2022;20:4975–83.
267. Alinejad-Rokny H, Heng JI, Forrest ARR. Brain-enriched coding and long non-coding RNA genes are overrepresented in recurrent neurodevelopmental disorder CNVs. *Cell Rep*. 2020;33(4):108307.
268. Larsen SJ, do Canto LM, Rogatto SR, Baumbach J. CoNVaQ: a web tool for copy number variation-based association studies. *BMC Genom*. 2018;19(1):369.
269. Abe-Hatano C, Iida A, Kosugi S, Momozawa Y, Terao C, Ishikawa K, et al. Whole genome sequencing of 45 Japanese patients with intellectual disability. *Am J Med Genet A*. 2021;185(5):1468–80.
270. Balagué-Dobón L, Cáceres A, González JR. Fully exploiting SNP arrays: a systematic review on the tools to extract underlying genomic structure. *Brief Bioinform*. 2022;23(2):bbac043.
271. Billingsley KJ, Ding J, Jerez PA, Illarionova A, Levine K, Grenn FP, et al. Genome-wide analysis of structural variants in Parkinson disease. *Ann Neurol*. 2023;93(5):1012–22.
272. de Almeida Santana MH, Junior GA, Cesar AS, Freua MC, da Costa GR, da Luz ESS, et al. Copy number variations and genome-wide associations reveal putative genes and metabolic pathways involved with the feed conversion ratio in beef cattle. *J Appl Genet*. 2016;57(4):495–504.
273. de Lemos MVA, Peripolli E, Berton MP, Feitosa FLB, Olivieri BF, Stafuzza NB, et al. Association study between copy number variation and beef fatty acid profile of Nellore cattle. *J Appl Genet*. 2018;59(2):203–23.
274. Fernandes AC, da Silva VH, Goes CP, Moreira GCM, Godoy TF, Ibelli AMG, et al. Genome-wide detection of CNVs and their association with performance traits in broilers. *BMC Genom*. 2021;22(1):354.
275. Getmantseva L, Kolosova M, Fede K, Korobeinikova A, Kolosov A, Romanets E, et al. Finding predictors of leg defects in pigs using CNV-GWAS. *Genes*. 2023;14(11):2054.
276. Glessner JT, Khan ME, Chang X, Liu Y, Otieno FG, Lemma M, et al. Rare recurrent copy number variations in metabotropic glutamate receptor interacting genes in children with neurodevelopmental disorders. *J Neurodev Disord*. 2023;15(1):14.
277. Glessner JT, Li J, Wang D, March M, Lima L, Desai A, et al. Copy number variation meta-analysis reveals a novel duplication at 9p24 associated with multiple neurodevelopmental disorders. *Genome Med*. 2017;9(1):106.
278. Han N, Oh JM, Kim IW. Combination of genome-wide polymorphisms and copy number variations of pharmacogenes in Koreans. *J Pers Med*. 2021;11(1):33.
279. Irvin MR, Wineinger NE, Rice TK, Pajewski NM, Kabagambe EK, Gu CC, et al. Genome-wide detection of allele specific copy number variation associated with insulin resistance in African Americans from the HyperGEN study. *PLoS ONE*. 2011;6(8):e24052.
280. Kaivola K, Chia R, Ding J, Rasheed M, Fujita M, Menon V, et al. Genome-wide structural variant analysis identifies risk loci for non-Alzheimer's dementias. *Cell Genom*. 2023;3(6):100316.
281. Kasak L, Rull K, Söber S, Laan M. Copy number variation profile in the placental and parental genomes of recurrent pregnancy loss families. *Sci Rep*. 2017;7:45327.
282. León LE, Benavides F, Espinoza K, Vial C, Alvarez P, Palomares M, et al. Partial microduplication in the histone acetyltransferase complex member KANSL1 is associated with congenital heart defects in 22q11.2 microdeletion syndrome patients. *Sci Rep*. 2017;7(1):1795.
283. Li D, Matsuoka LS, Donoghue S, Hou C, Strong A, McDonald-McGinn DM, et al. Modeling the long-range effect of an inversion downstream of EFNB1 concludes a 43-year molecular diagnostic odyssey for craniofrontonasal syndrome. *Eur J Hum Genet*. 2025;33(12):1684–9.
284. Oliveira P, Costa GNO, Damasceno AKA, Hartwig FP, Barbosa GCG, Figueiredo CA, et al. Genome-wide burden and association analyses implicate copy number variations in asthma risk among children and young adults from Latin America. *Sci Rep*. 2018;8(1):14475.
285. Pathak GA, Polimanti R, Silzer TK, Wendt FR, Chakraborty R, Phillips NR. Genetically-regulated transcriptomics & copy number variation of proctitis points to altered mitochondrial and DNA repair mechanisms in individuals of European ancestry. *BMC Cancer*. 2020;20(1):954.

286. Rambo-Martin BL, Mülle JG, Cutler DJ, Bean LJH, Rosser TC, Dooley KJ, et al. Analysis of copy number variants on chromosome 21 in down syndrome-associated congenital heart defects. *G3 (Bethesda)*. 2018;8(1):105–11.
287. Rymuza J, Kober P, Maksymowicz M, Nyc A, Mossakowska BJ, Woroniecka R, et al. High level of aneuploidy and recurrent loss of chromosome 11 as relevant features of somatotroph pituitary tumors. *J Transl Med*. 2024;22(1):994.
288. Sasaki S, Watanabe T, Ibi T, Hasegawa K, Sakamoto Y, Moriwaki S, et al. Identification of deleterious recessive haplotypes and candidate deleterious recessive mutations in Japanese Black cattle. *Sci Rep*. 2021;11(1):6687.
289. Sha Z, Sun KY, Jung B, Barzilay R, Moore TM, Almasy L, et al. Copy number variant architecture of child psychopathology and cognitive development in the ABCD study. *Am J Psychiatry*. 2025;182(8):763–78.
290. Silva VH, Regitano LC, Geistlinger L, Pértile F, Giachetto PF, Brassaloti RA, et al. Genome-wide detection of CNVs and their association with meat tenderness in Nelore cattle. *PLoS ONE*. 2016;11(6):e0157711.
291. Wang K, Cadzow M, Bixley M, Leask MP, Merriman ME, Yang Q, et al. A Polynesian-specific copy number variant encompassing the MICA gene associates with gout. *Hum Mol Genet*. 2022;31(21):3757–68.
292. Wu J, Wu T, Xie X, Niu Q, Zhao Z, Zhu B, et al. Genetic association analysis of copy number variations for meat quality in beef cattle. *Foods*. 2023;12(21):3986.
293. Wu Y, Adams K. Genome-wide analyses of copy number variants in 751 *Populus trichocarpa* individuals from natural populations. *Genome Biol Evol*. 2025;17(7):evaf136.
294. Xie X, Shi L, Hou G, Zhong Z, Wang Z, Pan D, et al. Genome wide detection of CNV and their association with body size in Danzhou chickens. *Poult Sci*. 2024;103(12):104266.
295. Zhang M, Li Q, Wang KL, Dong Y, Mu YT, Cao YM, et al. Lipolysis and gestational diabetes mellitus onset: a case-cohort genome-wide association study in Chinese. *J Transl Med*. 2023;21(1):47.
296. Zhang W, Yang P, Yang Y, Liu S, Xu Y, Wu C, et al. Genomic landscape and distinct molecular subtypes of primary testicular lymphoma. *J Transl Med*. 2024;22(1):414.
297. Zhang X, Du R, Li S, Zhang F, Jin L, Wang H. Evaluation of copy number variation detection for a SNP array platform. *BMC Bioinform*. 2014;15(1):50.
298. Nurk S, Koren S, Rhie A, Rautiainen M, Bizakdzé AV, Mikheenko A, et al. The complete sequence of a human genome. *Science*. 2022;376(6588):44–53.
299. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffari A, Hickey G, et al. Pangenome graphs. *Annu Rev Genomics Hum Genet*. 2020;21:139–62.
300. Chin CS, Behera S, Khalak A, Sedlazeck FJ, Sudmant PH, Wagner J, et al. Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nat Methods*. 2023;20(8):1213–21.
301. Groza C, Schwendinger-Schreck C, Cheung WA, Farrow EG, Thiffault I, Lake J, et al. Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nat Commun*. 2024;15(1):657.
302. Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat Genet*. 2022;54(4):518–25.
303. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. *Nature*. 2023;617(7960):312–24.
304. Eggertsson HP, Kristmundsdóttir S, Beyter D, Jonsson H, Skuladóttir A, Hardarson MT, et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat Commun*. 2019;10(1):5402.
305. Eggertsson HP, Jonsson H, Kristmundsdóttir S, Hjartarson E, Kehr B, Masson G, et al. GraphTyper enables population-scale genotyping using pangenome graphs. *Nat Genet*. 2017;49(11):1654–60.
306. Bai W-Y, Liu S, Duan Z, Yang J-J, Chen J, Hou J, et al. Genome-wide associations of structural variants with human traits through imputation from long-read assemblies. *Nat Genet*. 2026;58(6):1258–67.
307. Bergen SE, Ploner A, Howrigan D, O'Donovan MC, Smoller JW, Sullivan PF, et al. Joint contributions of rare copy number variants and common SNPs to risk for schizophrenia. *Am J Psychiatry*. 2019;176(1):29–35.
308. Gillani R, Collins RL, Crowdis J, Garza A, Jones JK, Walker M, et al. Rare germline structural variants increase risk for pediatric solid tumors. *Science*. 2025;387(6729):eadq0071.
309. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol*. 2020;38(3):276–8.
310. Porcu E, Rüeger S, Lepik K, Santoni FA, Raymond A, Kutalik Z. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat Commun*. 2019;10(1):3300.
311. Birney E, Vamathevan J, Goodhand P. Genomics in healthcare: GA4GH looks to 2022. *bioRxiv*. 2017:203554.
312. Hofmeister RJ, Ribeiro DM, Rubinacci S, Delaneau O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat Genet*. 2023;55(7):1243–9.
313. Collins RL, Glessner JT, Porcu E, Lepamets M, Brandon R, Lauricella C, et al. A cross-disorder dosage sensitivity map of the human genome. *Datasets*. Zenodo. 2022. <https://doi.org/10.5281/zenodo.6347673>.
314. Harris L, McDonagh EM, Zhang X, Fawcett K, Foreman A, Daneck P, et al. Genome-wide association testing beyond SNPs. Supplementary Information 2. *Datasets*. Springer Nature. 2025. https://static-content.springer.com/esm/art%3A10.1038%2F541576-024-00778-y/MediaObjects/41576_2024_778_MOESM2_ESM.xlsx.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.