Review

Progress in Methods for Copy Number Variation Profiling

Veronika Gordeeva [1,2,3,*], Elena Sharova [2] and Georgij Arapidi [2,3,4]

1   Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Federal Research and
    Clinical Center of Physical-Chemical Medicine of Federal Medical Biological Agency, 119435 Moscow, Russia
2   Federal Research and Clinical Center of Physical-Chemical Medicine of Federal Medical Biological Agency,
    119435 Moscow, Russia; sharova78@gmail.com (E.S.); arapidi@gmail.com (G.A.)
3   Moscow Institute of Physics and Technology, National Research University, Moscow Oblast,
    141701 Moscow, Russia
4   Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences,
    117997 Moscow, Russia
*   Correspondence: gordeeva.veronika@phystech.edu

Abstract: Copy number variations (CNVs) are the predominant class of structural genomic variations involved in the processes of evolutionary adaptation, genomic disorders, and disease progression. Compared with single-nucleotide variants, there have been challenges associated with the detection of CNVs owing to their diverse sizes. However, the field has seen significant progress in the past 20–30 years. This has been made possible due to the rapid development of molecular diagnostic methods which ensure a more detailed view of the genome structure, further complemented by recent advances in computational methods. Here, we review the major approaches that have been used to routinely detect CNVs, ranging from cytogenetics to the latest sequencing technologies, and then cover their specific features.

Keywords: copy number variation; karyotyping; chromosome microarray analysis; long-read and short-read sequencing

1. Introduction

Ever since Joe Hin Tjio and Albert Levan first identified the exact number of chromosomes in human cells in 1956, there have been massive advances in the understanding of the human genome and its structure [1]. In earlier times, due to the low resolution of methods for genetic material analysis, any form of chromosomal rearrangement was associated with disease or an abnormal condition. With advancements in the field after the completion of the human genome project, the use of comparative genomic hybridization (CGH) arrays has led to the discovery of an abundance of copy number variations (CNV), a structural variation of the DNA sequence which has a >50 bp multiplication and deletions of a particular segment of DNA, in the human genome [2,3]. It was shown that CNVs are widespread in human populations and comprise about 5–10% of the genome [4]. There is growing evidence showing the role of CNVs in causing various disorders. The list of diseases for which association with copy number variations has been established includes schizophrenia, type I diabetes, autism, cardiovascular diseases, congenital abnormalities, and neurodegenerative diseases [5]. However, due to the wide range of possible lengths and non-trivial estimation of the effect of CNVs on phenotype, there are difficulties in accurately screening and detecting CNVs. Due to limited resolution, modern technologies do not allow for complete CNV profile descriptions. However, each new approach to CNV detection has introduced new, valuable improvements. For example, chromosomal microarray analysis has allowed for large-scale genome studies with high resolution; whole-genome sequencing has provided the tool for the identification of all types of structural variations starting from 50 bp; and emerging sequencing technologies have opened up possibilities for the investigation of regions that were poorly accessible due to short reads.

An enormous amount of copy number analysis methods using various types of data have been elaborated, and most of them have been described in detail in early reviews dedicated to a particular single technology or several CNV detection technologies [6–9]. Recently, the ELIXIR human CNV Community was initiated and focuses primarily on detection, annotation, and variant interpretation issues, which will ultimately help in developing an optimal protocol for CNV identification [10].

Hence, the aim of this review is to highlight the major contributions made in the field to detect CNVs to date and describe their main principles and challenges. We will also describe the most widely used computational approaches and discuss burning issues in this research field.

2. Methods of Cytogenetics

The first approaches to CNV detection were based on the analysis of metaphase plates in cells. During metaphase in cell division, condensed chromosomes align along the cell equator, facilitating convenient viewing of chromosome structure through light microscopy. Until the early 1970s, the only method for chromosome staining was Giemsa staining. In such an approach, all chromosomes are stained evenly length-wise, allowing for changes in karyotype concerning chromosome number, shape, and size to be visualized. Giemsa staining was used to detect many chromosome aneuploidies, such as Down's syndrome [11], Klinefelter's syndrome [12], and Edwards syndrome [13], as well as the structural abnormalities of chromosomes in cancer cells [14].

After the advent of Giemsa staining, new methods of differential staining that allowed for visualization of certain chromosomal structures, e.g., C-banding for centromere or T-banding for telomere regions staining, as well as producing chromosome-specific sequences of dark and light bands along the length (R, Q, and G-banding) were developed. The most common was the G-banding technique, a method that implies preliminary treatment of chromosomes with trypsin for DNA denaturing and further staining of renatured chromosomes with Giemsa dye [15]. Pairs of homologous chromosomes are arranged according to their number (sex chromosomes on the end) with their short shoulders upwards, ensuring the centromeres are horizontally aligned [16] (Figure 1a). The technique has the potential to elucidate rather large abnormalities exceeding 5 Mb in length and is used today in cases of suspected or existing congenital pathologies, as well as family planning.
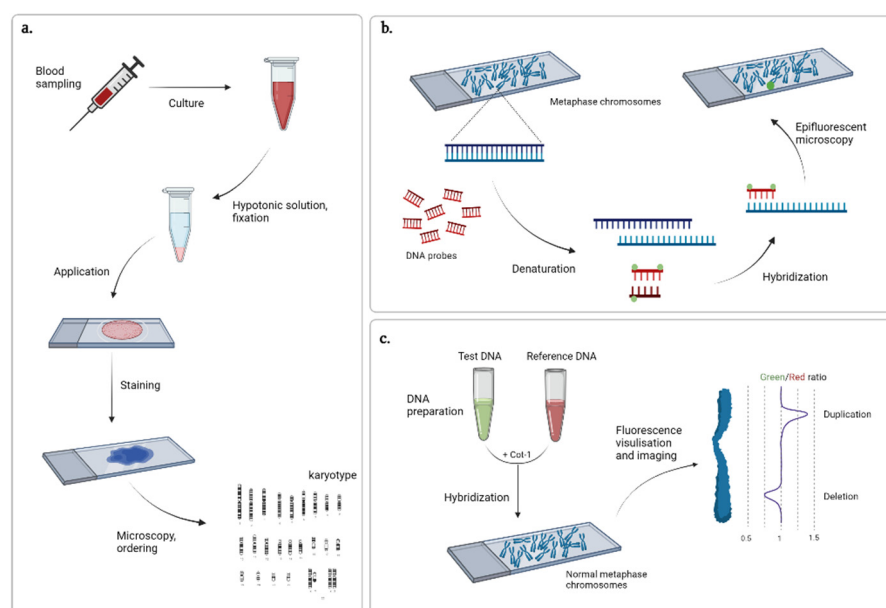
Figure 1. Cytogenetic techniques: (**a**) karyotyping, (**b**) FISH, and (**c**) comparative genome hybridization. Created with BioRender.com (accessed on 9 February 2022).

More detailed analyses of the genome became possible upon the development of DNA hybridization techniques. At first, tritium-labeled RNA probes were used to search for a specific DNA sequence and the hybrids were detected by autoradiography [17]. Later, the protocol was modified with fluorescent labels replacing the radioactive ones, thus improving safety and decreasing labor consumption [18,19]. Due to its targeted nature, fluorescence in situ hybridization (FISH) (Figure 1b) remains one of the most commonly used techniques to verify variations previously identified by real-time qPCR and multiplex ligation-dependent probe amplification (MLPA). However, its principle has been refined for use in other techniques, such as spectral karyotyping and multicolor FISH, offering new approaches to the visualization and analysis of chromosomes with the use of various combinations of fluorochromes and light filters [20,21].

For whole-genome analysis, comparative genome hybridization was developed in 1992 [22]. In contrast to FISH, where metaphase chromosomes obtained from a patient and DNA probes are complementary to the sequence of interest, the method proposed by Kallioniemi and colleagues isolated and hybridized differentially labeled genomic DNA from the tumor and normal tissue of a patient and utilized chromosome samples from peripheral blood lymphocytes of a healthy volunteer as a reference (Figure 1c). Further refinement of the protocol and development of specialized software for image analysis made the technique more accessible to many laboratories [23]. It has been widely used for the cytogenetic analysis of solid tumors [24]. However, utilization of metaphase chromosomes considerably limits the method resolution (the average size of detected CNVs is 5–10 Mb). Therefore, further research has been aimed at alternative representations of the cytogenetic map.

3. Chromosome Microarray Analysis (CMA)

The application of microarrays instead of metaphase chromosomes has provided a solution for large-scale genome studies. Hybridization occurs on multiple DNA probes attached to a solid surface, and physical position on a chip and specific nucleotide sequences of the probes are pre-determined, which allows visualization of relative fluorescence intensities of test and control samples along the genome. For example, in 1997, the matrix-based comparative genomic hybridization method was introduced [25], and soon after, array-based comparative genomic hybridization (aCGH) was introduced [26] (Figure 2a). Resolution of the method depends directly on the probe type, number, and distribution over the genome. At that time, a large set of bacterial artificial chromosome (BAC) clones 80–200 kb long spanning the genome in a fairly complete manner was created in the course of the human genome project [27]. This genomic library has been used for the construction of most CGH arrays capable of identifying variations exceeding 1 Mb [28,29]. In general, any nucleotide sequence can be used as a probe. Later, cDNA- [30,31] and PCR amplicon-based [32] arrays have been successfully used for DNA copy number analysis.

Further improvement of the method occurred upon the deciphering of the whole-genome sequence. This development led to the application of oligonucleotide probes (8–25 bp) that had been previously used for gene expression studies [33,34]. Oligonucleotide probes simplify the process of chip development in regard to both design customization and reproducibility. In addition, these microarrays provide better signal-to-noise ratio and event (CNV) confirmation with several probes due to a more complete genome coverage [35]. Array capacity has also improved, as the mechanical application of DNA onto a chip limited by approximately 50,000 probes has been replaced with oligonucleotide synthesis directly on the glass surface. Commercial technology proposed by Agilent—one of the leading comparative chromosomal analysis technologies today—synthesizes 60-nucleotide probes by an ink-jet technology [36]. Several chip formats have been developed (1×1 M, 2×400 K, 4×180 K, and 8×60 K) for whole-genome studies or for the targeted analysis of regions that have been strongly associated with various diseases (Table 1). A higher density of probes is typically used for regions of interest with the distance between

adjacent probes as little as several hundreds of nucleotides, and coverage of the rest of the genome provided by evenly spread backbone probes.
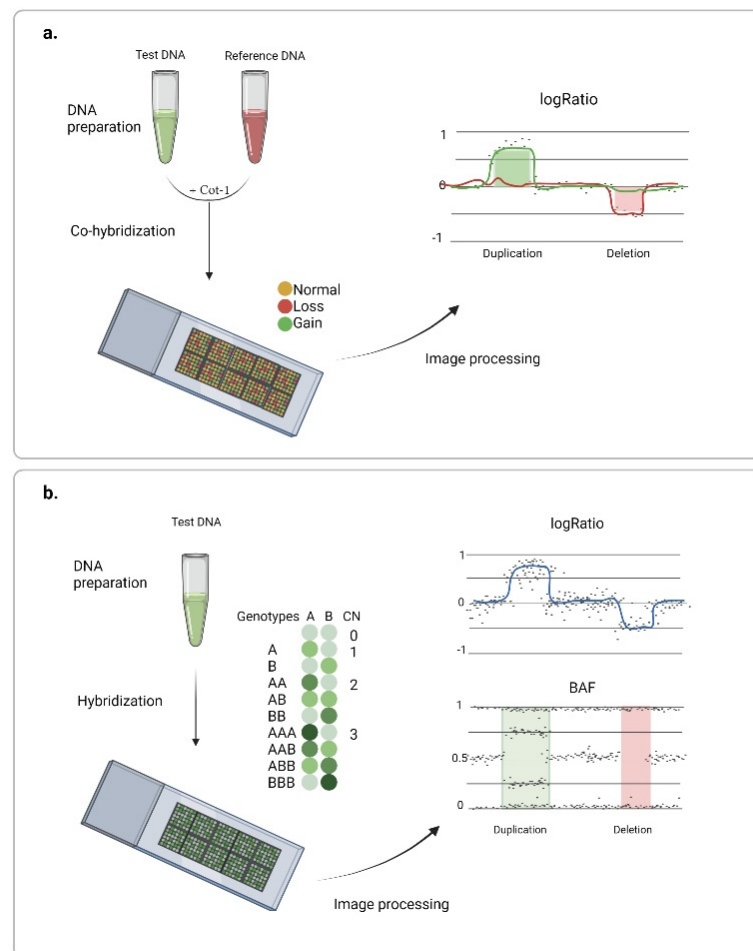
Figure 2. Chromosome microarray analysis: (**a**) array-based comparative genomic hybridization, and (**b**) DNA arrays for genotyping. Created with BioRender.com (accessed on 9 February 2022).

Another type of microarray used to search for regions of chromosome imbalance has been developed for genotyping and genome-wide association studies. Here, only fluorescently labeled DNA of the sample is applied to the chip, and copy number analysis is performed based on the rate of target DNA hybridization to allele-specific probes (Figure 2b). Such a technique was first demonstrated in 2004 using the Affymetrix chips [37]. Soon, Illumina introduced an alternative platform, Infinium BeadChip, in which after DNA hybridization occurs on complementary probes, SNPs are evaluated based on the brightness and color of fluorescent nucleotides attached to the probe. The platform could detect deletions and duplications of fragments shorter than 100 kb [38]. However, earlier versions of the chips limited the scope of the search for variation to the regions covering common polymorphisms. Later, the introduction of non-polymorphous probes allowed for more even genome coverage [39]. Today, a wide choice of DNA arrays equipped with 300,000 to several million probes is available on the market (Table 1), including arrays with the possibility of partial or complete customized designs. High genome coverage with short probes provides for higher resolution and more accurate determination of the breakpoints, which makes this type of array a convenient tool when searching for rare and short (starting from 500 bp) variations. Another advantage of the technology is the possibility of identifying mosaicism, loss of heterozygosity (LOH), and uniparent disomies. Soon, the inherent superiority of CMA replaced G-karyotyping as a first-line test in clinical diagnostics of multiple congenital anomalies [40].

Table 1. Modern platforms for chromosomal microarray analysis.

Array Platform	Specification *	Resolution **	Description
Agilent SurePrint G3 Human CGH	1 × 1 M	2.1 kb	enhanced coverage on known genes, promoters, miRNAs, PAR, and telomeric regions
	2 × 400 K	5.3 kb	
	4 × 180 K	13 kb	
	8 × 60 K	41 kb	
Agilent Human Genome CGH	2 × 105	35 kb	
	4 × 44 K	43 kb	
Agilent SurePrint G3 Human Genome CGH + SNP	2 × 400 K	7.2 Kb	
	4 × 180 K	25.3 kb	
Agilent SurePrint G3 Unrestricted CGH ISCA v2	4 × 180 K	25 kb	enhanced coverage on ISCA (International Standards for Cytogenomic Arrays) regions
	8 × 60 K	60 kb	
	4 × 44 K	75 kb	
Agilent SurePrint G3 ISCA v2 CGH + SNP	4 × 180 K	25.3 kb	
Agilent SurePrint G3 Human High-Resolution Discovery	1 × 1 M	2.6 kb	association studies
Agilent SurePrint G3 Human CNV	2 × 400 K	1 kb	
Agilent Human CNV Association	2 × 105 K	232 b	
Agilent SurePrint G3 CGH Postnatal Research	4 × 180 K	2.4 kb	regions identified by Baylor College of Medicine experts
	8 × 60 K	3.7 kb	
Agilent GenetiSure Postnatal Research CGH + SNP	2 × 400 K	9.8 kb	disease-associated regions (The Clinical Genome/ISCA database)
Agilent GenetiSure Pre-Screen	4 × 180 K	31 kb	CNV identification from embryo biopsies and single-cell samples; increased density on chromosomes 13, 18, 20, 21, 22, and X
	8 × 60 K	50 kb	
Agilent GenetiSure Cyto CGH	4 × 180 K	3.5 kb	disease-associated regions linked to developmental delay, intellectual disability, neuropsychiatric disorders, congenital anomalies, or dysmorphic features
	8 × 60 K	7.1 kb	
Agilent GenetiSure Cyto CGH + SNP	4 × 180 K	7.3 kb	

Table 1. *Cont.*

Array Platform	Specification *	Resolution **	Description
Agilent GenetiSure Cancer Research CGH + SNP	2 × 400 K	9.8 kb	cancer regions of the genome COSMIC (Catalogue of Somatic Mutation in Cancer) CGC (Cancer Genetics Consortium) databases
Illumina HumanCytoSNP	12 × 300 K	6.2 kb	enhanced coverage of ~250 disease regions, including subtelomeric regions, pericentromeric regions, and sex chromosomes
Illumina Infinium CytoSNP-850 K	8 × 850 K	1.8 kb	comprehensive coverage of cytogenetically relevant genes for congenital disorders and cancer research ICCG (International Collaboration for Clinical Genomics) and CCMC (Cancer Cytogenomics Microarray Consortium)
Illumina Infinium Core	24 × 300 K	5.8 kb	genome-wide tag SNPs found across diverse world populations
Illumina Infinium Exome	24 × 300 K	0.21 kb	comprehensive coverage of putative functional exonic variants (including markers representing a range of common conditions, such as type 2 diabetes, cancer, and metabolic, and psychiatric disorders)
Illumina Infinium CoreExome	24 × 600 K	1.82 kb	all of the markers from the Infinium Core-24 BeadChip and the Infinium Exome-24 BeadChip
Illumina Infinium Global Diversity Array	8 × 2 M	0.63 kb	common and low frequency variants in global populations, curated clinical research variants
Illumina Infinium Global Screening Array	24 × 700 K	2.3 kb	multiethnic genome-wide content, curated clinical research variants
Illumina Infinium Omni2.5	8 × 2.4 M	0.65 kb	common and rare SNP content from the 1000 Genomes Project (MAF > 2.5%)
Illumina Infinium Omni2.5Exome	8 × 2.7 M	0.56 kb	combined Infinum Omni2.5 and Infinium Exome-24 markers
Illumina Infinium Omni5	4 × 4.3 M	0.36 kb	comprehensive coverage of the genome including common, intermediate, and rare SNPs
Illumina Infinium Omni5 Exome	4 × 4.6 M	0.33 kb	comprehensive genome-wide backbone combined with putative functional exonic variants
Illumina Infinium OmniExpress	24 × 700 K	2.23 kb	high coverage of common variants for genome-wide association studies

Table 1. *Cont.*

Array Platform	Specification *	Resolution **	Description
Illumina Infinium OmniExpressExome	8 × 1 M	1.36 kb	tag SNPs and functional exonic content
Illumina Infinium OncoArray	24 × 500 K	5.4 kb	genetic variants associated with five common cancers
Illumina Infinium PsychArray	24 × 700 K	1.74 kb	genetic variants associated with common psychiatric disorders
Affymetrix Genome-Wide Human SNP Array 6.0	1 × 1.8 M	0.68 kb	comprehensive coverage of the genome
Affymetrix CytoScan XON Suite	24 × 6.85 M	0.5 kb	enhanced coverage in 7000 clinically relevant gene, exon-level copy number changes
Affymetrix CytoScan HD	24 × 2.7 M	1.3 kb	enhanced coverage on cytogenetic relevant region

* Samples × No. probes. ** Overall median probe spacing.

Not all chip platforms are equally fit for CNV detection. Although for any type of microarray (CGH, SNP) there is a dependence of the number of identified variations on the total number of probes, the design plays an equally important role [41]. Haraksingh and colleagues demonstrated that a suboptimal number of backbone probes or lack of probes in the intergenic regions results in a disability to identify large, potentially biologically important variations. Further, the comparison of existing platforms shows that additional enrichment with exome variants is not always optimal for CNV detection, as is the case with the Illumina Infinium Omni lineage. However, despite all platform-specific features, the choice of detection algorithm and settings wields the most influence on the results [41,42].

The aim of any algorithm is the analysis of signal intensity along the genome (logR) and elucidating any significant deviations. For this purpose, both simple thresholds, e.g., change of logR by 0.2–0.3 [43–45] or by more than 3–4 standard deviations [46,47], and more complex statistical models can be used. Data segmentation methods have found wide applications. For example, Olshen et al. adopted the method of binary segmentation to determine all sites dividing the genome into fragments of the same copy number [48]. In the modified version, two points are determined at each stage, limiting a region so that the t-statistics for the difference in mean probe intensity inside and outside the region are the highest. If the differences are statistically significant, the region is divided, and the procedure is repeated for each of the new three regions, if possible. An alternative method is based on a local genetic search. In this technique, the focus is on the optimization of a certain set of points chosen randomly by means of minimization of the negative logarithmic function of likeness and penalty for the high rate of genome fragmentation [49]. Later on, other methods for change-point detection have been suggested, e.g., ones that are based on the techniques of adaptive weight smoothing [50], fused lasso [51,52], and local search [52,53].

A large class of algorithms utilize hidden Markov models (HMM), allowing for the description of a system with unknown variables based on the observed ones. Changes in the number of copies (loss, gain, or maintenance of genetic material) [54] or absolute number of copies as such [55] are considered hidden states, while their most probable sequence is determined using the dynamic programming after preliminary optimization of the model parameters (the initial probabilities of the states, transition probabilities, and emission probabilities). HMMs are convenient as they work with several variables, as in the case of DNA microarrays. In the latter case, in addition to normalized probe intensities, the ratio of intensities between polymorphous alleles is also under investigation. Additional parameters, such as distance between the probes [56] and the population frequencies of SNPs [57], can be taken into account as well.

Today, multiple ways for microarray data to be processed have been proposed (Table 2). They differ by the number and size of CNVs detected and false-discovery rate [7,42,58–60]. As a result, it is recommended to use several algorithms to improve efficiency. Regardless, CMA remains one of the most in-demand methods for CNV detection in both research and clinical diagnostics due to its reliability, flexibility, and relatively low cost. To search for clinically important variants, aCGH or DNA microarrays developed specifically for cytogenetic studies (e.g., Affimetrix Cytoscan HD) and capable of highly reliable identification of variations exceeding 25 kb are preferred.

Table 2. The most widely used algorithms for CNV detection that use microarray data.

Tool	Description	aCGH	SNP-Array		Reference
			Affymetrix	**Illumina**	
ADM-2	search for intervals in which a Z-score based on the average weighted log ratio exceeds a user-specified threshold	✓			technical documentation (Agilent)
Birdsuite	integration of common CNP genotypes and CNVs discovered using HMM		✓		[61]
ChAS	HMM on the log2 ratios processed through a Bayes wavelet shrinkage estimator		✓		technical documentation (Affymetrix)
cnvPartition	recursive partitioning approach based on preliminary copy number estimates			✓	technical documentation (Illumina)
DNAcopy	circular binary segmentation	✓			[48]
GenoCN	estimation of HMM, parameters from data, germline, and somatic modes		✓	✓	[62]
iPattern	normalization of the total intensities across individuals, Gaussian mixture model fitting		✓	✓	[42]
Nexus	the probe's log-ratio rank segmentation	✓	✓	✓	[63]
PennCN	HMM, also counted for the population frequency of the B allele		✓	✓	[57]
QuantiSNP	objective Bayes-HMM, fixed rate of heterozygosity for each SNP		✓	✓	[64]

4. Sequencing Data

Methods of CNV detection gained momentum with the rise of high-throughput sequencing (new generation sequencing, or NGS) technology. Based on the analysis of millions of short readings, this was quickly deemed a revolutionary technique due to its high productivity, reproducibility, and accuracy. Data generated from this technique can also be used repeatedly in various types of studies, and this research domain has expanded considerably concerning some genomic variations that were previously difficult to detect. Firstly, the microarray technique had limitations in accessing balanced rearrangements. Secondly, as a consequence of multiple sequencing of random fragments, variation size is not strictly limited as is the case in CMA, where variation less than the distance between two probes cannot be resolved.

Detection algorithms are readily being developed in several directions since NGS data can be described by various signatures (Figure 3). One of the first approaches was based on the read pair concept (RP). It implies the presence of aberrations under the condition that distances between the mapped reads on a reference genome reads and/or their orientation is different from the expected ones [65,66]. To search for clusters of such abnormal RPs, two strategies were proposed: in the first case, the distance between the paired reads (insert size) is considered known and constant [67,68], and at least two discordant pairs are required to form a cluster; in the second case, the distribution of insert size for each region over the whole genome is taken into account [69,70]. In the latest studies, concordant read pairs are also taken into account, implying that smaller variations can thus be traced [71].
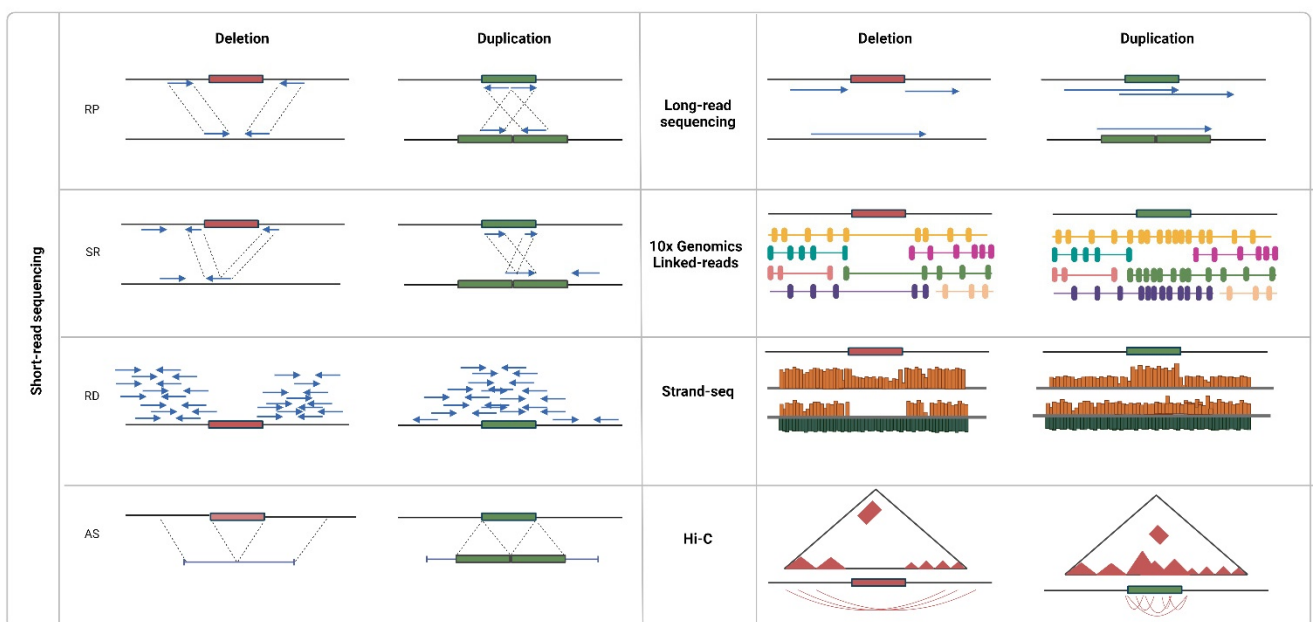
Figure 3. Approaches to CNV detection using sequencing data. Created with BioRender.com (accessed on 9 February 2022).

The split-read (SR) approach also originated from the incoherence property; however, in this case, read is not mapped onto the genome at all or is only partially mapped. The following repeated alignment of the read parts can indicate possible coordinates of the start and end of the variation [72]. The SR approach is suitable for both single and paired reads, but the latter ones impose additional limitations, which accelerate the search. For example, in the Pindel algorithm [73], alignment of the 3′ end of unmapped paired-ends is performed within the double insert size from the 3′ end of mapped paired-ends. The SR approach exists as a rare method to identify deletions 50–100 bp long. Due to its sensitivity to the quality of alignment, it is intended for studies of unique regions of the genome.

In contrast to previous methods indicating only the presence of a variation, the next approach was intended to evaluate the number of copies. The read-depth (RD) approach is based on the assumption that region coverage correlates with the number of its copies. Not limited by either read length or insert size, the approach is suitable for the identification of preferably large variations. The standard detection procedure consists of four stages: mapping of reads and calculation of coverage; normalization; segmentation; and evaluation of the copy number. In the first stage, a genome, or a sliding window, is usually used. The window size can be both chosen voluntarily (for example, 100 nucleotides was considered sufficient for the identification of small variations and accurate search for the breakpoints [74]) or selected based on the data and the desired confidence level of the CNV event [75,76]. Presumably, the number of reads per window is distributed normally; however, in reality, the coverage is shifted in the function of the GC content of the regions [77] and depends on mappability [50]. To take these factors into account, mean normalization methods [78–80], as well as various regression models [81–84], have been proposed. Segmentation and evaluation of copy numbers are performed by most CMA methods, including HMM, mixed Gaussian models, LASSO regression, and CBS.

The RD approach is applicable to the data of whole-exome or targeted sequencing [85]. Although identification of most variations does not seem possible, this type of analysis is convenient in the primary search for patterns specific to a disorder. In addition to the higher coverage of target regions, one should take into account that during genomic library preparation, the efficiency of the enrichment of target regions varies and some regions are over or under-represented. To describe exome data, various models have been proposed, including Gaussian [86], Poissonian [87], beta-binomial [88], and negative binomial [89] distributions (Table 3). In addition, the discrete structure of the data, with rare exceptions, does not allow an analysis of the exact breakpoints of the CNV; however, the analysis can be expanded by using the information on non-target regions that make up 30–40% of sequencing data and providing low genome coverage [78,90]. Another important issue considered in the framing of the problem is the choice of reference samples which are used at the stage of normalization with questions, such as how many reference samples are necessary for efficient detection and are all of them equally useful at being vitally important. The most frequently used strategies include having all available samples sequenced on the same platform with the same chemistry, having all samples in a single sequencing run, or using a set of the most coverage-correlated samples [91].

Recently, the team behind the GATK (Genome Analysis Toolkit), which is known as the most popular tool for analyzing short genomic variations, has also proposed a pipeline to call rare and common germline copy number variants. It uses negative-binomial factor analysis and HMM, and requires at least one-hundred samples to build a model (https://github.com/broadinstitute/gatk/blob/master/docs/CNV/germline-cnv-caller-model.pdf, accessed on 9 February 2022). A somatic mode is also available.

Table 3. The most widely used algorithms for whole-exome and targeted data sequencing.

Tool	Description	Data		Mode		Reference
		WES	Targeted	Germline	Somatic	
cn.MOPS	mixture Poisson model and Bayes approach	✓	✓	✓	✓	[92]
CNVkit	in- and off-target regions, rolling median bias correction, CBS	✓		✓	✓	[78]
CODEX	log-linear decomposition-based normalization, Poisson likelihood-based segmentation	✓	✓	✓	✓	[87]
CoNIFER	singular value decomposition-based normalization, ± 1.5 SVD-ZRPKM threshold	✓		✓		[93]
CoNVaDING	ratio scores and Z-scores of the sample of interest compared to the selected control		✓	✓		[94]
DECoN	ExomeDepth modification (the distance between exons is taken into account)		✓	✓		[95]
ExomeDepth	beta-binomial distribution, optimized reference set, HMM	✓	✓	✓		[88]
XHMM	principal component analysis normalization, HMM	✓		✓		[86]

The last approach to detect CNVs using NGS data implies that first, DNA fragments are assembled from overlapping reads de novo, and then the contigs are aligned onto a reference genome (AS). The approach does not require high coverage and is potentially fit for the identification of all types of structural variations, especially new ones. The assembly is performed using graph models (overlapping graphs built by the overlap layout consensus (OLC) method and de Bruijn graph [96] are most often used). Searches for variations can be performed without a reference genome; in such a case, the graph is constructed for several samples and is then analyzed for bifurcations and copy number [97,98].

Despite progress, each of the four described approaches alone is not able to identify the whole range of variations; therefore, the next step was the development of combined methods (Table 4). The most frequently used combinations were RP plus SR or RP plus RD, which are capable of the identification of variations of different lengths and more accurate prediction of breakpoints. Later, algorithms started to take into account the advantages of all three signatures [99–103], which allows a decreased number of false-positive identifications.

Table 4. Combinations of approaches to the analysis of whole-genome sequencing and the most frequently implemented algorithms.

Approach	Tool	Description	Reference
RP	BreakDancer	search for regions that include more anomalous read pairs than expected	[67]
SR	Pindel	pattern growth approach for breakpoint identification	[73]
RD	CNVnator	mean-shift technique, multiple-bandwidth partitioning, and GC correction	[79]
AS	Cortex	bubble-calling in the colored de Bruijn graph	[98]
RP + RD	GenomeSTRiP	connected components algorithm for read pair clustering, Gaussian mixture model for read depth genotyping	[104]
RP + SR	DELLY	graph-based paired-end clustering, breakpoints refinement using split-read alignment	[105]
RP + AS	Hydra	assembly of discordant mate pairs and aligned to the reference genome with MEGABLAST	[106]
RP + SR + AS	Manta	breakend graph construction, independent for each edge variation hypothesis refinement and scoring with diploid model	[103]
RP + SR + RD	Lumpy	probabilistic representation of an SV breakpoint	[102]
Ensemble	MetaSV	merging calls from tools (BreakDancer, CNVnator, BreakSeq, Pindel), breakpoint refinement by aligning the assembled CNV regions	[107]

Often, CNV detection is narrowed down to classification problems solved by methods of machine learning. Along with various signatures, a number of additional factors, for example, mapping quality (MAPQ) or nucleotide content, are considered [108]. In addition, it is possible to take into account specific features of the formed validation sets, which typically contain intermediate-sized variants. For example, a one-class model trained on a representative set of regions with normal copy numbers searches for regions unlike those in the set, thus covering variations of varying type and size [109]. The latest developments include the DeepSV algorithm based on the analysis of mapped read images [110].

Integration of the data can proceed not only at the level of the signatures, but also at the level of the variants predicted by multiple individual algorithms. Today, the so-called

ensemble-based approach is not standardized in any way. Various methods are used for combining and evaluating the variants, including coordinate overlapping, distance between the breakpoints, signature prioritization, agreement of the algorithms, number of confirmations of the event, confidence intervals of the breakpoints, FDR cutoff, and metaheuristics [111–113]. Despite the improvement in the accuracy of predictions, all these methods are limited by the characteristics of the input data (short read or insert size), and they do not allow a comprehensive analysis of complex genome regions.

On the contrary, long reads (above 5 kb) produced by third-generation sequencing machines can be used to solve the problem, despite their lack of accuracy. The Pacbio platform performs real-time sequencing of a single molecule through the synthesis of a new strand using a polymerase bound to the well bottom and registration of fluorescence of each newly added nucleotide. In turn, the Oxford nanopore technology is based on the evaluation of change in an electrical current induced by a single-stranded DNA molecule passing through a nanopore. In either case, to detect variations, signatures inside the reads completely covering variations and signatures indicating the presence of a variation based on discrepancies between the reads (orientation, size, or location) are analyzed [114,115]. Moreover, calling efficiency depends more on coverage than on the read length or error rate [116].

As an alternative to long reads, the sequencing of bound molecules can be used. For example, using additional barcodes to indicate association to a single DNA molecule, synthetic long reads can be efficient in the detection of large variations. The identification of variants proceeds based on the evaluation of the density of molecule coverage by paired reads, the excessive increase or decrease of which implies a CNV event in a certain region [117,118]. In Strand-seq technology, DNA strands are sequenced independently, and large deletions or duplications are identified based on the evaluation of coverage rate [119]. The size of predicted variations is often limited, which is due to the considerable decrease of coverage resulting from the many purification steps during library preparation. Another method, Hi-C, has been developed to study the 3D structure of chromatin, in particular, to determine the nucleotide sequences that are separately located in the genome but still interact with each other. The matrix of contacts between any two loci can be transformed into the coverage profile with a certain resolution; then, signature-typical analysis methods are applied (normalization, segmentation, and copy number evaluation) [119,120].

In addition to the sequence-based methods mentioned above, optical mapping that aims to determine the physical location of specific sequence motifs or enzymes has great potential for CNV calling. The method first demonstrated in 1993 on the example of construction restriction maps of *Saccharomyces cerevisiae* chromosomes has undergone some modifications [121]. Today's workflow includes the isolation of high molecular weight DNA, labeling of specific sequences across the entire genome, single-molecule DNA linearization in nanochannels, and imaging using high-resolution fluorescent microscopy. The data obtained can be used for genome assembly improvement [122], haplotype phasing [123], and searches for large structural variations [124]. A CNV event is defined by changes in the density or the distance between restriction sites compared with a reference map obtained from the in silico digestion of the reference sequence. Errors most often occur due to excessive or insufficient stretching of the molecule in the channel, non-specific enzymatic cuts, and incomplete enzyme digestion [125].

5. Conclusions

Despite considerable progress, the identification of copy number variations remains a difficult task. Each of the proposed approaches, from cytogenetics to emerging sequencing technologies, capable of the analysis of the so-called dark genome regions has its own limitations. In an attempt to neutralize the latter, over one-hundred methods of analysis have been developed. Particular attention is paid to the development of algorithms aimed at exome and targeted sequencing as optimal tools for applied methods of genome analysis in regard to information load and cost [85], as well as integration of data of any kind, including

the use of ensemble models [126–128]. Therefore, researchers face a huge solution space from which they usually choose established algorithms, even though they may be less ideal than newer approaches. Algorithms are mainly shaped to detect certain types of variations or length ranges, which should be considered when choosing the approach [129].

Evaluations of the efficiency of existing CNV detection methods and understanding their advantages and limitations are complicated by the lack of a comprehensive validation set. The only available approach for the description of many genomic variations today is the integration of the results of several platforms, as proposed by Zook et al., for simple deletions and insertions [130]. Expansion of the set of platforms used for the analysis and improvement of their accuracy, as well as the development of protocols for integration of the whole bulk of information, are the key problems in this area of research. The design of a variation profile is necessary not only to be used as a reference in the choice of appropriate detection method, but also for basic research focuses, such as the study of the role and function of CNVs, evaluation of their effect on pathogenesis, and many others.

References

1. Hollox, E.J.; Zuccherato, L.W.; Tucci, S. Genome Structural Variation in Human Evolution. *Trends Genet.* **2022**, *38*, 45–58. [CrossRef] [PubMed]
2. Iafrate, A.J.; Feuk, L.; Rivera, M.N.; Listewnik, M.L.; Donahoe, P.K.; Qi, Y.; Scherer, S.W.; Lee, C. Detection of Large-Scale Variation in the Human Genome. *Nat. Genet.* **2004**, *36*, 949–951. [CrossRef] [PubMed]
3. Redon, R.; Ishikawa, S.; Fitch, K.R.; Feuk, L.; Perry, G.H.; Andrews, T.D.; Fiegler, H.; Shapero, M.H.; Carson, A.R.; Chen, W.; et al. Global Variation in Copy Number in the Human Genome. *Nature* **2006**, *444*, 444–454. [CrossRef] [PubMed]
4. Zarrei, M.; MacDonald, J.R.; Merico, D.; Scherer, S.W. A Copy Number Variation Map of the Human Genome. *Nat. Rev. Genet.* **2015**, *16*, 172–183. [CrossRef] [PubMed]
5. Shaikh, T.H. Copy Number Variation Disorders. *Curr. Genet. Med. Rep.* **2017**, *5*, 183. [CrossRef]
6. Roca, I.; González-Castro, L.; Fernández, H.; Couce, M.L.; Fernández-Marmiesse, A. Free-Access Copy-Number Variant Detection Tools for Targeted next-Generation Sequencing Data. *Mutat. Res./Rev. Mutat. Res.* **2019**, *779*, 114–125. [CrossRef]
7. Seiser, E.L.; Innocenti, F. Hidden Markov Model-Based CNV Detection Algorithms for Illumina Genotyping Microarrays. *Cancer Inform.* **2015**, *13*, CIN-S16345. [CrossRef] [PubMed]
8. Carter, N.P. Methods and Strategies for Analyzing Copy Number Variation Using DNA Microarrays. *Nat. Genet.* **2007**, *39*. [CrossRef]
9. Mahmoud, M.; Gobet, N.; Cruz-Dávalos, D.I.; Mounier, N.; Dessimoz, C.; Sedlazeck, F.J. Structural Variant Calling: The Long and the Short of It. *Genome Biol.* **2019**, *20*, 1–14. [CrossRef]
10. Salgado, D.; Armean, I.M.; Baudis, M.; Beltran, S.; Capella-Gutierrez, S.; Carvalho-Silva, D.; Del Angel, V.D.; Dopazo, J.; Furlong, L.I.; Gao, B.; et al. The ELIXIR Human Copy Number Variations Community: Building Bioinformatics Infrastructure for Research. *F1000Research* **2020**, *9*, 1229. [CrossRef]
11. Roizen, N.J.; Patterson, D. Down's Syndrome. *Lancet* **2003**, *361*, 1281–1289. [CrossRef]
12. Lanfranco, F.; Kamischke, A.; Zitzmann, M.; Nieschlag, E. Klinefelter's Syndrome. *Lancet* **2004**, *364*, 273–283. [CrossRef]
13. Cereda, A.; Carey, J.C. The Trisomy 18 Syndrome. *Orphanet J. Rare Dis.* **2012**, *7*, 81. [CrossRef] [PubMed]
14. Nowell, P.C. The Minute Chromosome (Phl) in Chronic Granulocytic Leukemia. *Blut* **1962**, *8*, 65–66. [CrossRef]
15. Bayani, J.; Squire, J.A. Traditional Banding of Chromosomes for Cytogenetic Analysis. *Curr. Protoc. Cell Biol.* **2004**, *22*, 22.3.1–22.3.7. [CrossRef]
16. Swansbury, J. Introduction to the Analysis of the Human G-Banded Karyotype. *Methods Mol. Biol.* **2003**, *220*, 259–269. [PubMed]
17. Gall, J.G.; Pardue, M.L. Formation and Detection of RNA-DNA Hybrid Molecules in Cytological Preparations. *Proc. Natl. Acad. Sci. USA* **1969**, *63*, 378–383. [CrossRef]

18. Rudkin, G.T.; Stollar, B.D. High Resolution Detection of DNA-RNA Hybrids in Situ by Indirect Immunofluorescence. *Nature* **1977**, *265*, 472–473. [CrossRef]

19. Bauman, J.G.; Wiegant, J.; Borst, P.; van Duijn, P. A New Method for Fluorescence Microscopical Localization of Specific DNA Sequences by in Situ Hybridization of Fluorochromelabelled RNA. *Exp. Cell Res.* **1980**, *128*. [CrossRef]

20. Schrock, E.; du Manoir, S.; Veldman, T.; Schoell, B.; Wienberg, J.; Ferguson-Smith, M.A.; Ning, Y.; Ledbetter, D.H.; Bar-Am, I.; Soenksen, D.; et al. Multicolor Spectral Karyotyping of Human Chromosomes. *Science* **1996**, *273*, 494–497. [CrossRef] [PubMed]

21. Speicher, M.R.; Ballard, S.G.; Ward, D.C. Karyotyping Human Chromosomes by Combinatorial Multi-Fluor FISH. *Nat. Genet.* **1996**, *12*, 368–375. [CrossRef]

22. Kallioniemi, A.; Kallioniemi, O.P.; Sudar, D.; Rutovitz, D.; Gray, J.W.; Waldman, F.; Pinkel, D. Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors. *Science* **1992**, *258*, 818–821. [CrossRef] [PubMed]

23. Kallioniemi, O.P.; Kallioniemi, A.; Piper, J.; Isola, J.; Waldman, F.M.; Gray, J.W.; Pinkel, D. Optimizing Comparative Genomic Hybridization for Analysis of DNA Sequence Copy Number Changes in Solid Tumors. *Genes Chromosomes Cancer* **1994**, *10*, 231–243. [CrossRef]

24. Gebhart, E. Comparative Genomic Hybridization (CGH): Ten Years of Substantial Progress in Human Solid Tumor Molecular Cytogenetics. *Cytogenet. Genome Res.* **2004**, *104*, 352–358. [CrossRef] [PubMed]

25. Solinas-Toldo, S.; Lampel, S.; Stilgenbauer, S.; Nickolenko, J.; Benner, A.; Döhner, H.; Cremer, T.; Lichter, P. Matrix-Based Comparative Genomic Hybridization: Biochips to Screen for Genomic Imbalances. *Genes Chromosomes Cancer* **1997**, *20*, 399–407. [CrossRef]

26. Pinkel, D.; Segraves, R.; Sudar, D.; Clark, S.; Poole, I.; Kowbel, D.; Collins, C.; Kuo, W.-L.; Chen, C.; Zhai, Y.; et al. High Resolution Analysis of DNA Copy Number Variation Using Comparative Genomic Hybridization to Microarrays. *Nat. Genet.* **1998**, *20*, 207–211. [CrossRef] [PubMed]

27. Osoegawa, K.; Mammoser, A.G.; Wu, C.; Frengen, E.; Zeng, C.; Catanese, J.J.; de Jong, P.J. A Bacterial Artificial Chromosome Library for Sequencing the Complete Human Genome. *Genome Res.* **2001**, *11*, 483. [CrossRef]

28. Cowell, J.K.; Nowak, N.J. High-Resolution Analysis of Genetic Events in Cancer Cells Using Bacterial Artificial Chromosome Arrays and Comparative Genome Hybridization. *Adv. Cancer Res.* **2003**, *90*, 91–125. [CrossRef]

29. Malan, V.; Chevallier, S.; Soler, G.; Coubes, C.; Lacombe, D.; Pasquier, L.; Soulier, J.; Morichon-Delvallez, N.; Turleau, C.; Munnich, A.; et al. Array-Based Comparative Genomic Hybridization Identifies a High Frequency of Copy Number Variations in Patients with Syndromic Overgrowth. *Eur. J. Hum. Genet.* **2009**, *18*, 227–232. [CrossRef]

30. Pollack, J.R.; Perou, C.M.; Alizadeh, A.A.; Eisen, M.B.; Pergamenschikov, A.; Williams, C.F.; Jeffrey, S.S.; Botstein, D.; Brown, P.O. Genome-Wide Analysis of DNA Copy-Number Changes Using cDNA Microarrays. *Nat. Genet.* **1999**, *23*, 41–46. [CrossRef]

31. Bashyam; Bair, R.; Kim, Y.H.; Wang, P.; Hernandez-Boussard, T.; Karikari, C.A.; Tibshirani, R.; Maitra, A.; Pollack, J.R. Array-Based Comparative Genomic Hybridization Identifies Localized DNA Amplifications and Homozygous Deletions in Pancreatic Cancer. *Neoplasia* **2005**, *7*, 556-IN16. [CrossRef] [PubMed]

32. Dhami, P.; Coffey, A.J.; Abbs, S.; Vermeesch, J.R.; Dumanski, J.P.; Woodward, K.J.; Andrews, R.M.; Langford, C.; Vetrie, D. Exon Array CGH: Detection of Copy-Number Changes at the Resolution of Individual Exons in the Human Genome. *Am. J. Hum. Genet.* **2005**, *76*, 750. [CrossRef] [PubMed]

33. Schena, M.; Shalon, D.; Heller, R.; Chai, A.; Brown, P.O.; Davis, R.W. Parallel Human Genome Analysis: Microarray-Based Expression Monitoring of 1000 Genes. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 10614–10619. [CrossRef] [PubMed]

34. DeRisi, J.; Penland, L.; Brown, P.O.; Bittner, M.L.; Meltzer, P.S.; Ray, M.; Chen, Y.; Su, Y.A.; Trent, J.M. Use of a cDNA Microarray to Analyse Gene Expression Patterns in Human Cancer. *Nat. Genet.* **1996**, *14*, 457–460. [CrossRef] [PubMed]

35. Lucito, R.; Healy, J.; Alexander, J.; Reiner, A.; Esposito, D.; Chi, M.; Rodgers, L.; Brady, A.; Sebat, J.; Troge, J.; et al. Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation. *Genome Res.* **2003**, *13*, 2291–2305. [CrossRef] [PubMed]

36. Barrett, M.T.; Scheffer, A.; Ben-Dor, A.; Sampas, N.; Lipson, D.; Kincaid, R.; Tsang, P.; Curry, B.; Baird, K.; Meltzer, P.S.; et al. Comparative Genomic Hybridization Using Oligonucleotide Microarrays and Total Genomic DNA. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 17765–17770. [CrossRef]

37. Bignell, G.R.; Huang, J.; Greshock, J.; Watt, S.; Butler, A.; West, S.; Grigorova, M.; Jones, K.W.; Wei, W.; Stratton, M.R.; et al. High-Resolution Analysis of DNA Copy Number Using Oligonucleotide Microarrays. *Genome Res.* **2004**, *14*, 287–295. [CrossRef]

38. Peiffer, D.A.; Le, J.M.; Steemers, F.J.; Chang, W.; Jenniges, T.; Garcia, F.; Haden, K.; Li, J.; Shaw, C.A.; Belmont, J.; et al. High-Resolution Genomic Profiling of Chromosomal Aberrations Using Infinium Whole-Genome Genotyping. *Genome Res.* **2006**, *16*, 1136–1148. [CrossRef]

39. Shen, F.; Huang, J.; Fitch, K.R.; Truong, V.B.; Kirby, A.; Chen, W.; Zhang, J.; Liu, G.; McCarroll, S.A.; Jones, K.W.; et al. Improved Detection of Global Copy Number Variation Using High Density, Non-Polymorphic Oligonucleotide Probes. *BMC Genet.* **2008**, *9*, 27. [CrossRef]

40. Miller, D.T.; Adam, M.P.; Aradhya, S.; Biesecker, L.G.; Brothman, A.R.; Carter, N.P.; Church, D.M.; Crolla, J.A.; Eichler, E.E.; Epstein, C.J.; 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*, 749. [CrossRef] [PubMed]

41. Haraksingh, R.R.; Abyzov, A.; Urban, A.E. Comprehensive Performance Comparison of High-Resolution Array Platforms for Genome-Wide Copy Number Variation (CNV) Analysis in Humans. *BMC Genom.* **2017**, *18*, 321. [CrossRef] [PubMed]

42. Pinto, D.; Darvishi, K.; Shi, X.; Rajan, D.; Rigler, D.; Fitzgerald, T.; Lionel, A.C.; Thiruvahindrapuram, B.; MacDonald, J.R.; Mills, R.; et al. Comprehensive Assessment of Array-Based Platforms and Calling Algorithms for Detection of Copy Number Variants. *Nat. Biotechnol.* **2011**, *29*, 512–520. [CrossRef]

43. Veltman, J.A.; Fridlyand, J.; Pejavar, S.; Olshen, A.B.; Korkola, J.E.; DeVries, S.; Carroll, P.; Kuo, W.-L.; Pinkel, D.; Albertson, D.; et al. Array-Based Comparative Genomic Hybridization for Genome-Wide Screening of DNA Copy Number in Bladder Tumors. *Cancer Res.* **2003**, *63*, 2872–2880. [PubMed]

44. Array-Based Comparative Genomic Hybridization for the Genomewide Detection of Submicroscopic Chromosomal Abnormalities. *Am. J. Hum. Genet.* **2003**, *73*, 1261–1270. [CrossRef] [PubMed]

45. Comparative Genomic Hybridization–Array Analysis Enhances the Detection of Aneuploidies and Submicroscopic Imbalances in Spontaneous Miscarriages. *Am. J. Hum. Genet.* **2004**, *74*, 1168–1174. [CrossRef] [PubMed]

46. Shaw-Smith, C.; Redon, R.; Rickman, L.; Rio, M.; Willatt, L.; Fiegler, H.; Firth, H.; Sanlaville, D.; Winter, R.; Colleaux, L.; et al. Microarray Based Comparative Genomic Hybridisation (array-CGH) Detects Submicroscopic Chromosomal Deletions and Duplications in Patients with Learning Disability/mental Retardation and Dysmorphic Features. *J. Med. Genet.* **2004**, *41*, 241–248. [CrossRef] [PubMed]

47. Schwaenen, C.; Nessling, M.; Wessendorf, S.; Salvi, T.; Wrobel, G.; Radlwimmer, B.; Kestler, H.A.; Haslinger, C.; Stilgenbauer, S.; Döhner, H.; et al. Automated Array-Based Genomic Profiling in Chronic Lymphocytic Leukemia: Development of a Clinical Tool and Discovery of Recurrent Genomic Alterations. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 1039–1044. [CrossRef] [PubMed]

48. :Olshen, A.B.; Venkatraman, E.S.; Lucito, R.; Wigler, M. Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data. *Biostatistics* **2004**, *5*, 557–572. [CrossRef] [PubMed]

49. Jong, K.; Marchiori, E.; van der Vaart, A.; Ylstra, B.; Weiss, M.; Meijer, G. Chromosomal Breakpoint Detection in Human Cancer. In Proceedings of the Applications of Evolutionary Computing, Essex, UK, 14–16 April 2003; Springer: Berlin/Heidelberg, Germany, 2003; pp. 54–65.

50. Hupé, P.; Stransky, N.; Thiery, J.P.; Radvanyi, F.; Barillot, E. Analysis of Array CGH Data: From Signal Ratio to Gain and Loss of DNA Regions. *Bioinformatics* **2004**, *20*, 3413–3422. [CrossRef] [PubMed]

51. Tibshirani, R.; Wang, P. Spatial Smoothing and Hot Spot Detection for CGH Data Using the Fused Lasso. *Biostatistics* **2008**, *9*, 18–29. [CrossRef]

52. Jeng, X.J.; Cai, T.T.; Li, H. Optimal Sparse Segment Identification with Application in Copy Number Variation Analysis. *J. Am. Stat. Assoc.* **2010**, *105*, 1156–1166. [CrossRef] [PubMed]

53. Niu, Y.S.; Zhang, H. The screening and ranking algorithm to detect dna copy number variations. *Ann. Appl. Stat.* **2012**, *6*, 1306–1326. [CrossRef]

54. de Vries, B.B.A.; Pfundt, R.; Leisink, M.; Koolen, D.A.; Vissers, L.E.L.; Janssen, I.M.; van Reijmersdal, S.; Nillesen, W.M.; Huys, E.H.L.P.; de Leeuw, N.; et al. Diagnostic Genome Profiling in Mental Retardation. *Am. J. Hum. Genet.* **2005**, *77*, 606. [CrossRef]

55. Zhao, X.; Li, C.; Guillermo Paez, J.; Chin, K.; Jänne, P.A.; Chen, T.-H.; Girard, L.; Minna, J.; Christiani, D.; Leo, C.; et al. An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. *Cancer Res.* **2004**, *64*, 3060–3071. [CrossRef]

56. Picard, F.; Robin, S.; Lavielle, M.; Vaisse, C.; Daudin, J.-J. A Statistical Approach for Array CGH Data Analysis. *BMC Bioinform.* **2005**, *6*, 27. [CrossRef]

57. Wang, K.; Li, M.; Hadley, D.; Liu, R.; Glessner, J.; Grant, S.F.A.; Hakonarson, H.; Bucan, M. PennCNV: An Integrated Hidden Markov Model Designed for High-Resolution Copy Number Variation Detection in Whole-Genome SNP Genotyping Data. *Genome Res.* **2007**, *17*, 1665–1674. [CrossRef] [PubMed]

58. Dellinger, A.E.; Saw, S.M.; Goh, L.K.; Seielstad, M.; Young, T.L.; Li, Y.J. Comparative Analyses of Seven Algorithms for Copy Number Variant Identification from Single Nucleotide Polymorphism Arrays. *Nucleic Acids Res.* **2010**, *38*, e105. [CrossRef] [PubMed]

59. Roy, S.; Motsinger, R.A. Evaluation of Calling Algorithms for Array-CGH. *Front. Genet.* **2013**, *4*, 4. [CrossRef] [PubMed]

60. Winchester, L.; Yau, C.; Ragoussis, J. Comparing CNV Detection Methods for SNP Arrays. *Brief. Funct. Genom.* **2009**, *8*, 353–366. [CrossRef]

61. Korn, J.M.; Kuruvilla, F.G.; McCarroll, S.A.; Wysoker, A.; Nemesh, J.; Cawley, S.; Hubbell, E.; Veitch, J.; Collins, P.J.; Darvishi, K.; et al. Integrated Genotype Calling and Association Analysis of SNPs, Common Copy Number Polymorphisms and Rare CNVs. *Nat. Genet.* **2008**, *40*, 1253–1260. [CrossRef]

62. Sun, W.; Wright, F.A.; Tang, Z.; Nordgard, S.H.; Van Loo, P.; Yu, T.; Kristensen, V.N.; Perou, C.M. Integrated Study of Copy Number States and Genotype Calls Using High-Density SNP Arrays. *Nucleic Acids Res.* **2009**, *37*, 5365–5377. [CrossRef] [PubMed]

63. Darvishi, K. Application of Nexus Copy Number Software for CNV Detection and Analysis. *Curr. Protoc. Hum. Genet.* **2010**, *65*, 4–14. [CrossRef]

64. Colella, S.; Yau, C.; Taylor, J.M.; Mirza, G.; Butler, H.; Clouston, P.; Bassett, A.S.; Seller, A.; Holmes, C.C.; Ragoussis, J. QuantiSNP: An Objective Bayes Hidden-Markov Model to Detect and Accurately Map Copy Number Variation Using SNP Genotyping Data. *Nucleic Acids Res.* **2007**, *35*, 2013–2025. [CrossRef]

65. 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*, S1. [CrossRef]

66. Korbel, J.O.; Urban, A.E.; Affourtit, J.P.; Godwin, B.; Grubert, F.; Simons, J.F.; Kim, P.M.; Palejev, D.; Carriero, N.J.; Du, L.; et al. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science* **2007**, *318*, 420–426. [CrossRef] [PubMed]

67. Chen, K.; Wallis, J.W.; McLellan, M.D.; Larson, D.E.; Kalicki, J.M.; Pohl, C.S.; McGrath, S.D.; Wendl, M.C.; Zhang, Q.; Locke, D.P.; et al. BreakDancer: An Algorithm for High-Resolution Mapping of Genomic Structural Variation. *Nat. Methods* **2009**, *6*, 677–681. [CrossRef] [PubMed]

68. Korbel, J.O.; Abyzov, A.; Mu, X.J.; Carriero, N.; Cayting, P.; Zhang, Z.; Snyder, M.; Gerstein, M.B. PEMer: A Computational Framework with Simulation-Based Error Models for Inferring Genomic Structural Variants from Massive Paired-End Sequencing Data. *Genome Biol.* **2009**, *10*, R23. [CrossRef] [PubMed]

69. Lee, S.; Hormozdiari, F.; Alkan, C.; Brudno, M. MoDIL: Detecting Small Indels from Clone-End Sequencing with Mixtures of Distributions. *Nat. Methods* **2009**, *6*, 473–474. [CrossRef] [PubMed]

70. Hayes, M.; Pyon, Y.S.; Li, J. A Model-Based Clustering Method for Genomic Structural Variant Prediction and Genotyping Using Paired-End Sequencing Data. *PLoS ONE* **2012**, *7*, e52881. [CrossRef]

71. Marschall, T.; Costa, I.G.; Canzar, S.; Bauer, M.; Klau, G.W.; Schliep, A.; Schönhuth, A. CLEVER: Clique-Enumerating Variant Finder. *Bioinformatics* **2012**, *28*, 2875–2882. [CrossRef] [PubMed]

72. Trappe, K.; Emde, A.-K.; Ehrlich, H.-C.; Reinert, K. Gustaf: Detecting and Correctly Classifying SVs in the NGS Twilight Zone. *Bioinformatics* **2014**, *30*, 3484–3490. [CrossRef] [PubMed]

73. Ye, K.; Schulz, M.H.; 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*, 2865–2871. [CrossRef] [PubMed]

74. 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*, 1586–1592. [CrossRef] [PubMed]

75. Xie, C.; Tammi, M.T. CNV-Seq, a New Method to Detect Copy Number Variation Using High-Throughput Sequencing. *BMC Bioinform.* **2009**, *10*, 80. [CrossRef] [PubMed]

76. Gusnanto, A.; Taylor, C.C.; Nafisah, I.; Wood, H.M.; Rabbitts, P.; Berri, S. Estimating Optimal Window Size for Analysis of Low-Coverage next-Generation Sequence Data. *Bioinformatics* **2014**, *30*, 1823–1829. [CrossRef]

77. Benjamini, Y.; Speed, T.P. Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing. *Nucleic Acids Res.* **2012**, *40*, e72. [CrossRef]

78. Talevich, E.; Hunter Shain, A.; Botton, T.; Bastian, B.C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **2016**, *12*, e1004873. [CrossRef]

79. Abyzov, A.; Urban, A.E.; 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*, 974–984. [CrossRef] [PubMed]

80. Dharanipragada, P.; Vogeti, S.; Parekh, N. iCopyDAV: Integrated Platform for Copy Number variations—Detection, Annotation and Visualization. *PLoS ONE* **2018**, *13*, e0195334. [CrossRef]

81. Wang, W.; Wang, W.; Sun, W.; Crowley, J.J.; Szatkiewicz, J.P. Allele-Specific Copy-Number Discovery from Whole-Genome and Whole-Exome Sequencing. *Nucleic Acids Res.* **2015**, *43*, e90. [CrossRef]

82. Xi, R.; Lee, S.; Xia, Y.; Kim, T.-M.; Park, P.J. Copy Number Analysis of Whole-Genome Data Using BIC-seq2 and Its Application to Detection of Cancer Susceptibility Variants. *Nucleic Acids Res.* **2016**, *44*, 6274–6286. [CrossRef]

83. Boeva, V.; Zinovyev, A.; Bleakley, K.; Vert, J.P.; Janoueix-Lerosey, I.; Delattre, O.; Barillot, E. Control-Free Calling of Copy Number Alterations in Deep-Sequencing Data Using GC-Content Normalization. *Bioinformatics* **2011**, *27*, 268–269. [CrossRef]

84. Miller, C.A.; Hampton, O.; Coarfa, C.; Milosavljevic, A. ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads. *PLoS ONE* **2011**, *6*, e16327. [CrossRef]

85. Gordeeva, V.; Sharova, E.; Babalyan, K.; Sultanov, R.; Govorun, V.M.; Arapidi, G. Benchmarking Germline CNV Calling Tools from Exome Sequencing Data. *Sci. Rep.* **2021**, *11*, 14416. [CrossRef]

86. Fromer, M.; Moran, J.L.; Chambert, K.; Banks, E.; Bergen, S.E.; Ruderfer, D.M.; Handsaker, R.E.; McCarroll, S.A.; O'Donovan, M.C.; Owen, M.J.; et al. Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth. *Am. J. Hum. Genet.* **2012**, *91*, 597–607. [CrossRef] [PubMed]

87. Jiang, Y.; Oldridge, D.A.; Diskin, S.J.; Zhang, N.R. CODEX: A Normalization and Copy Number Variation Detection Method for Whole Exome Sequencing. *Nucleic Acids Res.* **2015**, *43*, e39. [CrossRef]

88. Plagnol, V.; Curtis, J.; Epstein, M.; Mok, K.Y.; Stebbings, E.; Grigoriadou, S.; Wood, N.W.; Hambleton, S.; Burns, S.O.; Thrasher, A.J.; et al. A Robust Model for Read Count Data in Exome Sequencing Experiments and Implications for Copy Number Variant Calling. *Bioinformatics* **2012**, *28*, 2747–2754. [CrossRef]

89. Love, M.I.; Myšičková, A.; Sun, R.; Kalscheuer, V.; Vingron, M.; Haas, S.A. Modeling Read Counts for CNV Detection in Exome Sequencing Data. *Stat. Appl. Genet. Mol. Biol.* **2011**, *10*. [CrossRef]

90. 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*, e154. [CrossRef] [PubMed]

91. Kuśmirek, W.; Szmurło, A.; Wiewiórka, M.; Nowak, R.; Gambin, T. Comparison of kNN and K-Means Optimization Methods of Reference Set Selection for Improved CNV Callers Performance. *BMC Bioinform.* **2019**, *20*, 266. [CrossRef]

92. Klambauer, G.; Schwarzbauer, K.; Mayr, A.; Clevert, D.-A.; Mitterecker, A.; Bodenhofer, U.; Hochreiter, S. 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*, e69. [CrossRef] [PubMed]

93. Krumm, N.; Sudmant, P.H.; Ko, A.; O'Roak, B.J.; Malig, M.; Coe, B.P.; Quinlan, A.R.; Nickerson, D.A.; Eichler, E.E. Copy Number Variation Detection and Genotyping from Exome Sequence Data. *Genome Res.* **2012**, *22*, 1525–1532. [CrossRef]

94. Johansson, L.F.; van Dijk, F.; de Boer, E.N.; van Dijk-Bos, K.K.; Jongbloed, J.D.; van der Hout, A.H.; Westers, H.; Sinke, R.J.; Swertz, M.A.; Sijmons, R.H.; et al. CoNVaDING: Single Exon Variation Detection in Targeted NGS Data. *Hum. Mutat.* **2016**, *37*, 457–464. [CrossRef] [PubMed]

95. Fowler, A.; Mahamdallie, S.; Ruark, E.; Seal, S.; Ramsay, E.; Clarke, M.; Uddin, I.; Wylie, H.; Strydom, A.; Lunter, G.; et al. Accurate Clinical Detection of Exon Copy Number Variants in a Targeted NGS Panel Using DECoN. *Wellcome Open Res.* **2016**, *1*, 20. [CrossRef] [PubMed]

96. Pevzner, P.A.; Tang, H.; Waterman, M.S. An Eulerian Path Approach to DNA Fragment Assembly. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 9748–9753. [CrossRef] [PubMed]

97. Nijkamp, J.F.; van den Broek, M.A.; Geertman, J.-M.A.; Reinders, M.J.T.; Daran, J.-M.G.; de Ridder, D. De Novo Detection of Copy Number Variation by Co-Assembly. *Bioinformatics* **2012**, *28*, 3195–3202. [CrossRef]

98. Iqbal, Z.; Caccamo, M.; Turner, I.; Flicek, P.; McVean, G. De Novo Assembly and Genotyping of Variants Using Colored de Bruijn Graphs. *Nat. Genet.* **2012**, *44*, 226–232. [CrossRef] [PubMed]

99. Cameron, D.L.; Schröder, J.; Penington, J.S.; Do, H.; Molania, R.; Dobrovic, A.; Speed, T.P.; Papenfuss, A.T. GRIDSS: Sensitive and Specific Genomic Rearrangement Detection Using Positional de Bruijn Graph Assembly. *Genome Res.* **2017**, *27*, 2050–2060. [CrossRef]

100. Wang, J.; Mullighan, C.G.; Easton, J.; Roberts, S.; Heatley, S.L.; Ma, J.; Rusch, M.C.; Chen, K.; Harris, C.C.; Ding, L.; et al. CREST Maps Somatic Structural Variation in Cancer Genomes with Base-Pair Resolution. *Nat. Methods* **2011**, *8*, 652–654. [CrossRef] [PubMed]

101. Using ERDS to Infer Copy-Number Variants in High-Coverage Genomes. *Am. J. Hum. Genet.* **2012**, *91*, 408–421. [CrossRef] [PubMed]

102. Layer, R.M.; Chiang, C.; Quinlan, A.R.; Hall, I.M. LUMPY: A Probabilistic Framework for Structural Variant Discovery. *Genome Biol.* **2014**, *15*, e1004572. [CrossRef]

103. Chen, X.; Schulz-Trieglaff, O.; Shaw, R.; Barnes, B.; Schlesinger, F.; Källberg, M.; Cox, A.J.; Kruglyak, S.; Saunders, C.T. Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications. *Bioinformatics* **2015**, *32*, 1220–1222. [CrossRef] [PubMed]

104. Handsaker, R.E.; Korn, J.M.; Nemesh, J.; McCarroll, S.A. Discovery and Genotyping of Genome Structural Polymorphism by Sequencing on a Population Scale. *Nat. Genet.* **2011**, *43*, 269–276. [CrossRef]

105. Rausch, T.; Zichner, T.; Schlattl, A.; Stütz, A.M.; Benes, V.; Korbel, J.O. DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis. *Bioinformatics* **2012**, *28*, i333–i339. [CrossRef]

106. Quinlan, A.R.; Clark, R.A.; Sokolova, S.; Leibowitz, M.L.; Zhang, Y.; Hurles, M.E.; Mell, J.C.; Hall, I.M. Genome-Wide Mapping and Assembly of Structural Variant Breakpoints in the Mouse Genome. *Genome Res.* **2010**, *20*, 623–635. [CrossRef]

107. Mohiyuddin, M.; Mu, J.C.; Li, J.; Bani Asadi, N.; Gerstein, M.B.; Abyzov, A.; Wong, W.H.; Lam, H.Y.K. MetaSV: An Accurate and Integrative Structural-Variant Caller for next Generation Sequencing. *Bioinformatics* **2015**, *31*, 2741–2744. [CrossRef]

108. Michaelson, J.J.; Sebat, J. forestSV: Structural Variant Discovery through Statistical Learning. *Nat. Methods* **2012**, *9*, 819–821. [CrossRef]

109. Parikh, H.; Mohiyuddin, M.; Lam, H.Y.K.; Iyer, H.; Chen, D.; Pratt, M.; Bartha, G.; Spies, N.; Losert, W.; Zook, J.M.; et al. Svclassify: A Method to Establish Benchmark Structural Variant Calls. *BMC Genom.* **2016**, *17*, 64. [CrossRef] [PubMed]

110. Cai, L.; Wu, Y.; Gao, J. DeepSV: Accurate Calling of Genomic Deletions from High-Throughput Sequencing Data Using Deep Convolutional Neural Network. *BMC Bioinform.* **2019**, *20*, 665. [CrossRef]

111. Mills, R.E.; Walter, K.; Stewart, C.; Handsaker, R.E.; Chen, K.; Alkan, C.; Abyzov, A.; Yoon, S.C.; Ye, K.; Cheetham, R.K.; et al. Mapping Copy Number Variation by Population-Scale Genome Sequencing. *Nature* **2011**, *470*, 59–65. [CrossRef]

112. Sudmant, P.H.; Rausch, T.; Gardner, E.J.; Handsaker, R.E.; Abyzov, A.; Huddleston, J.; Zhang, Y.; Ye, K.; Jun, G.; Fritz, M.H.-Y.; et al. An Integrated Map of Structural Variation in 2504 Human Genomes. *Nature* **2015**, *526*, 75–81. [CrossRef]

113. Collins, R.L.; Brand, H.; Karczewski, K.J.; Zhao, X.; Alföldi, J.; Francioli, L.C.; Khera, A.V.; Lowther, C.; Gauthier, L.D.; Wang, H.; et al. A Structural Variation Reference for Medical and Population Genetics. *Nature* **2020**, *581*, 444–451. [CrossRef] [PubMed]

114. Cretu Stancu, M.; van Roosmalen, M.J.; Renkens, I.; Nieboer, M.M.; Middelkamp, S.; de Ligt, J.; Pregno, G.; Giachino, D.; Mandrile, G.; Espejo Valle-Inclan, J.; et al. Mapping and Phasing of Structural Variation in Patient Genomes Using Nanopore Sequencing. *Nat. Commun.* **2017**, *8*, 1326. [CrossRef] [PubMed]

115. English, A.C.; Salerno, W.J.; Reid, J.G. PBHoney: Identifying Genomic Variants via Long-Read Discordance and Interrupted Mapping. *BMC Bioinform.* **2014**, *15*, 180. [CrossRef]

116. Jiang, T.; Liu, S.; Cao, S.; Liu, Y.; Cui, Z.; Wang, Y.; Guo, H. Long-Read Sequencing Settings for Efficient Structural Variation Detection Based on Comprehensive Evaluation. *BMC Bioinform.* **2021**, *22*, 552. [CrossRef] [PubMed]

117. Spies, N.; Weng, Z.; Bishara, A.; McDaniel, J.; Catoe, D.; Zook, J.M.; Salit, M.; West, R.B.; Batzoglou, S.; Sidow, A. Genome-Wide Reconstruction of Complex Structural Variants Using Read Clouds. *Nat. Methods* **2017**, *14*, 915–920. [CrossRef]

118. Elyanow, R.; Wu, H.T.; Raphael, B.J. Identifying Structural Variants Using Linked-Read Sequencing Data. *Bioinformatics* **2018**, *34*, 353–360. [CrossRef]
119. Hills, M.; O'Neill, K.; Falconer, E.; Brinkman, R.; Lansdorp, P.M. BAIT: Organizing Genomes and Mapping Rearrangements in Single Cells. *Genome Med.* **2013**, *5*, 82. [CrossRef]
120. Wang, S.; Lee, S.; Chu, C.; Jain, D.; Kerpedjiev, P.; Nelson, G.M.; Walsh, J.M.; Alver, B.H.; Park, P.J. HiNT: A Computational Method for Detecting Copy Number Variations and Translocations from Hi-C Data. *Genome Biol.* **2020**, *21*, 73. [CrossRef]
121. Schwartz, D.C.; Li, X.; Hernandez, L.I.; Ramnarain, S.P.; Huff, E.J.; Wang, Y.K. Ordered Restriction Maps of Saccharomyces Cerevisiae Chromosomes Constructed by Optical Mapping. *Science* **1993**, *262*, 110–114. [CrossRef]
122. Latreille, P.; Norton, S.; Goldman, B.S.; Henkhaus, J.; Miller, N.; Barbazuk, B.; Bode, H.B.; Darby, C.; Du, Z.; Forst, S.; et al. Optical Mapping as a Routine Tool for Bacterial Genome Sequence Finishing. *BMC Genom.* **2007**, *8*, 321. [CrossRef]
123. Pendleton, M.; Sebra, R.; Pang, A.W.C.; Ummat, A.; Franzen, O.; Rausch, T.; Stütz, A.M.; Stedman, W.; Anantharaman, T.; Hastie, A.; et al. Assembly and Diploid Architecture of an Individual Human Genome via Single-Molecule Technologies. *Nat. Methods* **2015**, *12*, 780–786. [CrossRef] [PubMed]
124. Optical Genome Mapping Enables Constitutional Chromosomal Aberration Detection. *Am. J. Hum. Genet.* **2021**, *108*, 1409–1422. [CrossRef]
125. Li, L.; Leung, A.K.-Y.; Kwok, T.-P.; Lai, Y.Y.Y.; Pang, I.K.; Chung, G.T.-Y.; Mak, A.C.Y.; Poon, A.; Chu, C.; Li, M.; et al. OMSV Enables Accurate and Comprehensive Identification of Large Structural Variations from Nanochannel-Based Single-Molecule Optical Maps. *Genome Biol.* **2017**, *18*, 230. [CrossRef] [PubMed]
126. Zhang, Z.; Cheng, H.; Hong, X.; Di Narzo, A.F.; Franzen, O.; Peng, S.; Ruusalepp, A.; Kovacic, J.C.; Bjorkegren, J.L.M.; Wang, X.; et al. EnsembleCNV: An Ensemble Machine Learning Algorithm to Identify and Genotype Copy Number Variation Using SNP Array Data. *Nucleic Acids Res.* **2019**, *47*, e39. [CrossRef] [PubMed]
127. Pounraja, V.K.; 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*, 1134–1143. [CrossRef]
128. Akbarinejad, S.; Hadadian Nejad Yousefi, M.; Goudarzi, M. SVNN: An Efficient PacBio-Specific Pipeline for Structural Variations Calling Using Neural Networks. *BMC Bioinform.* **2021**, *22*, 335. [CrossRef]
129. 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*, 1–18. [CrossRef]
130. Zook, J.M.; Hansen, N.F.; Olson, N.D.; Chapman, L.; Mullikin, J.C.; Xiao, C.; Sherry, S.; Koren, S.; Phillippy, A.M.; Boutros, P.C.; et al. A Robust Benchmark for Detection of Germline Large Deletions and Insertions. *Nat. Biotechnol.* **2020**, *38*, 1347–1355. [CrossRef]