# Structural variation in the sequencing era

*Steve S. Ho*[1], *Alexander E. Urban*[2,3] *and Ryan E. Mills*[1,4]*

Abstract | Identifying structural variation (SV) is essential for genome interpretation but has been historically difficult due to limitations inherent to available genome technologies. Detection methods that use ensemble algorithms and emerging sequencing technologies have enabled the discovery of thousands of SVs, uncovering information about their ubiquity, relationship to disease and possible effects on biological mechanisms. Given the variability in SV type and size, along with unique detection biases of emerging genomic platforms, multiplatform discovery is necessary to resolve the full spectrum of variation. Here, we review modern approaches for investigating SVs and proffer that, moving forwards, studies integrating biological information with detection will be necessary to comprehensively understand the impact of SV in the human genome.

Structural variations
(SVs). Operationally defined as sequence variants >50 bp in size. The most recognized forms of SV include deletions, duplications, inversions, insertions and translocations.

Complex rearrangements
A structural variant that consists of multiple combinations of structural variant types nested or clustered with one another.

Read signatures
Specific marks that result from reads that map discordantly to the reference genome.

[1]*Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA.*

[2]*Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA.*

[3]*Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA.*

[4]*Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA.*

**e-mail: remills@umich.edu*

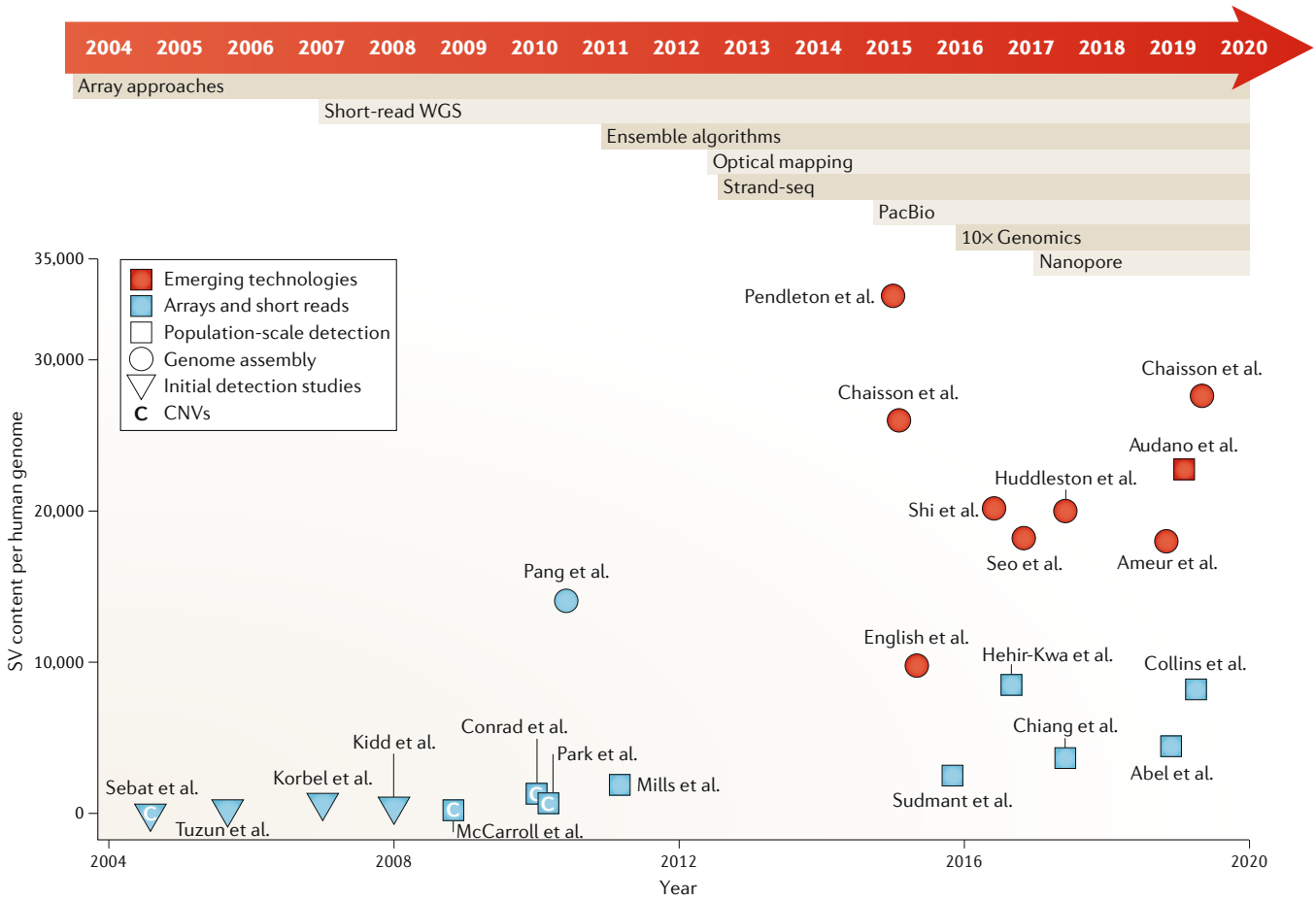
https://doi.org/10.1038/s41576-019-0180-9

Widespread application of whole-genome high-throughput sequencing (HTS) for the detection of genetic variants has shown that differences between individuals are typically present as single-nucleotide variants (SNVs), small insertions and deletions (indels; <50 bp), and structural variations (SVs)[1]. SVs are extremely diverse in type and size, ranging anywhere from ~50 bp to well over megabases of sequence, affecting more of the genome per nucleotide changes than any other class of sequence variant[2–6]. They comprise a myriad of subclasses that consist of unbalanced copy number variants (CNVs), which include deletions, duplications and insertions of genetic material, as well as balanced rearrangements, such as inversions and interchromosomal and intrachromosomal translocations. Additionally, SVs include mobile element insertions, multi-allelic CNVs of highly variable copy number, segmental duplications and complex rearrangements that consist of multiple combinations of these described events. SVs are present in every human genome and affect molecular and cellular processes, regulatory functions, 3D structure and transcriptional machinery[5,7,8]. Thus, increasing our knowledge of SV structure and prevalence is necessary to discern the genomics of physiological and pathophysiological processes.

Many of the prevalent tools and algorithms to detect SVs use short read signatures to infer the presence of SVs compared with a reference genome[9]. Although short-read approaches are highly effective at resolving SNVs, SV detection is unable to completely overcome the limited sequence and insert sizes of standard short-read HTS[10]. There are still considerable limitations on what can be achieved in SV analysis owing to technical difficulties in resolving exact structures of SVs given their substantial diversity and proximity to repetitive regions[5,9,11–13]. SNVs detected by short-read technologies can be sequence-resolved during the discovery stage owing to their smaller size, whereas most SVs require computational inference post hoc as they span multiple short reads. Because of this, the degree to which contemporary genomics has studied SNVs compared with SVs is significantly skewed. Specifically, standardized best practices, robust detection platforms, high-quality reference sets and extensive functional data from genome-wide association studies are available for SNV research[14–20]. Comparatively, progress in SV analysis is notably lagging behind, as detection is suboptimal and reference sets are lacking in diversity, sample size and depth.

A considerable increase in the development and availability of novel sequencing technologies that leverage specialized flow cells, advanced microfluidics and protein pores, among others, has led to platforms that produce reads several orders of magnitude longer than those generated from short-read HTS, enabling the direct detection of many previously indiscernible SVs[21] (BOX 1). In this Review, we discuss methods for resolving SVs in human genomes that bypass the limitations of individual short-read approaches through algorithmic ensembles and by leveraging new technologies. In particular, we discuss the findings of applying new technologies to genome assembly and population-scale variant mapping as they relate to germline SVs (for recent reviews on somatic SVs, see REFS[22,23]). Along with integrating SV algorithms, we consider integrating data generated from multiple genomic platforms as a way to comprehensively detect the broad range of SVs. As each

Box 1 | **From microarrays to short-read sequencing and beyond**




The prevalence of structural variation (SV) in human genomes has historically been determined by the resolution of available technologies. Molecular cytogenetics techniques, particularly chromosome banding and fluorescence in situ hybridization, powered seminal work involving the detection of microscopic chromosomal aberrations but were unable to identify submicroscopic variants (for brief historical perspectives on cytogenetic-based SV detection, see REFS[22,155]). Microarrays then became the primary technology to identify copy number variants (CNVs) in the 2000s due to improved resolution over karyotype-based analysis. Array-comparative genomic hybridization enabled the first reports of global SV, identifying ~300 copy number-variable loci and informing the wide presence of SVs in phenotypically normal human genomes[56,156]. One of the first sequence mapping approaches performed with a single fosmid library reported a similar number of SVs, ~300 variants[11]. These numbers were highly preliminary as SNP arrays would soon detect 1,447 and 1,320 CNVs across 270 individuals[157,158]. At this time, sequencing-based approaches were dropping in cost — their proof-of-principle studies exhibited similar sensitivity compared with arrays but with significantly fewer samples; one study employed paired-end 454 pyrosequencing in two human genomes[24] and another used a fosmid-clone-based mapping approach in nine human genomes[159] to detect ~1,700 and ~1,300 SVs, respectively. Large, population-scale detection efforts then started to emerge. In 2010, high-density microarrays employing millions of probes ascertained 11,700 CNVs across 450 individuals[2]. A sequencing based-approach proved to be more comprehensive in 2011; this study applied an ensemble approach to ~4× short-read high-throughput sequencing (HTS) of 185 individuals to detect a three-fold increase of SVs in comparison[4].

Throughout these studies, two main advantages made short-read HTS superior to microarrays for exhaustive SV detection: the detection of balanced variants and sequences not in the reference (novel insertions), which are missed by arrays; and higher overall resolution. Thus, short-read HTS has been the major driver of progress in SV detection over the past decade given its improved sensitivity over array platforms, although arrays are still regularly used for their low cost and high throughput[160]. Improvements in short-read technology have enabled the detection of millions of variants, improving the number of detectable SVs from ~2,100 to ~8,000 SVs per human genome[5,43]. The emerging sequencing technologies discussed in this Review push these estimates further, to >25,000 SVs per individual. Shown (see the figure) are selected studies that either estimate the extent of SV content or provide estimates of detectable SVs according to technology within phenotypically healthy human genomes, showing the relationship between detectable SVs and available technologies.

For a more comprehensive overview of the methods and algorithms used to detect SVs before adoption of the technologies discussed in this Review, we suggest the following references: molecular cytogenetics techniques, REF.[161]; the application of molecular cytogenetics to understand clinical disorders, REF.[162]; array and clone-based approaches to detect SVs, REF.[155]; a comprehensive survey of the first SV detection studies, REF.[163]; short-read discovery and genotyping, REFS[9,164,165]; detecting complex SVs, REF.[166]; and clinically relevant CNVs and SV detection from whole-exome sequencing, REFS[167–169].

Citations for the studies listed in the figure: REFS[2,4,5,11,24,39,43–45,61,95,97,102–105,108,142,156,158,159,170,171]. CNV, copy number variant; PacBio, Pacific Biosciences; SV, structural variation; WGS, whole-genome sequencing.

Short-read HTS
(Short-read high-throughput sequencing). Standard sequencing where libraries are fragmented to ~600–800 bp in length. Two ends are sequenced ~100–250 bp with an unsequenced insert size of ~100–600 bp.

Flow cells
Glass slides containing fluidic channels for sequencing reactions to occur.

Microfluidics
Devices that precisely manipulate and control small amounts of fluids.

SV callers
An algorithm designed to detect structural variations (SVs). Each putative SV detected by a caller is an individual 'call'. 'Call' derives from computer science, meaning to invoke a particular task; detected SVs are the result of each performed 'task'.

Sensitivity
The ability to detect known variants correctly. Low sensitivity implies low ability to detect bona fide variants.

Reference data sets
High-resolution structural variation data sets typically derived from de novo genome assemblies, population-scale sequencing or projects employing multiple orthogonal detection methods. Reference sets are used to benchmark detection algorithms and determine the novelty and rarity of structural variation calls.

Ensemble algorithm
A detection method that combines the resulting call sets from multiple independent algorithms.

False-discovery rate
The expected number of calls that should be false but are marked as true within the final call set.

Coordinate overlap
The number of base pairs that are identical between two different variant calls.

approach has different strengths, we highlight individual strategies, their applications and recent findings. We discuss future directions and consider incorporating multimodal biological information as a way to interpret the impact of SVs in their molecular contexts.

## Ensemble algorithms

Sequencing-based SV detection leverages primarily signatures that result from mapping discordance between a sample read and the reference genome: read-pair approaches assess the orientation and distance of paired ends; read-depth methods detect deletions or duplications based on divergences in mapping depth; split-read approaches leverage alignments that map over SV breakpoints; and, alternatively, de novo or local assembly reassembles contigs before pairwise comparison with a reference[24–26]. Many early SV callers, such as BreakDancer[27], CNVnator[28] and PEMer[29], specialized in leveraging only one of four approaches, which inherently limits detection (reviewed in REF.[9]). Hybrid-signature algorithms, such as DELLY[30], Genome STRiP[31], LUMPY[32] and Manta[33], among others[34–36], mitigate the limited scope of single-approach algorithms, improving sensitivity by integrating two or more disparate signatures to call putative SVs based on combined supporting evidence. However, even with signal integration, no individual caller has been shown to be capable of identifying the complete range of SV owing to the large diversity in viable detection approaches and the variability in SV subtype and size[37–39]. One strategy to attenuate this issue involves detecting SVs using multiple discrete algorithms on the same sequence data and integrating the variant calls to generate a unified call set (FIG. 1A). Combining multiple algorithms improves detection by leveraging the different heuristic approaches of each individual caller and has been shown to increase the concordance of SV calls when compared with reference data sets (BOX 2, TABLE 1) developed by large consortium projects[40–42]. From here onwards, we refer to an 'ensemble algorithm' as the combination and integration of multiple independent SV detection algorithms.

Most ensemble algorithms have been developed in-house, meaning the combination of algorithms and heuristic filters are unique to individual projects and thus non-standardized. However, one or several algorithms are typically used to cover each signature type; for example, CNVnator can be combined with BreakDancer and Pindel to cover read depth, read pair and split reads, respectively, although recent approaches use hybrid-signature callers as well. Following multi-algorithm detection, the resultant calls are merged, combining potentially duplicate SVs with delineating SVs called uniquely by each algorithm. The methods to integrate, combine and score calls vary markedly between studies and thus far have used breakpoint confidence interval overlap, breakpoint distances, false-discovery rate (FDR) cut-off thresholds, read-signature prioritization (split reads > read pair > read depth), caller concordance and supporting signatures' thresholds[4,5,43–46] (FIG. 1B). A seventh factor, coordinate overlap, is considered by all ensemble algorithm methods to varying degrees. Depending on the level of sensitivity a project aims to achieve, applications will either intersect calls or take a union, decreasing and increasing sensitivity while decreasing and increasing the FDR, respectively.

Stand-alone tools for ensemble algorithms help standardize these integrative pipelines. SpeedSeq[47] employs LUMPY and CNVnator to cover split-read, paired-end and read-depth detection before validating calls with a Bayesian likelihood genotyper (SVTyper), an approach that is also implemented in the population scale-specific svtools[48]. HugeSeq[41], iSVP[49], Parliament2 (REF.[50]) and SVMerge[40] are ensemble algorithm callers that primarily intersect by coordinate overlap, which require that a call is detected by multiple callers, whereas MetaSV[51] takes the union and does not require caller overlap. SVMerge and MetaSV both validate their consensus calls with local reassembly, but MetaSV prioritizes SV signatures with higher resolution (for example, split reads over read pairs). Parliament2 allows users to decide on a combination of six short-read algorithms before merging calls with SURVIVOR[52] and genotyping with SVTyper[47]. Ensemble algorithm callers are beginning to implement meta-level heuristics to improve precision beyond using simple overlap: Parliament2 scores each SV call with a caller concordance metric trained on the HG002 (also known as NA24385) reference genome[50]; FusorSV[53] implements a data-mining method to learn how well different SV algorithms perform compared with a truth set to promote the most complementary and highest performing ensemble; and CN-Learn[54], an algorithm for whole-exome data, extracts features from a truth set and uses these features to train a random forest classifier that differentiates CNV calls as true or false.

### *Population-scale SV detection*

Ensemble algorithm approaches have been widely used in studies characterizing SV across populations. The 1000 Genomes Project (1KGP) initially integrated 19 algorithms to detect SVs in European, Han, Japanese and Yoruban individuals to create a sequencing-based SV reference map[4]. This early work provided one of the first frameworks for using ensemble approaches to detect SVs at the population scale and revealed 51 SV hotspots in the genome, 80% of which were dominated by a single formation mechanism, non-allelic homologous recombination, some at loci associated with known genetic conditions. At the completion of phase 3, the 1KGP had sequenced 2,504 individuals across 26 populations and investigated all major SV classes in contrast to the deletion focus of the phase 1 marker paper[5] (TABLE 1). The authors generated one of the most comprehensive and diverse reference sets of human SVs to date and estimated that typical human genomes contain between 2,100 and 2,500 SVs that affect ~20 million nucleotides (BOX 1). Moreover, they found that SVs are enriched for expression quantitative trait loci (eQTLs) up to 50-fold compared with SNVs. Although the 1KGP set the stage for large-scale SV detection by sequencing, the fairly low coverage (~6–7×) per sample limited power to detect rare variants[55].

SV projects with larger and deeper data sets have emerged to improve on the 1KGP reference set. One study applied svtools to ~18,000 human genomes,

Purifying selection
A process of natural selection where strongly deleterious alleles are selectively removed from a population.

detecting 118,973 and 241,426 SVs from data sets aligned to Genome Reference Consortium Human Build 37 (GRCh37) and GRCh38, respectively[44]. The authors estimated a mean of 4,442 high-confidence SVs per human genome (BOX 1) and notably found that: ~4 out of 4,442 SVs alter exons directly; and ~19 out of 4,442 SVs are rare non-coding deletions that, using predictive functional annotation, were up to 800 times more likely to be strongly deleterious than rare SNVs, exhibiting levels of purifying selection comparable with those of small loss-of-function variants. To improve rare SV detection, the Genome Aggregation Database (gnomAD) systematically processed data from fewer individuals (~15,000) but at increased mean coverage (~32× versus 20×)[43]. The authors detected 498,257 SVs from an ensemble of four algorithms, finding an average of 8,202 SVs per human genome (BOX 1) non-uniformly distributed throughout the genome by SV subclass. This study revealed that 253 out of 8,202 SVs in the average genome are intragenic and eight out of 8,202 are rare SVs that

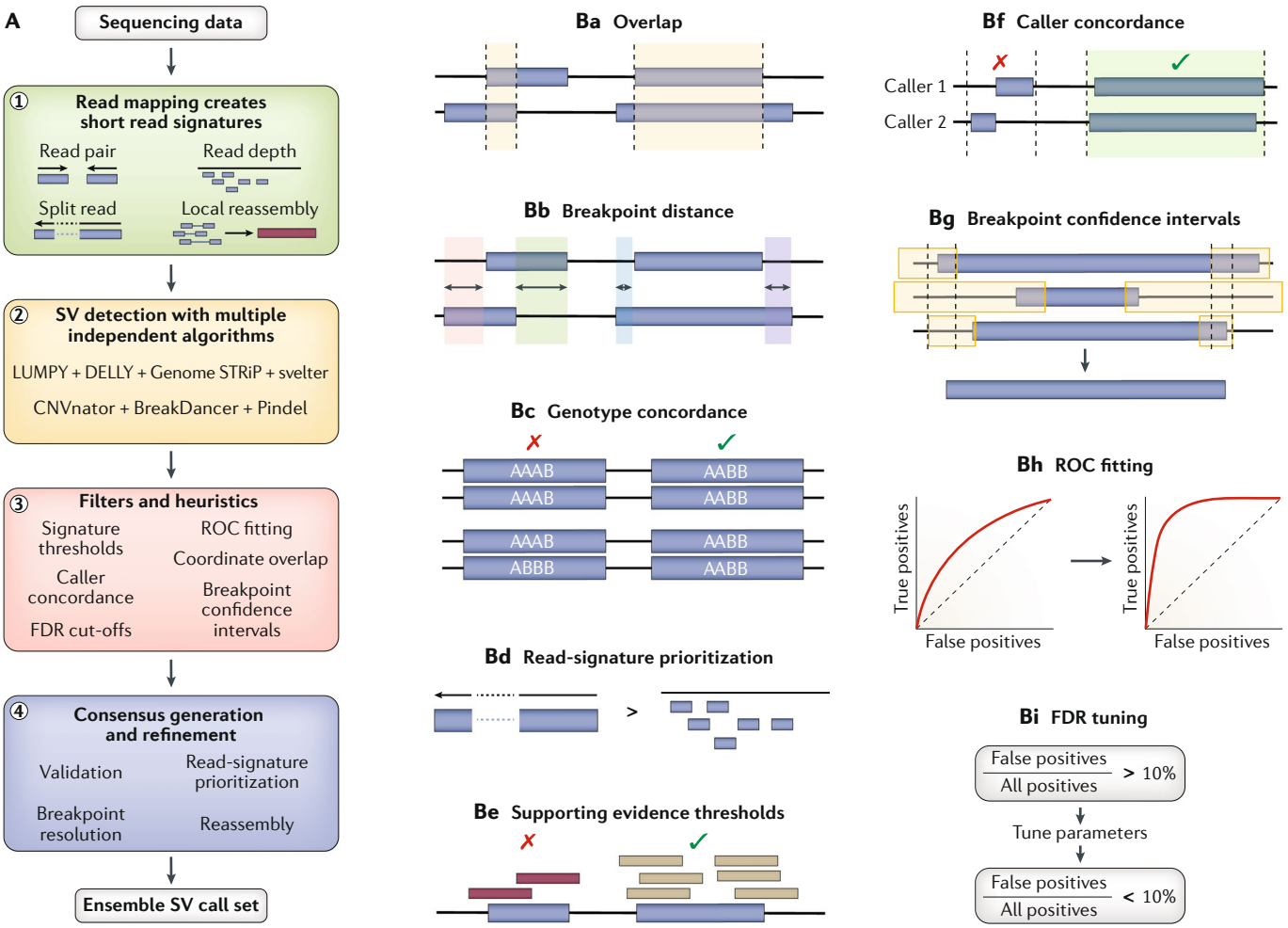



Fig. 1 | **Overview of ensemble algorithms. A** | Flowchart outlining the major steps in an ensemble algorithm. Step 1, discordantly mapped reads result in signatures that are used to infer structural variations (SVs). Step 2, multiple independent algorithms detect SVs in parallel. Step 3, filters and heuristics based on the project aims are applied to remove false-positives and merge calls. Step 4, final decisions are made to designate and preserve high-confidence calls, and they are output as a consolidated list of putative variants. **B** | Factors in integrating SV calls. As detection methods vary substantially in their resolution and approach, a large variety of heuristics have been applied to merge calls derived from different algorithms. **Ba** | Almost all integration methods consider the immediate intuitive option, overlap, with a common requirement of 50% reciprocity. Overlap analysis can require a minimum or maximum length difference between the called SVs to improve stringency. Alternatively to coordinate overlap, one can use sequence similarity, as employed by the Genome in a Bottle consortium[60]. **Bb** | Computing the distance between breakpoints as opposed to overlap is useful for higher-resolution methods such as split-read analysis. **Bc** | Algorithms may require that calls to be merged have consistent genotypes for additional accuracy. **Bd** | Read signatures are often prioritized such that if two calls overlap, the call supported with a higher-resolution read signature is chosen. **Be** | Calls may be required to have support from a minimum number of reads containing a given signature before merging. **Bf** | Intersection, or caller concordance, requires that calls are detected by a minimum number of multiple algorithms, most often two. This opposes taking the union of calls, which requires no caller overlap. **Bg** | Breakpoint confidence intervals were estimated by local reassembly in the 1000 Genomes Project phase 1 (REF.[4]) and by comparisons with high-quality long-read SVs[39]. In both studies, calls were merged if their breakpoint confidence intervals overlapped. **Bh** | Parameters of individual callers can be adjusted to better fit a receiver operating characteristic (ROC) curve by benchmarking against a truth set of choice, although high-confidence calls within a given call set have also been used as a benchmark[43]. **Bi** | Projects with orthogonal data can adjust caller parameters to keep the false-discovery rate (FDR) at a certain threshold (typically <10%) before merging calls[5]. These factors and techniques have been primarily considered for short-read integration but they carry over to multiplatform approaches as well.

Box 2 | **Structural variation reference sets**

Reference data sets are essential for the development of structural variation (SV) discovery methods. Many algorithms validate detection ability by benchmarking against or training with data sets released by population-scale sequencing, de novo genome assemblies or projects that perform comprehensive discovery with multiple orthogonal platforms[5,39,43–45,58,60,61,75,97,103–105,107,108,119,142,172]. The type of chosen reference sets should be appropriate for each application; for example, highly curated discovery sets are appropriate for benchmarking detection methods, whereas population-scale sets are useful for determining call set novelty or rarity. These data sets differ in sample size, ancestry, depth, platform, merging methodology, sensitivity and specificity, all of which should be considered before deciding which set is right to utilize, as biases influenced by these choices are inherently passed to the applications that employ them. Reference sets also vary widely when it comes to orthogonal validation, whereby some reference sets employ multiple orthogonal platforms but others perform none, opting to maximize quality metrics instead. Given this large variation, projects often use more than one reference set to maximize inclusivity and avoid overfitting. Reference sets undergo an iterative process where newer data sets are typically more sensitive and exhaustive due to technological improvements. Thus, developing algorithms should focus their benchmarks on more recent resources to avoid confounding issues stemming from technological limitations in legacy data. Indeed, a recent study found numerous batch effects within the 1000 Genomes Project release set[173]. Selected sequencing-driven reference data sets representing phenotypically 'normal' individuals are listed in TABLE 1. We chose data sets that include SV calls, focus on collections with available raw data and list orthogonal data from multiple sources for some reference sets. Additional resources can be found in dbVar[174].

Phased SVs
(Phased structural variations). Variants that are assigned to a paternal haplotype, often computed using family trio or heterozygous single-nucleotide variant data.

Receiver operating characteristic curves
Plots of the true positive rate against the false positive rate showing the relationship between sensitivity and specificity.

Connected-molecule strategies
Genomic methods that connect shorter reads of a DNA molecule together to provide long-range information.

Sequence coverage
The average number of times a given locus is covered by a sequence read.

Physical coverage
The average number of times a given locus is covered by the cumulative length of the reads, including unsequenced inserts.

Single-molecule strategies
Genomic methods that read the entirety of long strands of DNA.

Specificity
The ability to detect the absence of variants correctly. Low specificity implies many false positives.

likely alter gene function. Strikingly, they found that 57% of the human reference genome GRCh37 is covered by at least one CNV. The 1KGP and subsequent population-scale SV analyses show the potential for SVs to affect gene expression and reveal the prodigious ubiquity of SVs far beyond the ~12 CNVs per human genome estimated in 2004 (REF.[56]).

In contrast to global approaches, some projects focus on detecting SVs from populations that derive from a recent common ancestry. SVs were twice analysed in ~750 genomes derived from 250 Dutch families, once for de novo SVs and then for phased SVs (note that SVs were defined as variants >20 bp in this project), revealing Dutch-specific SVs and SV hotspots undetected by the 1KGP[45,57]. Similar work used an ensemble algorithm to detect SVs in 1,070 Japanese individuals to develop a Japanese-specific reference panel[58]. An increase in similar population-specific SV detection projects will be necessary to shift the diversity gap in genetics research and help identify rare SVs specific to ancestral backgrounds[59]. Indeed, some groups are still extremely under-represented; for example, Hispanic and Latin American individuals make up only 7.8%[43] and 16%[44] of recent data sets, respectively.

### *Limitations*

Studies that use ensemble algorithms are confounded by highly variable coverage, which has ranged from 3× to 90× in different projects, leading to the application of ad hoc heuristics and filtering which appreciably influence sensitivity and detection outcomes. Projects employ anywhere from three to 19 distinct algorithms — variations in sensitivity and precision between algorithm choices directly affect the consensus call set, as the accuracy of ensembles is highly influenced by algorithm combinations[38]. The truth sets used to benchmark calls and the filters applied for stringency are also highly variable, leading to parameterizations that sacrifice precision for recall, or vice versa. Additionally, stand-alone ensemble algorithm tools are largely immature and mostly rely on simple overlap. Larger projects optimize ensemble algorithms with truth sets generated from validation data, implementing FDR cut-offs and tuning receiver operating characteristic curves. However, stand-alone methods do not possess specifically generated benchmarks, making it difficult to implement these methods. The development of standardized variant benchmarks is an active area of research that may help formalize the development of ensemble algorithms by providing high-quality reference data sets that are thoroughly validated computationally and experimentally[42,60]. Furthermore, ensemble algorithms focused on integrating only short-read data do not overcome the limitations of short-insert sizes; they continue to poorly detect small insertions and suffer in repetitive regions[39,61,62].

## Emerging genomic technologies

A plethora of emerging technologies seek to expand beyond the capabilities of short reads. Connected-molecule strategies, such as linked reads, Strand-seq and high-throughput chromosome conformation capture (Hi-C), expand upon short reads by inferring long connections between distally mapped short-read pairs. These strategies are similar to long-insert short-read libraries (reviewed elsewhere[63]), which trade lowered sequence coverage for high physical coverage, improving and decreasing power to detect large and small variants, respectively. Alternatively, single-molecule strategies generate contiguous reads tens to hundreds of kilobases long, thus enabling direct detection of many SVs and improving alignment of unique reads in repetitive regions. Single-molecule strategies exist in two dominant forms: long-read sequencing by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT); and optical mapping by Bionano. Comparatively, connected-molecule strategies have high specificity for defined size ranges and SV subtypes, whereas single-molecule strategies have higher overall sensitivity. Many of the above technologies are thoroughly reviewed in REF.[21].

### *Connected-molecule strategies*

***Linked reads.*** Numerous methods, such as pooled-clone sequencing and Illumina Synthetic Long Reads, represent 'synthetic long reads' or linked reads, which use specific library preparations to infer long-range information from existing short-read sequences[64,65]. The 10x Genomics Linked-Reads (LR) platform is currently the most commonly used synthetic long-read platform. This technology partitions and barcodes diluted high-molecular-weight DNA using a microfluidic device prior to short-read sequencing; the origin of the short-read fragments can be determined from their respective barcodes, and long-range information is reconstructed in silico[66]. Additionally, linked reads retain their underlying short-read information and have greatly increased physical coverage resulting from coverage of the constructed molecule combined with coverage of each underlying short fragment. The physical coverage makes linked

Table 1 | **A list of currently available reference data sets**

| Selected reference data sets | Reference type, platform and coverage | Raw data publicly available | Sample number | SVs detected | Description; orthogonal validation if applicable |
|---|---|---|---|---|---|
| 1KGP phase 3[5] | Population-scale<br>Illumina short-read, 7.4 | Y | 2,504 | 68,818 | Individuals across 26 populations; PCR, orthogonal short-read platforms, PacBio and microarrays |
| 1KGP — high coverage | Population-scale<br>Illumina short-read, ~30 | N/A | 2,504 | N/A | High coverage sequencing of the individuals from phase 3 of the 1KGP |
| Genome of the Netherlands release 6.1[45] | Population-scale<br>Illumina short-read, 12 | N | 769 | 59,358* (>20 bp) | 769 individuals from 250 Dutch families; PCR amplification of breakpoint junctions followed by Sanger or short-read sequencing |
| Tohoku Medical Megabank Organization, 1KJPN[58] | Population-scale<br>Illumina short-read, 32.4 | N | 1,070 | 56,697* (>100 bp) | Individuals of Japanese ancestries; digital droplet PCR |
| GTEx[142] | Population-scale<br>Illumina short-read, 49.9 | N | 147 | 23,602 | SVs detected across 13 different human tissues; microarray data |
| Abel et al.[44] | Population-scale<br>Illumina short-read, ≥20 | N | 17,795 | 118,973 (GRCh37)<br>241,426 (GRCh38) | Individuals of African American, Latino, Finnish European, non-Finnish European, East Asian, Pacific Islander and South Asian ancestries |
| Sherman et al.[207] | Population-scale<br>Illumina short-read, 30–40 | Y | 910 | 125,715 | Novel insertion detection in individuals of African ancestries |
| gnomAD-SV[43] | Population-scale<br>Illumina short-read, 32 | N/A | 14,216 | 498,257 | Individuals of African, East Asian, European, Latino and admixed ancestries |
| Venter/HuRef[170,208,209] | Highly curated<br>Sanger reads, 7.5<br>10x Genomics LR, 42<br>Illumina short-read, 92, 36<br>Illumina 2 kb mate-pair, 7<br>Illumina 5 kb mate-pair, 6<br>Illumina 12 kb mate-pair, 3 | Y | 1 | 808,346* | De novo assembly of a European–American adult man; Sanger sequencing-based assembly, a wide suite of microarray data, and BAC and fosmid libraries |
| CHM1[61,97] | Highly curated<br>PacBio, ~40<br>PacBio, 62.4 | Y | 1 | 20,602 | De novo assembly of a haploid human hydatidiform mole; short-reads and Sanger capillary-based sequencing; target sequencing of BAC clones, de novo PacBio assemblies, Sanger sequencing and targeted PCR |
| CHM13[97,210] | Highly curated<br>PacBio, 66.3<br>ONT, 32<br>10x Genomics LR, 50<br>Bionano OM, 430<br>Hi-C, 40<br>Illumina short-read, ~30 | Y | 1 | 20,470 | Haploid human hydatidiform mole; target sequencing of BAC clones, de novo PacBio assemblies, Sanger sequencing and targeted PCR |
| HX1[103] | Highly curated<br>PacBio, 103<br>Bionano OM, 101<br>Illumina short-read, 143 | Y | 1 | 20,175 | De novo assembly of a Chinese adult man |
| AK1[104] | Highly curated<br>PacBio, 101<br>Bionano OM, 97 & 108<br>10x Genomics LR, 30<br>Illumina short-read, 72 | Y | 1 | 18,210 | De novo assembly of a Korean adult man; BAC clone assembly |
| Audano et al.[108] | Population-scale<br>PacBio, ~57 | Y | 15 | 99,604 | Individuals of African, Asian, European, American and South Asian ancestries; BAC and fosmid libraries |

Table 1 (cont.) | **A list of currently available reference data sets**

| Selected reference data sets | Reference type, platform and coverage | Raw data publicly available | Sample number | SVs detected | Description; orthogonal validation if applicable |
|---|---|---|---|---|---|
| Swe1 & Swe2[105] | Highly curated | N | 2 | 17,936 (Swe1) | One Swedish man and one Swedish woman |
| | PacBio, 78.8 (Swe1) | | | 17,687 (Swe2) | |
| | PacBio, 77.8 (Swe2) | | | | |
| | Bionano OM, >100 | | | | |
| Levy-Sakin et al.[119] | Population-scale | Y | 156 | 15,601 | 156 samples from the 1KGP; concordance with 10x Genomics LR |
| | Bionano OM, 79 | | | | |
| | 10x Genomics LR, 60 | | | | |
| Pendleton et al. & Jain et al., NA12878[102,172] | Highly curated | Y | 1 | 34,237 | Two separate de novo assemblies of a white, adult woman; PCR |
| | PacBio, 22 and 24 | | | | |
| | Bionano OM, 80 | | | | |
| | ONT, 26[a] | | | | |
| Genome in a Bottle, NA12878 | Highly curated | Y | 1 | 10,594 | One white, adult woman |
| | PacBio, ~44 | | | | |
| Wong et al.[75] | Population-scale | Y | 17 | 1,842 | De novo assembly and non-reference insertion detection in individuals of African, American, East Asian, European and South Asian ancestries; insertions >2kb were validated with OM |
| | 10x Genomics LR, 60 | | | | |
| Genome in a Bottle HG005 (son), HG006 (father), HG007 (mother)[211,212] | Highly curated | Y | 3 | 59,973 | A preliminary call set containing deletions and insertions from a Han Chinese family trio |
| | Illumina short-read, 300 (son), 100 (parent) | | | | |
| | Complete Genomics, 98 | | | | |
| | Ion Proton, 1,036 | | | | |
| | Bionano OM, 57 | | | | |
| | PacBio, 60 (son), 30 (parents) | | | | |
| Genome in a Bottle HG002 (son), HG003 (father), HG004 (mother)[60] | Highly curated | Y | 3 | 12,745 | Contains high-confidence deletions and insertions from an Ashkenazi family trio; concordance across multiple trios |
| | Illumina short-read, ~300, ~14.5, ~25, ~208.5, ~101, ~100 | | | | |
| | 10x Genomics LR, 47 (mother), 36 (father), 86 (son) | | | | |
| | Complete Genomics, ~101, 100 | | | | |
| | Ion Proton, 1,020 | | | | |
| | Bionano OM, 92 (mother), 87 (father), 112 (son) | | | | |
| | PacBio, ~31 (parent), 69 (son) | | | | |
| | ONT, 0.017 (son) | | | | |
| Human Genome Structural Variation Consortium[39] | Highly curated | Y | 3 (data available for 9) | 103,985 | Three family trios of Han Chinese, Puerto Rican and Yoruban Nigerian ancestries; concordance across multiple genomic platforms |
| | PacBio. ~40 | | | | |
| | ONT, 18.9 | | | | |
| | Illumina short-read, 74.5 | | | | |
| | Illumina 3 kb mate-pair, 3 | | | | |
| | Illumina 7 kb mate-pair, 1.1 | | | | |
| | 10x Genomics LR, 82.4 | | | | |
| | Bionano, N/A | | | | |
| | Tru-Seq SLR, 3.47 | | | | |
| | Strand-seq, N/A | | | | |
| | Hi-C, 19.49 | | | | |

A version of this table with additional information can be found online as Supplementary Table 1. 1KGP, 1000 Genomes Project; BAC, bacterial artificial chromosome; GTEx, genotype–tissue expression; Hi-C, high-throughput chromosome conformation capture; LR, linked reads; N/A, not available; OM, optical mapping; ONT, Oxford Nanopore technologies; PacBio, Pacific Biosciences; SV, structural variation. [a]Median coverage depth.

reads well suited for SV detection, whereas the low error rate and long molecule length (up to 100 kb) make the method useful for haplotype phasing[67]. Detection methods such as Long Ranger[66,68] and GROC-SVs[69] leverage read clouds, which are clusters of short reads thought to derive from the same underlying molecule due to identical barcodes. Read-cloud methods look at two criteria: the density of overlapping barcodes, where sudden increases or drops in barcode 'coverage' determine SV breakpoints; and distant genomic loci that share more barcode overlap than would occur by chance (FIG. 2). GROC-SVs additionally performs local reassembly to detect complex SVs 10–100 kb in length. A second approach analyses split alignments within long-read 'molecules' reconstructed from shared barcodes, analogous to split reads. LinkedSV[70], NAIBR[71] and VALOR2 (REFS[72,73]) are SV callers that use split-molecule approaches to detect SVs, whereas ZoomX[74] considers discrepancies in molecule coverage.

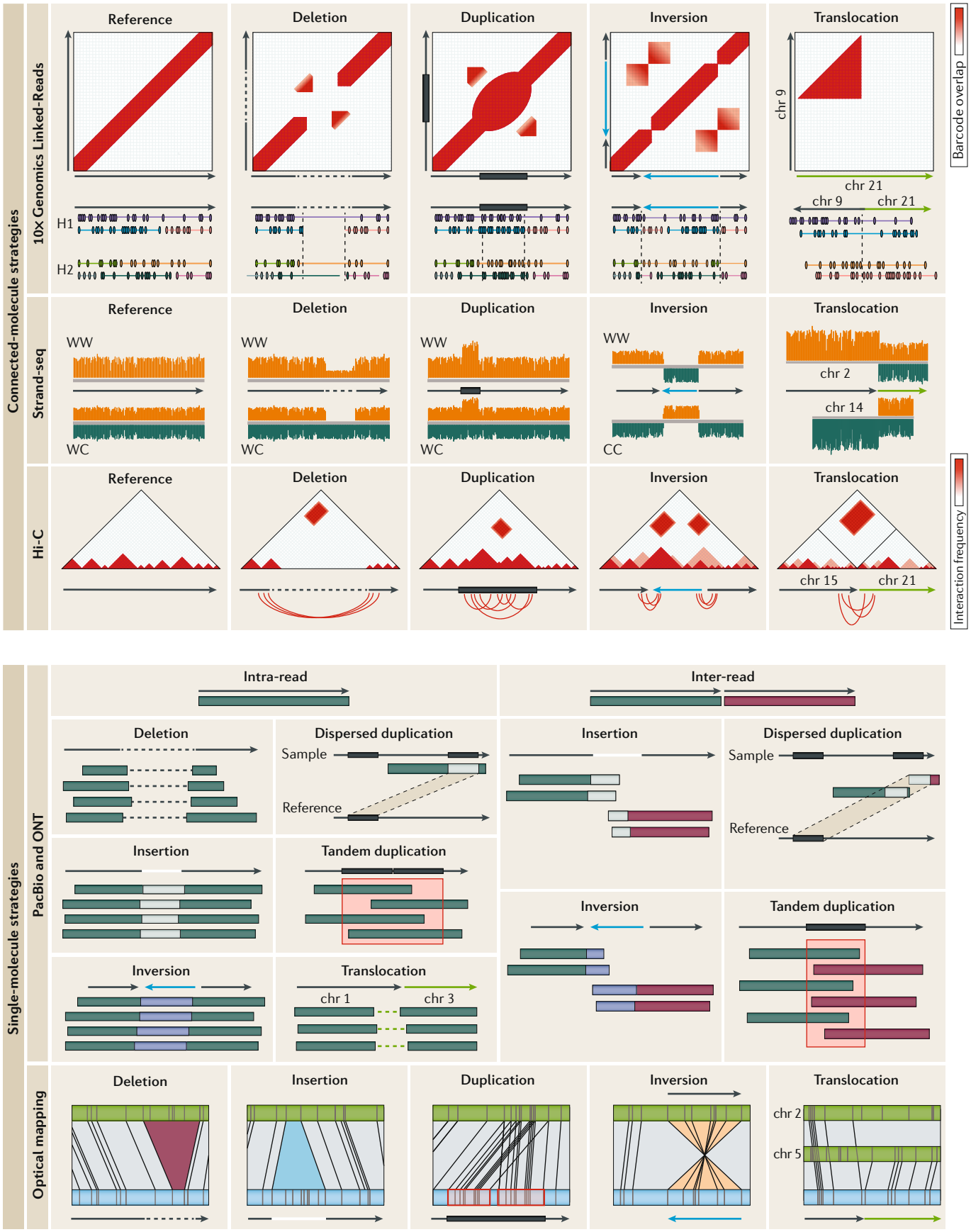
Linked-read approaches have various strengths owing to their barcoding, a key feature being the ability to determine whether fragments mapping to distant genomic loci derive from the same molecule, which makes the visual interpretation of translocations and large SVs exceptionally effective[66]. Linked reads are able to detect similar amounts of deletions compared with single-molecule approaches but there is a discrepancy in detectable insertions[68]. Whereas assembly-based linked-read studies have found megabases of novel insertional sequence across different populations[75,76], single-molecule approaches will typically detect more insertion events[77]. This may result from the fact that linked reads have a coverage drawback compared with single-molecule reads: no molecule within a read cloud has complete coverage of the DNA fragment. Hence, there are substantial gaps between the read pairs underlying each molecule, decreasing mappability in repetitive regions. The decrease in insertion detection may also result from the higher algorithmic difficulty of calling insertions through mapping versus assembly approaches, which use simple pairwise comparisons[78]. Indeed, one of the most widely used algorithms, Long Ranger, cannot currently call insertions. However, recent efforts to develop algorithms that augment the mapping of linked reads to repetitive regions are improving the ability of linked reads to detect novel sequence insertions[77,79].

***Strand-seq.*** Strand-seq independently sequences template DNA strands by incorporating bromodeoxyuridine into the non-template strand during replication, followed by UV-induced photolysis at bromodeoxyuridine sites to selectively ablate the nascent strand[80]. As libraries only contain independent parental strands, Strand-seq is especially suited for haplotype phasing. The inherent directionality enables highly efficient detection of inversions, which manifest as segments of opposing strand orientation[39,81] (FIG. 2). Indeed, Strand-seq has been used to identify polymorphic inversions, showing that they are enriched in certain chromosomes over others, and revealing that the reference genome carries the minor allele or is misoriented at many inverted loci[81].

Fig. 2 | **Structural variation signatures in single-molecule and connected-molecule strategies.** Emerging technologies vary in how they detect structural variations (SVs). 10x Genomics Linked-Reads detect SVs based on barcode overlap between genomic loci. Split-molecule approaches infer SVs from splitting of linked reads, examples of which are displayed below each barcode matrix (each colour represents a shared barcode and linked molecules are separated by haplotype; only homozygous variants are shown for simplicity). Strand-seq determines SVs based on read depth or sudden changes in mapping orientation. For deletions and duplications, only two of four possible daughter cell configurations are shown for simplicity (Watson–Watson (WW) and Watson–Crick (WC); Crick–Crick not shown). For inversions, only a homozygous inversion in Watson–Watson and Crick–Crick daughter cells are shown as Watson–Crick daughter cells mask homozygous inversions (homozygous for simplicity; for more detail on inversion detection, see REF.[81]). High-throughput chromosome conformation capture (Hi-C) detects SVs by looking for unusually high-frequency contacts between genomic loci. Underneath each interaction matrix is a schematic of the expected chromosomal contacts resulting from each SV. Single-molecule sequencing methods infer SVs based on discordant mapping signatures that can involve one (intra-read) or many (inter-read) reads. SVs derive from intra-read signatures, which result from reads that span an entire SV, or inter-read signatures, which require multiple reads to cover the event. Insertions differ from deletions by an increase in the expected distance between the two split pairs marked by the white soft-clip between the reads, and inversions involve reads that map best to the complementary strand. Optical maps detect SVs based on increased presence, absence or change in the orientation of restriction enzyme sites compared with a reference (blue, sample; green, reference). Resolution is dependent on the distribution of restriction enzyme sites. chr, chromosome; H1, haplotype 1; H2, haplotype 2; ONT, Oxford Nanopore Technologies; PacBio, Pacific Biosciences.

Large deletions and duplications can be detected by read-depth approaches, whereas translocations are detected as changes in template state, as implemented in BAIT[82]. However, Strand-seq requires many enzymatic clean-up steps that ultimately reduce sequence coverage to an average of 0.01–0.05× per library, which makes it inappropriate to detect smaller-sized SVs until improvements in single-cell library preparation are made[83]. Additionally, as inversions and translocations in Strand-seq look similar to sister chromatid exchanges, events must be consistent across multiple libraries for identification; thus, SV detection by Strand-seq requires preparation of many individual single cells.

***Hi-C.*** Hi-C involves sequencing crosslinked chromatin to provide information about DNA sequences that may be distant in the linear genome but proximal in 3D space[84]. Hi-C read pairs can span megabases, making the method useful for detecting large SVs, especially translocations. However, as Hi-C relies on the presence of digestion sites kilobases apart in the linear genome, its resolution is limited. Hi-C also relies on underlying read pairs and suffers from low sequence coverage,


Connected-molecule strategies
10X Genomics Linked-Reads
Reference
Deletion
Duplication
Inversion
Translocation
chr 9
chr 21
H1
H2
Barcode overlap
Strand-seq
WW
WC
CC
chr 2
chr 14
Hi-C
chr 15
Interaction frequency
Single-molecule strategies
PacBio and ONT
Intra-read
Inter-read
Sample
Reference
Insertion
Tandem duplication
Optical mapping
Partial read alignment
Split read that maps best to complementary strand
Missing sequence

Base-calling error
Errors in determining the respective nucleotide from raw signals during sequencing.

Circular consensus sequencing
A single-molecule real-time (SMRT) sequencing method that improves accuracy through multiple passes of the template molecule.

as do linked reads and Strand-seq. Chromosomal interactions derived from Hi-C are represented in a contact-frequency heat map across all possible pairs of genomic loci. Interactions between proximal loci are shown in the diagonal, and contacts off of the diagonal are indicative of long-range interactions. Unusually elevated contact frequencies between distal loci represent possible deletions, inversions and translocations, whereas elevated contact frequencies at proximal loci are indicative of duplications[85] (FIG. 2). Although Hi-C has been used to detect predominantly translocations within cancer cells, methods to detect other SVs, such as HiCNV, which uses read coverage to detect CNVs, are starting to emerge[85–89]. Delineating potential SVs from regular fluctuations in 3D structure remains a notable challenge. Recent work shows that large CNVs can affect chromatin organization across the chromosome, further confounding the ability to differentiate between variation in chromatin interaction and putative rearrangements[90]. To address this problem, Hi-C Breakfinder uses a probabilistic model that incorporates information about expected spatial features when determining aberrant contact frequencies[91]. However, most of the intrachromosomal SVs detected by this method are >2 Mb in size, as distinction from local interactions is still difficult. Additionally, Hi-C currently requires cell culture of millions of cells, although recent developments aim to decrease this limitation[92]. A deeper understanding of 3D architecture will be necessary before Hi-C can reliably call SVs independent of orthogonal support.

### *Single-molecule strategies*

***Single-molecule real-time sequencing.*** PacBio single-molecule real-time (SMRT) sequencing leverages a stationary polymerase attached to the bottom of a nano-sized well and passages single DNA strands through the enzyme to produce long reads that significantly improve unambiguous mappability across the genome[93]. Algorithms detect SVs from SMRT data by leveraging intra-read and inter-read signatures (FIG. 2). Intra-read signatures enable the direct detection of SVs and are derived from reads spanning entire SV events, resulting in a missing sequence (deletion) or a soft-clip (insertion) within properly aligned flanking sequences. Inter-read signatures involve multiple reads and detect SVs from inconsistencies in orientation, location and size during mapping, analogous to short read signatures. After signature detection, callers typically cluster and merge similar signatures from multiple reads, delineate proximal but different signatures and choose the highest quality reads that support the putative SV. CORGi[94], PBHoney[95], pbsv, Sniffles[96], SMRT-SV[61,97] and SVIM[98] detect SVs through combinations of intra-read and inter-read signatures but differ in their discovery heuristics. Sniffles filters SVs by evaluating similarities between breakpoint position and size, and additionally clusters SVs supported by the same set of reads to detect nested SVs[96]. SVIM evaluates how signature clusters overlap each other or nearby breakpoints to differentiate between interspersed duplications, tandem duplications and novel sequence insertions[98]. Some methods, such as CORGi and SMRT-SV, locally reassemble loci with SV signatures and call SVs based on consensus sequences derived from these assemblies[61,94,97]. NextSV integrates Sniffles and PBHoney analogous to the ensemble algorithm approaches discussed above[99].

Single-molecule sequencing studies have so far been used to investigate fewer genomes than short-read studies due to higher operational costs, a large input DNA requirement and lower sample throughput. Thus, although many short-read studies sequence across numerous genomes, long reads have been mostly applied to single-genome assemblies. Although the base-calling error rate for PacBio sequencing is higher than for short reads, this can be overcome by increasing coverage or utilizing circular consensus sequencing[100]. It is pertinent to note that higher SMRT coverage results in more accurate consensus sequences but at a trade-off for shorter median read lengths due to enzyme degradation — researchers must determine the ideal coverage according to project aims[101]. Nonetheless, these single-molecule applications are challenging the SV detection landscape and its reliance on short-read technology. Sequencing of the CHM1 human hydatidiform mole genome served as a proof of concept for using long reads to resolve SVs, detecting >20,000 SVs in this haploid genome compared with ~2,500 SVs per diploid genome in the 1KGP[5,61] (BOX 1). A recent analysis found that PacBio long reads were approximately three times more sensitive than a short-read ensemble maximized for sensitivity, implying that a large subset of SVs, many 50–2,000 bp in length, are unresolvable without long reads[39]. Approximately half of the novel variants detectable by long reads are insertions ~500 bp in length embedded within mobile elements and tandem repeats (BOX 3). SMRT assembly or SV detection in 19 other human genomes found comparably large magnitudes of SVs and exhibited the corresponding insertional bias[39,60,96,97,102–107]. As it is impossible to tell the difference between a novel insertion or missing sequence in the reference, the magnitude of SVs that have been detected questions the completeness of the human reference genome. To investigate this issue, one study performed SV discovery in 15 individuals sequenced using long-read technology to an average ~57× and found 86,761 SVs absent from the 1KGP and the Genomes of the Netherlands project data sets[108]. A substantial amount of the SVs shared between these 15 genomes are not present in the GRCh38 version of the human reference sequence, which implies it may contain errors or minor alleles at many SV loci. Remarkably, ~50% of the detected SVs intersect genes or regulatory elements[108]. Overall, long-read technology enables detection of previously unresolvable SVs and may be pivotal in deciding how the field of genomics evolves from using a single human reference genome.

***Nanopore sequencing.*** Algorithms to detect SVs from nanopore sequencing are still emerging but have gradually become available, primarily through studies utilizing ONT. During nanopore sequencing, a single-stranded DNA is threaded through a protein pore, and DNA sequences are discriminated based on the changes in electric current elicited by different bases[109,110]. As nanopore sequencing is a variation of single-molecule

Box 3 | **Confounding complexity**

The detection studies discussed have revealed that structural variations (SVs) consisting of complex arrangements are more prevalent than previously perceived in both phenotypically 'normal' individuals and individuals with disease[5,43,45,61,102,119,127,128,132,146,148,152,153]. Additionally, new technologies have identified substantial amounts of SVs in areas that are difficult to resolve with short reads. These loci are either extremely low in complexity, such as tandem repeats, telomeres and mobile element insertions, or high in complexity, such as segmental duplications, centromeres, the major histocompatability complex and other areas of high olymorphism[5,39,61,97,103–105,108,114,117–119,125,175]. Indeed, mechanisms behind SV formation, such as non-allelic homologous recombination and replication-based mechanisms, are dependent on local repeat structures, which leads to breakpoints within repetitive regions (reviewed in REF.[176]). 'Complexity' confounds detection in two ways: in terms of complex SV events and in terms of the variable complexity at genomic loci. It is essential to consider specialized methods that can leverage new technologies to detect SVs in complex regions and detect SVs of complex arrangements, and methods that reassemble complex regions to decrease unambiguous mapping. Indeed, specific tools — such as SDA[177], which resolves segmental duplications, CORGI[94], which resolves complex events, and rMETL[178], which detects mobile element insertions, and other tools taking a specificity-first approach — will help in resolving difficult-to-detect SVs that cannot be ascertained from generalized whole-genome approaches due to complicated genomic loci or irregular compounded structure (see the table). Eventually, generalized SV detection methods should implement the strategies used from specialized callers or be utilized concurrently for a more comprehensive assessment of genome-wide SV.

| Method | Detection | URL |
|---|---|---|
| Sniffles[96] | Complex SVs | https://github.com/fritzsedlazeck/Sniffles |
| CORGi[94] | Complex SVs | https://github.com/zstephens/CORGi |
| HySA[132] | Complex SVs | https://bitbucket.org/xianfan/hybridassemblysv |
| GROC-SVs[69] | Complex SVs | https://github.com/grocsvs/grocsvs |
| TSD[179] | Complex SVs | https://github.com/menggf/tsd |
| local-rearrangements[180] | Complex SVs | https://github.com/mcfrith/local-rearrangements |
| gemtools[181] | Complex SVs, SV phasing | https://github.com/sgreer77/gemtools |
| SDA[177] | Segmental duplications | https://github.com/mvollger/SDA |
| rMETL[178] | Mobile element insertions | https://github.com/hitbc/rMETL |
| adVNTR[182] | Variable number tandem repeats | https://github.com/mehrdadbakhtiari/adVNTR |
| PacmonsTR[183] | Tandem repeats | https://github.com/alibashir/pacmonstr |
| RepeatHMM[184] | Microsatellites | https://github.com/WGLab/RepeatHMM |
| nplvn[185] | NAHR-mediated inversions | https://github.com/haojingshao/nplnv |
| VALOR2 (REF.[73]) | Segmental duplications | https://github.com/BilkentCompGen/valor |
| PALMER | Mobile element insertions | https://github.com/mills-lab/PALMER |
| tandem-genotypes[186] | Tandem repeats | https://github.com/mcfrith/tandem-genotypes |

NAHR, non-allelic homologous recombination; SV, structural variation.

sequencing, the signatures to detect SVs are similar to those used in PacBio data (FIG. 2). Callers that detect SVs from nanopore data include NanoSV[111], Picky[112], Sniffles[96] and SVIM[98]; the latter three also detect SVs from PacBio data. Both NanoSV and Picky leverage split reads to detect SVs and apply heuristics that consider coordinates, orientation and breakpoint sites. NanoSV iteratively clusters all reads that support a breakpoint junction, whereas Picky stitches together split reads with surrounding reads and calls SVs from the best alignments. Studies that use ONT find similar numbers of SVs as PacBio detection but show many nanopore-specific small deletions[111,112]. However, one study found the overwhelming majority of ~10,000 unique ONT SVs to be small deletions located within repeat regions and likely derived from base-calling errors, compared with ~800 unique PacBio SVs, of which ~40% overlapped repeats[96]. Another study found that ONT SV algorithms detect small SVs poorly[113]. ONT provides improved read lengths, lower adaptation costs and higher throughput than PacBio, while still being effective at detecting many SVs. However, reduced specificity owing to high error rates make ONT less suitable for smaller SVs (<100–200 bp), although recent improvements aiming to reduce base-calling errors may mitigate this issue. Overall, the single-molecule approaches provided by PacBio and ONT enable highly sensitive SV detection and are the most powerful methods to detect novel sequence insertions.

***Optical mapping.*** Optical mapping, an alternative to sequencing-based technologies, linearizes single DNA strands in nanochannels and intermittently marks them with a nicking endonuclease to create physical maps known as genome maps[114–116]. Optical mapping-based methods call SVs by comparing divergences in the nicks of DNA strands against an in silico digested reference: missing or extra labels and the spacing between labels are used to determine deletions or insertions; repeated labels indicate repeats and copy number changes; the presence of unique nicks on non-reference loci indicate translocations; and reversed nicking patterns indicate

inversions (FIG. 2). The generated DNA fragments are up to 1 Mb long, making optical mapping well suited to detect large genomic rearrangements, particularly insertions, and effective at identifying SVs within repetitive regions[75,117–119]. Optical mapping excels at deconvoluting zygosity as long as there is sufficient coverage such that molecules spanning each haplotype can be directly observed[118]. Due to reliance on restriction enzyme sites, optical mapping does not produce a sequence and therefore lacks base-pair resolution, instead providing breakpoint estimations based on the most proximal nicks. As a result, optical mapping detects substantially fewer SVs than long-read methods and is typically limited to sizes of ~6 kb and larger, although newer applications improve resolution by utilizing more than one restriction enzyme[21,75,107,118–120]. Thus, most optical mapping applications detect large SVs through de novo assembly of genome maps but use short-read sequencing to detect smaller variants[103,104]. New detection algorithms such as OMSV[120] and Bionano Solve[121] call SVs without de novo assembly by using alignment-based strategies. It is important to note that optical mapping suffers from a high error rate, whereby errors manifest as missing or extra labels from incomplete and uneven stretching of individual molecules in their nanochannels[117,120]. Resolution and error rate notwithstanding, optical mapping is amplification-free and significantly cheaper than HTS even at 60× coverage, making it an economical choice to investigate large cohorts[118]. Recent work used optical mapping on 154 genomes from the 1KGP to find ~60 Mb of sequence not present in the reference genome as well as 55 loci in the genome that are both structurally complex and harbour complex SVs[119].

## Multiplatform discovery

Currently, no single method or technology has been shown to be comprehensive enough to detect all SV within a genome. Multiplatform approaches that combine strengths of various genomic platforms to enhance detection of SVs across all types and sizes have emerged as a result. The platforms discussed can be employed in combination to complement strengths and mitigate weaknesses[60]. Due to their high base-calling accuracy, bioinformatic maturity and affordability, short reads are regularly used to correct errors in long reads, a process known as 'polishing' (reviewed and evaluated elsewhere[78,122–124]), whereas newer technologies are used for exhaustive variant detection and resolution of complex structures. A practical example includes combining short-read sequencing at higher coverage (>30×) with lower-coverage single-molecule sequencing (~10×) to optimize economy and sensitivity. The use of individual technologies depends on logistical variables such as cost, required resolution and project scope. Technical variables including sensitivity, variant size, repetitive nature of the target region and haplotype information must be considered as well. An overview of each technology is provided in TABLE 2, with additional information on advantages and disadvantages provided in Supplementary Table 2.

Multiplatform discovery is often used to investigate SVs in cancer (BOX 4). Two studies on leukaemia and prostate cancer genomes integrated short reads with optical mapping and found that many SVs detected uniquely by optical mapping have breakpoints within regions of low mappability, whereas SVs detected uniquely by short reads are typically smaller and below the resolution of optical mapping[125,126]. An analysis combining an ensemble algorithm, linked reads and long-insert libraries to detect and phase SVs in the K562 and HepG2 cell cancer genomes identified thousands of calls unique to each platform[127,128]. Similarly, combining optical mapping, short reads and Hi-C to detect SVs in eight different cancer genomes reported that only 20% of interchromosomal translocations were detected by two or more platforms, demonstrating the necessity of multiplatform discovery to detect all variants[91]. In another study, short reads were used not to improve sensitivity across the detection size spectrum but to resolve ambiguity in unique, unaligned optical mapping fragments from a liposarcoma genome[129]. Whereas optical mapping was necessary to reveal large fragments, the short read signatures provided the necessary resolution to reveal ~6 SV breakpoints within the unaligned maps, suggesting that the fragments consisted of complex SVs.

Genome assemblies typically integrate platforms when detecting SVs to increase sensitivity and produce orthogonal validation, known as a hybrid assembly. In one example, assembly of the HG001 genome (also known as NA12878) merged PacBio contigs with optical genome maps to create highly contiguous scaffolds with an N50 of 28.8 Mb (REF.[102]). As 55% of inversions called from these scaffolds were enriched for arrangement complexity and colocalization with other SVs, they would be difficult to detect without the improved contiguity from integration. A similar approach was used in another study[105]. Also, a team generated short-read and long-read sequences for the HS1011 genome and detected SVs by combining an ensemble algorithm, PacBio and hybrid local reassembly[130]. Although the authors found many SVs overlapping from the three approaches, they revealed bona fide SVs that were unique to their respective detection method. Additionally, hybrid reassembly detection performed with an FDR <10%, whereas popular short-read callers (BreakDancer, CNVnator, DELLY and Pindel) exhibited FDRs between 31% and 80%, showing greatly improved detection with integration. A recent comprehensive multiplatform discovery of SVs integrated nine platforms across three family trios, discovering ~27,622 SVs per genome[39]. This study combined an ensemble algorithm, PacBio, optical mapping, Strand-seq and long-insert libraries to detect deletions, insertions and inversions, with additional technologies applied for phasing, assembly and orthogonal validation (TABLE 1). PacBio contributed the highest number of unique deletions and insertions, and Strand-seq contributed the highest number of inversions; each platform identified high-confidence unique calls. Each of these studies illustrates that combining platforms is necessary for comprehensive detection across the full range of SVs.

Integration of SV calls from differing technologies is analogous to ensemble algorithm approaches. Most methods are in-house and consider coordinate overlap, breakpoint proximity, mapping orientation, read

Hybrid assembly
A genome assembly that leverages sequencing data from multiple platforms to reconstruct the original sequence, using the orthogonal data to extend the contig lengths or to branch contigs to one another.

N50
A number that denotes the minimum contig size for which 50% of the nucleotide sequence is contained within. A larger N50 implies a more contiguous assembly.

Table 2 | **Algorithms to detect genome-wide SVs from ensemble, single-molecule and connected-molecule approaches**

| Platform | Method | Approach | SVs detected |
|---|---|---|---|
| Ensemble algorithms | SVMerge[40] | PE, SR and RD signals with integration of two specialized insertion callers; calls are merged on overlap with coordinate thresholds and validated by local reassembly | DEL, INS, INV, CNG, CPX |
| | HugeSeq[41] | PE, SR and RD signals, along with breakpoint junction mapping; calls are merged by 50% reciprocal coordinate overlap | DEL, DUP, INS, INV |
| | iSVP[49] | PE, SR and RD signals; additional calls are made with GATK HaplotypeCaller, which uses local reassembly; calls are merged by overlap | DEL |
| | MetaSV[51] | PE, SR and RD signals, along with breakpoint junction mapping; calls are merged by overlap that prioritizes read signatures by their respective resolution and are refined with local reassembly | DEL, DUP, INS, INV, TRX |
| | SpeedSeq[47] | PE and SR signals, along with a Bayesian likelihood genotyper; uses an RD caller to annotate the copy number at each variant locus | DEL, DUP, INS, INV, TRX, CNG |
| | Parliament2 (REF.[50]) | User choice of six individual callers; calls are merged based on coordinate overlap and scored with a precision metric trained on HG002 | DEL, DUP, INS, INV, TRX |
| | FusorSV[53] | Fits a model that determines which combination of eight individual callers performs best according to a user-input truth set | Dependent on input truth set |
| 10x Genomics Linked-Reads | Long Ranger[66] | Read-pair barcode overlap between distant loci and changes in barcode density | DEL, DUP, INV, TRX |
| | GROC-SVs[69] | Read-pair barcode overlap between distant loci and changes in barcode density; SVs are reconstructed with local reassembly | Reports reconstructed breakpoints that can derive from any SV type |
| | LinkedSV[70] | Molecule barcode overlap between distant loci and barcodes from two distance loci mapped to adjacent positions | DEL, DUP, INV, TRX |
| | VALOR2 (REFS[72,73]) | SR signatures from linked molecules, read-pair signatures and molecule depth for filtering | DEL, DUP, INV, TRX, INV–DUP, INV–TRX |
| | Novel-X[77] | Assembly of unmapped reads with reads of associated barcodes to obtain anchors in unique sequence; these reassembled contigs are then mapped to the reference | INS |
| | NAIBR[71] | Combines SR signatures from linked molecules with the PE signatures from the underlying short reads into a probabilistic model | DEL, DUP, INS, INV, TRX |
| | ZoomX[74] | Changes in linked molecule coverage | DEL, DUP, INV, TRX |
| Strand-Seq | BAIT[82] | Changes in the ratio of reads mapped in opposing directionality and sudden changes in template state that are consistent across loci | DEL, DUP, INV, TRX |
| | Invert.R[81] | Changes in the ratio of reads mapped to opposing directionalities | INV |
| Hi-C | HiCNV + HiCtrans[89] | RD of restriction enzyme fragments and high-frequency interchromosomal contacts | DEL, DUP, TRX |
| | Hi-C Breakfinder[91] | Clusters of interaction frequencies that deviate from expected | DEL, DUP, INV, TRX |
| PacBio | PBHoney[95] | Unmapped split-read tails (PBHoney-Tails) and intra-read discordance (PBHoney-Spots) | DEL, INS, INV, TRX |
| | pbsv | SR and intra-read signatures | DEL, DUP, INS, INV, TRX |
| | SMRT-SV[61,97] | Local assembly at loci with intra-read or inter-read signatures; SVs subsequently called from consensus sequences derived from each assembly | DEL, DUP, INS, INV |
| | Sniffles[96,a] | SR and intra-read signatures | DEL, DUP, INS, INV, CPX, TRX |
| | NextSV[99] | Combines calls from PBHoney and Sniffles by union (sensitive call set) or intersect (stringent call set) | DEL, DUP, INS, INV, CPX, TRX |
| | CORGi[94] | Chooses the highest scoring putative SV from a collection of possible SVs generated by realigning loci with split-read and intra-read signatures multiple times | DEL, DUP (tandem, dispersed), INS, INV, CPX, CNG |
| | SVIM[98,a] | SR and intra-read signatures | DEL, DUP (tandem, dispersed), INS, INV |
| Oxford Nanopore | NanoSV[111] | SR signatures and evidence from reads that map to putative breakpoint junctions | DEL, DUP, INS, INV, TRX |
| | Picky[112,b] | SR signatures from long-read alignments that are linked together to improve coverage | DEL, DUP, INS, INV, TRX |
| Optical mapping | OMSV[120] | Discordance in the number of and distances between restriction label sites | DEL, DUP, INS, INV, TRX |
| | Bionano Solve | Discordance in the number of and distances between restriction label sites | DEL, DUP, INS, INV, TRX |
| Multiplatform | MultiBreak-SV[131] | Clusters all possible short-read and long-read alignments that support a putative SV into a combined probabilistic model | DEL, INV, TRX |
| | HySA[132] | Clusters short reads with PE and SR signals with long reads; SVs are called from contigs assembled by the reads in each cluster | DEL, INS, CPX |

A version of this table with additional information is available as Supplementary Table 2. CNG, copy number gain; CPX, complex rearrangement; DEL, deletions; DUP, duplications; Hi-C, high-throughput chromosome conformation capture; INS, insertion; INV, inversion; PacBio, Pacific Biosciences; PE, paired end; RD, read depth; SR, split read; SV, structural variation; TRX, translocation. [a]Also able to detect SVs from Oxford Nanopore data. [b]Also able to detect SVs from PacBio data.

Box 4 | **Detecting structural variation in disease**

Structural variations (SVs) are associated with diverse diseases and are a notable hallmark of cancer genomes[187]. Long reads, linked reads, high-throughput chromosome conformation capture (Hi-C) and optical mapping resolve structures that short reads struggle to detect in the majority of cancers such as interchromosomal and intrachromosomal translocations, complex rearrangements, chromoplexy, chromothripsis, chained fusions and extremely large (>30 kb) SVs[69,74,87,127–129,145,188–192]. Pacific Biosciences (PacBio) reads were used to analyse the breast cancer cell line SKBR3, detecting >17,000 SVs, including SVs that overlap COSMIC, which are somatically acquired cancer-specific mutations[148]. The single-molecule approach detected 76% more SVs than an ensemble of three short-read callers (with two-caller concordance), most of which derive from repetitive regions. The long reads enabled identification of clustered, complex translocations and inverted duplications that amplified the oncogene *ERBB2* to >32 copies[148], as later confirmed in a separate long-read analysis[96], providing insight into a possible breast cancer-specific mechanism. Linked reads have been used to detect and phase translocations and gene fusions in cancer genomes, finding loci where heterozygous SVs have an impact on allele-specific expression[127,128]. Another linked-reads study resolved an extremely complex haplotype-specific SV in a lung cancer cell line where one haplotype harbours an *EML4–ALK* gene fusion and the other an *ALK–PTPN3* fusion[66]. Another study also used linked reads to study the genomic architecture of the *AR* oncogene in castration-resistant prostate cancer and found that SVs were likely to inactivate tumour suppressor genes in complex patterns where each haplotype could harbour a different type of inactivating SV[153]. Each of these findings are examples of complex genomic architectures now resolvable through the improved resolution of emerging technologies.

Copy number variants (CNVs) and de novo mutations play pertinent roles in the aetiology of several neuropsychiatric diseases such as intellectual disability, schizophrenia and, particularly, autism spectrum disorder (ASD)[193–195]. Application of ensemble algorithms in ASD family genomes has revealed CNVs that disrupt known neurodevelopmental genes; clustering of de novo SNVs proximal to de novo CNV regions[196]; an abundance of complex duplication-associated SVs[197]; and elevated numbers of de novo CNVs compared with unaffected individuals[198]. However, it is pertinent to note the challenges and disagreement in extrapolating the association between rare non-coding variants and ASD risk; a dearth of both rigorous analytical approaches and replicated associations between studies notably hampers the interpretation of non-coding SVs in these diseases[46,199,200].

Emerging methods have additionally been applied to Mendelian disorders, clinical phenotypes and structural haplotypes to identify SVs that are traditionally difficult to characterize. For example, optical mapping was effective at detecting the D4Z4 repeats in facioscapulohumeral muscular dystrophy, which are challenging to resolve with classic techniques due to their size[150,188]. In individuals in whom short reads were uninformative, PacBio sequencing was able to detect disease-causal SVs, such as a de novo ~2.1-kb SV overlapping *PRKAR1A* in Carney complex[143] and a 4.6-kb repeat expansion[201] and 12.4-kb deletion[202] in benign adult familial myoclonic epilepsy located in GA-rich and GC-rich regions, respectively. Similarly, in a patient with glycogen storage disease type 1a, for whom whole-exome and Sanger sequencing failed to determine a genetic cause, nanopore sequencing detected a compound heterozygous structure containing a point mutation and a 7.1-kb deletion in *G6PC* on separate alleles[144]. New detection methods have also identified complex SVs that are insufficiently resolved with short reads in patients with congenital abnormalities and severe quality-of-life disorders; they contain numerous breakpoints, cluster closely with other SVs, affect considerable nucleotides and are flanked by repetitive sequences[111,148,203–206]. In a final example, optical mapping was used to construct and determine the frequency of segmental duplication haplotypes LCR22A and LCR22D, which are involved in 22q11 deletion syndrome and escape short-read resolution. The large fragment sizes of optical mapping enabled the authors to find extensive CNVs, differing by up to 1.75 Mb between individuals, and revealed that the reference genome does not represent the major allele at this locus[175].

support, putative SV type and resolution of the underlying technology. There are few stand-alone multiplatform detection tools; most combine short and long reads, for example, MultiBreak-SV[131] and HySA[132]. MultiBreak-SV considers all possible short-read and long-read alignments that support a putative SV in a combined probabilistic model, whereas HySA clusters short reads with paired-end and split-read signals with the long reads that support them, before calling SVs from contigs assembled with the long reads in each cluster. New platform ensemble tools are expected to develop as the cost of sequencing continues to drop and access to new technologies improves.

## Integrating SVs with biological information

Despite the computational and technological improvements described, we are still unable to interpret the functional consequences of the vast majority of variants. Strategies to ascertain functional impact are necessary now more than ever given the expansive increase in detectable and novel SVs. Moving forwards, integrating SV detection across layers of biological information shows promise for elucidating the biological impact of variants.

Studies using short reads have shown the potential of integrative frameworks in interpreting SV function, and now a subset of studies employing the emerging detection methods discussed are starting to integrate SVs with layered biological data, such as gene expression, epigenetics and 3D genome structure, to understand the effects of SVs holistically[133–140]. Building on seminal work showing that CNVs affect expression phenotypes distinctly from SNVs[141], a recent study detected SVs with an ensemble algorithm before mapping SV expression quantitative trait loci. This study found that SVs had a larger median effect and were up to 53 times more likely to affect gene expression compared with SNVs or indels[142]. Indeed, other studies integrating emerging detection methods with expression data, long-read transcriptome sequencing and assembly have revealed the high potential for rearrangements to affect genes, demonstrating differential expression, alternatively spliced transcripts and complex gene fusions resulting from novel SVs[75,104,126,143–149]. Although the transcriptome is often integrated with SV calls, given its immediacy to the genome, more efforts to integrate the methylome are emerging and so far have revealed inconsistent methylation patterns around SVs[127,128,150,151], suggesting complex regulatory consequences. Another data type that should be considered with SVs are small variants and their effects. For example, nanopore sequencing analysis identified a heterozygous point mutation and an exon-disrupting deletion in an individual, when the typical disease genotype involved biallelic point mutations[144]. Additionally, a study investigating non-recurrent SVs with arrays, short reads and long reads found enrichment of de novo SNVs and indels near SV breakpoints, the majority of which were intragenic[152]. These studies imply and show the potential for multimodal integration to provide insight into the biological mechanisms affected by SVs.

Ideally, the field moves towards integration across multiple layers that can reveal relationships and reconstruct molecular contexts (for a strategy that can be generalized to functionally interpret SVs within multiple molecular contexts, see figure 6 in REF.[8]). Linked reads found that the androgen receptor gene *AR* was co-amplified with upstream tandem duplications in cancer cells[153]. Whereas DNase I hypersensitivity peaks and increased nucleosome spacing predicted

an enhancer within the duplicated region, Hi-C data showed that the duplications and *AR* lie within the same topologically associating domain (TAD), and paired RNA sequencing (RNA-seq) revealed increased expression of *AR* in samples with the upstream SV. Together, the findings were indicative of duplication of a distal enhancer element that resulted in upregulation of the oncogene (FIG. 3). In another example, investigators combined short reads, optical mapping and Hi-C to detect large and complex SVs in cancer cells, which can possibly disrupt the TAD structure[8,91,154]. RNA-seq analysis of cancer genes within disrupted TADs revealed that TADs containing an SV show greater allelic bias and altered gene expression in *cis*, suggesting that the SVs

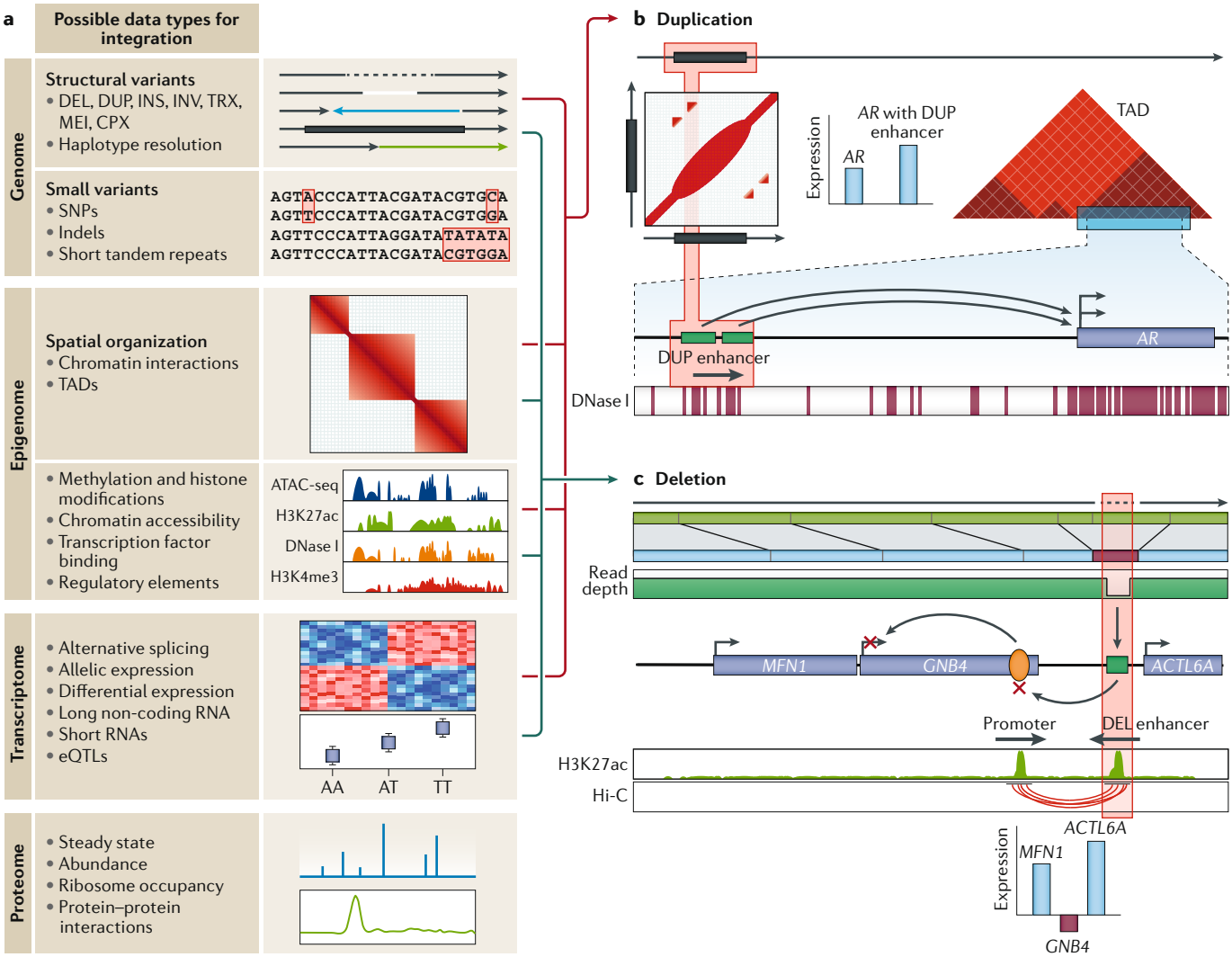



Fig. 3 | **Resolving the molecular context behind structural variants by integrating multimodal information. a** | Layers of biological data that can be integrated with structural variation (SV) calls to interpret a possible mechanistic chain of events. Each layer possesses quantifiable readouts that can be tested for association with genomic variants. Studies have focused less on the integration with more distal layers, such as the proteome, metabolome and microbiome (later two not shown), but future efforts focused here should have just as much potential to be informative. **b** | Linked reads detect tandem duplications (DUPs) upstream of *AR*[153]. Previous studies showed that this region contains an enhancer (green boxes) for *AR*, which is consistent with DNase I hypersensitivity peaks. Hi-C (high-throughput chromosome conformation capture) analysis showed that both the enhancer and the gene body are located within the same topologically associating domain (TAD), further suggesting their interaction. Paired expression data from multiple samples showed that DUP of the enhancer leads to increased *AR* expression when compared with cases without the DUP. Integration of layered data suggests that tandem DUPs cause gain of an enhancer element that drives *AR* expression in castration-resistant prostate cancer. **c** | A 3.4-kb deletion (DEL) was detected in the T47D cancer cell line by optical mapping and the read depth from short-read high-throughput sequencing[91]. The authors used histone 3 lysine 27 acetylation (H3K27ac) chromatin immunoprecipitation followed by sequencing (ChIP–seq) data to determine that the DEL overlapped an enhancer element (green box) and Hi-C data to determine that the enhancer interacts with an upstream promoter (yellow oval) to regulate *GNB4*. Comparisons of expression data against human mammary epithelial cells (HMECs) revealed increased expression of nearby genes, but *GNB4* expression was notably decreased in T47D cells. This information taken together illustrates that decreased expression of *GNB4* may result from DEL of a downstream enhancer in spite of amplification of the gene body. ATAC-seq, assay for transposase-accessible chromatin using sequencing; CPX, complex rearrangement; eQTLs, expression quantitative trait loci; H3K4me3, methylation of lysine 4 on histone H3; Indels, insertions and deletions; INS, insertion; INV, inversion; MEI, mobile-element insertion; SNP, single-nucleotide polymorphism; TRX, translocation. Part **b** adapted with permission from REF.[153], Elsevier. Part **c** adapted from REF.[91], Springer Nature Limited.

Topologically associating domain
A spatial partition of the genome where segments within these domains are enriched for interactions with each other when compared with interactions with segments outside the domain.

Allelic bias
Gene expression that is biased towards one allele over the other.

create new TADS (so-called neo-TADs) that rewire regulatory environments. In a final example, optical mapping and short reads detected a 3.4-kb deletion in a copy number-amplified region[91]. Histone 3 lysine 27 acetylation (H3K27ac) chromatin immunoprecipitation followed by sequencing (ChIP–seq) peaks predicted that part of the removed sequence acted as an enhancer. Hi-C linked the deleted enhancer to upstream *GNB4*, and RNA-seq revealed decreased expression of *GNB4* but increased expression of all other proximal genes[91]. These relationships, discovered by integrating multimodal data, paint a clearer picture of the role of this variant in perturbing biological mechanisms (FIG. 3). These studies show immense potential and provide frameworks to interpret the effects of SVs but rely largely on manual curation.

## Conclusions

Tremendous improvements in variant calling have made the ubiquity, complexity and pertinence of SVs in human genomes clearer than ever. Many advancements have contributed to an explosion in detection, including the application of ensemble algorithms, which have been essential in characterizing SVs across populations[4,5,43–45,58], and single-molecule and connected-molecule strategies, which have enabled the detection of thousands of previously undiscoverable variants[61,66,80,85,111,114]. Indeed, we now estimate that each human genome contains >20,000 SVs, many of which are located in regions that are unmappable using short reads[61,97,103,104,108]. Each emerging platform possesses unique strengths, but they also exhibit inherent biases. A philosophical ideal would involve sequencers that read entire genomes, without bias, as a contiguous whole. Until this is possible, the integration of multiple platforms will be necessary to resolve all SVs within a given human genome. Although there are no human genomes for which all classes of SVs have been completely resolved, multiplatform discovery approaches are closing this gap[39,60].

Detection is essential to characterize individual genomes, but detection alone is not enough. Indeed, the technologies and methods discussed have resulted in an influx of detectable variants, but there is little ability to assign impact. Lists of thousands of newly detected SVs will be more useful for the field if we are able to interpret their functional effects. Thus, we believe that the field should consider concurrent detection and integration. We anticipate that moving from manual curation to the development of multivariate models generalizable to projects with layered data has great potential to provide insight into the complex genomic architecture affected by SV. Ultimately, detecting SVs is a piece of the larger puzzle that is understanding the genome, its disparate parts and all of its connections. Improvements in, and applications of, new emerging genomic technologies, and the integration of variants with disparate layers of biological information, will pave the way for a future where we understand the possible function and effects of every nucleotide in the human genome.

Published online: 15 November 2019


1. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
3. Sudmant, P. H. et al. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
4. Mills, R. E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
**This study provides one of the first frameworks for using an ensemble approach to detect structural variants as part of phase 1 for the 1KGP.**
5. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
**This paper describes the development of the 1KGP phase 3 release set, which is currently one of the largest and most diverse reference sets.**
6. Sudmant, P. H. et al. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
7. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
8. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
9. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
10. Lappalainen, T., Scott, A. J., Brandt, M. & Hall, I. M. Genomic analysis in the age of human genome sequencing. *Cell* **177**, 70–84 (2019).
11. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
12. Sharp, A. J. et al. Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
13. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**, 551–564 (2009).
14. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
15. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
16. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
17. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
18. UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
19. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
20. Exome Aggregation Consortium et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
21. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
22. Macintyre, G., Ylstra, B. & Brenton, J. D. Sequencing structural variants in cancer for precision therapeutics. *Trends Genet.* **32**, 530–542 (2016).
23. Yi, K. & Ju, Y. S. Patterns and mechanisms of structural variations in human cancer. *Exp. Mol. Med.* **50**, 98 (2018).
24. Korbel, J. O. et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
25. 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.* **19**, 1586–1592 (2009).
26. Hajirasouliha, I. et al. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics* **26**, 1277–1283 (2010).
27. Chen, K. et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 (2009).
28. 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.* **21**, 974–984 (2011).
29. Korbel, J. O. et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* **10**, R23 (2009).
30. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
31. 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.* **43**, 269–276 (2011).
32. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
33. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 32, 1220–1222 (2016).
34. Sindi, S. S., Önal, S., Peng, L. C., Wu, H.-T. & Raphael, B. J. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol.* **13**, R22 (2012).
35. Zhao, X., Emery, S. B., Myers, B., Kidd, J. M. & Mills, R. E. Resolving complex structural genomic rearrangements using a randomized approach. *Genome Biol.* **17**, 126 (2016).
36. Michaelson, J. J. & Sebat, J. forestSV: structural variant discovery through statistical learning. *Nat. Methods* **9**, 819–821 (2012).

37. Telenti, A. et al. Deep sequencing of 10,000 human genomes. *Proc. Natl Acad. Sci. USA* **113**, 11901–11906 (2016).
38. Kosugi, S. et al. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **20**, 117 (2019).
**This paper extensively compares the sensitivity of SV detection algorithms and the combinations of these algorithms**.
39. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
**This study generates one of the most comprehensive multiplatform haplotype-specific SV discovery sets and provides potential frameworks for their integration**.
40. Wong, K., Keane, T. M., Stalker, J. & Adams, D. J. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* **11**, R128 (2010).
41. Lam, H. Y. K. et al. Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat. Biotechnol.* **30**, 226–229 (2012).
42. Parikh, H. et al. svclassify: a method to establish benchmark structural variant calls. *BMC Genom.* **17**, 64 (2016).
43. Collins, R. L. et al. An open resource of structural variation for medical and population genetics. *bioRxiv* https://doi.org/10.1101/578674 (2019).
44. Abel, H. J. et al. Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. *bioRxiv* https://doi.org/10.1101/508515 (2018).
45. Hehir-Kwa, J. Y. et al. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* **7**, 12989 (2016).
46. Werling, D. M. et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50**, 727–736 (2018).
47. Chiang, C. et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
48. Larson, D. E. et al. svtools: population-scale analysis of structural variation. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz492 (2019).
49. Mimori, T. et al. iSVP: an integrated structural variant calling pipeline from high-throughput sequencing data. *BMC Syst. Biol.* **7**, S8 (2013).
50. Zarate, S. et al. Parliament2: fast structural variant calling using optimized combinations of callers. *bioRxiv* https://doi.org/10.1101/424267 (2018).
51. Mohiyuddin, M. et al. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* **31**, 2741–2744 (2015).
52. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
53. Becker, T. et al. FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biol.* **19**, 38 (2018).
54. 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.* **29**, 1134–1143 (2019).
55. Huddleston, J. & Eichler, E. E. An incomplete understanding of human genetic variation. *Genetics* **202**, 1251–1254 (2016).
56. Iafrate, A. J. et al. Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
57. Kloosterman, W. P. et al. Characteristics of de novo structural changes in the human genome. *Genome Res.* **25**, 792–801 (2015).
58. Nagasaki, M. et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.* **6**, 8018 (2015).
59. Morales, J. et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, 21 (2018).
60. Zook, J. M. et al. A robust benchmark for germline structural variant detection. *bioRxiv* https://doi.org/10.1101/664623 (2019).
**This study integrates multiple platforms to develop a gold standard reference set for SV benchmarking**.
61. Chaisson, M. J. P. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
**This is one of the first papers using PacBio for comprehensive SV discovery, detecting thousands of previously undetectable SVs, including small insertions in tandem repeats and mobile elements**.
62. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2012).
63. Medvedev, P., Stanciu, M. & Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6**, S13–S20 (2009).
64. Kitzman, J. O. et al. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29**, 59–63 (2011).
65. McCoy, R. C. et al. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLOS ONE* **9**, 13 (2014).
66. Zheng, G. X. Y. et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
**This paper is the first major study using linked reads to detect SVs in human genomes and demonstrates the ability of linked reads in phasing large haplotype blocks and detecting gene fusions**.
67. Bishara, A. et al. Read clouds uncover variation in complex regions of the human genome. *Genome Res.* **25**, 1570–1580 (2015).
68. Marks, P. et al. Resolving the full spectrum of human genome variation using linked-reads. *Genome Res.* **29**, 635–645 (2019).
69. Spies, N. et al. Genome-wide reconstruction of complex structural variants using read clouds. *Nat. Methods* **14**, 915–920 (2017).
70. Fang, L. et al. LinkedSV: detection of mosaic structural variants from linked-read exome and genome sequencing data. *bioRxiv* https://doi.org/10.1101/409789 (2019).
71. Elyanow, R., Wu, H.-T. & Raphael, B. J. Identifying structural variants using linked-read sequencing data. *Bioinformatics* **34**, 353–360 (2018).
72. Eslami Rasekh, M. et al. Discovery of large genomic inversions using long range information. *BMC Genom.* **18**, 65 (2017).
73. Karaoglanoglu, F. et al. Characterization of segmental duplications and large inversions using linked-reads. *bioRxiv* https://doi.org/10.1101/394528 (2018).
74. Xia, L. C. et al. Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acids Res.* **46**, e19 (2018).
75. Wong, K. H. Y., Levy-Sakin, M. & Kwok, P.-Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nat. Commun.* **9**, 3040 (2018).
76. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).
77. Meleshko, D., Marks, P., Williams, S. & Hajirasouliha, I. Detection and assembly of novel sequence insertions using linked-read technology. *bioRxiv* https://doi.org/10.1101/551028 (2019).
78. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).
**This Review discusses the main bioinformatics challenges faced by many of the described technologies. Topics include phasing, assembly, long-range expression and methylation**.
79. Shajii, A., Numanagić, I., Whelan, C. & Berger, B. Statistical binning for barcoded reads improves downstream analyses. *Cell Syst.* **7**, 219–226.e5 (2018).
80. Falconer, E. et al. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).
**This is the first major study showing the utility of Strand-seq for the detection of chromosomal rearrangements, along with the first application of this method in human genomes**.
81. Sanders, A. D. et al. Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Res.* **26**, 1575–1587 (2016).
**This paper is the first major work using Strand-seq to detect inversions and reveals numerous inverted loci of interest within the human genome**.
82. Hills, M., O'Neill, K., Falconer, E., Brinkman, R. & Lansdorp, P. M. BAIT: organizing genomes and mapping rearrangements in single cells. *Genome Med.* **5**, 82 (2013).
83. Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C. J. & Lansdorp, P. M. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat. Protoc.* **12**, 1151–1176 (2017).
84. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
85. Harewood, L. et al. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol.* **18**, 125 (2017).
**This is the first study detecting both large chromosomal rearrangements and copy number changes with Hi-C**.
86. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
87. Steininger, A. et al. Genome-wide analysis of interchromosomal interaction probabilities reveals chained translocations and overrepresentation of translocation breakpoints in genes in a cutaneous T-cell lymphoma cell line. *Front. Oncol.* **8**, 183 (2018).
88. Seaman, L. et al. Nucleome analysis reveals structure–function relationships for colon cancer. *Mol. Cancer Res.* **15**, 821–830 (2017).
89. Chakraborty, A. & Ay, F. Identification of copy number variations and translocations in cancer cells from Hi-C data. *Bioinformatics* **34**, 338–345 (2018).
90. Zhang, X. et al. Local and global chromatin interactions are altered by large genomic deletions associated with human brain development. *Nat. Commun.* **9**, 5356 (2018).
91. Dixon, J. R. et al. Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).
**This study integrates three platforms, showing that their combination is necessary to detect the range of SVs in cancer genomes, and describes the only algorithm that currently detects most SV types with Hi-C**.
92. Díaz, N. et al. Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. *Nat. Commun.* **9**, 4938 (2018).
93. Lee, H. & Schatz, M. C. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics* **28**, 2097–2105 (2012).
94. Stephens, Z., Wang, C., Iyer, R. K. & Kocher, J.-P. Detection and visualization of complex structural variants from long reads. *BMC Bioinform.* **19**, 508 (2018).
95. English, A. C., Salerno, W. J. & Reid, J. G. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinform.* **15**, 180 (2014).
96. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
97. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
98. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).
99. Fang, L., Hu, J., Wang, D. & Wang, K. NextSV: a meta-caller for structural variants from low-coverage long-read sequencing data. *BMC Bioinform.* **19**, 180 (2018).
100. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
101. Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genom. Proteom. Bioinform.* **13**, 278–289 (2015).
102. Pendleton, M. et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
103. Shi, L. et al. Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* **7**, 12065 (2016).
104. Seo, J.-S. et al. De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).
105. Ameur, A. et al. De novo assembly of two Swedish genomes reveals missing segments from the human GRCh38 reference and improves variant calling of population-scale sequencing data. *Genes* **9**, 486 (2018).

106. Kronenberg, Z. N. et al. High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018).
107. Nagasaki, M. Construction of JRG (Japanese reference genome) with single-molecule real-time sequencing. *Hum. Genome Var.* **6**, 27 (2019).
108. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675.e19 (2019).
**This study is the most comprehensive PacBio-based SV discovery project to date, detecting variants over 15 deeply sequenced individuals and creating a call-set reference with major shared SVs**.
109. Clarke, J. et al. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).
110. Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
111. Cretu Stancu, M. et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.* **8**, 1326 (2017).
**This is the first major paper using nanopore sequencing to detect SVs in human genomes and describes the NanoSV algorithm**.
112. Gong, L. et al. Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nat. Methods* **15**, 455–460 (2018).
113. De Coster, W. et al. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Res.* **29**, 1178–1187 (2019).
114. Lam, E. T. et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012).
**This is the first major study using Bionano optical mapping to detect SVs in human genomes, leveraging the long molecules to characterize the highly polymorphic major histocompatibility complex**.
115. Schwartz, D. et al. Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262**, 110–114 (1993).
116. Teague, B. et al. High-resolution human genome structure by single-molecule analysis. *Proc. Natl Acad. Sci. USA* **107**, 10848–10853 (2010).
117. Cao, H. et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *GigaScience* **3**, 34 (2014).
118. Mak, A. C. Y. et al. Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics* **202**, 351–362 (2016).
119. Levy-Sakin, M. et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* **10**, 1025 (2019).
120. Li, L. et al. OMSV enables accurate and comprehensive identification of large structural variations from nanochannel-based single-molecule optical maps. *Genome Biol.* **18**, 230 (2017).
121. Hastie, A. R. *et al*. Rapid automated large structural variation detection in a diploid genome by nanochannel based next-generation mapping. *bioRxiv* https://doi.org/10.1101/102764 (2017).
122. Lima, L. *et al*. Comparative assessment of long-read error correction software applied to nanopore RNA-sequencing data. *Brief. Bioinform*. https://doi.org/10.1093/bib/bbz058 (2019).
123. Fu, S., Wang, A. & Au, K. F. A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biol.* **20**, 26 (2019).
124. Zhang, H., Jain, C. & Aluru, S. A comprehensive evaluation of long read error correction methods. *bioRxiv* https://doi.org/10.1101/519330 (2019)
125. Jaratlerdsiri, W. et al. Next generation mapping reveals novel large genomic rearrangements in prostate cancer. *Oncotarget* **8**, 23588–23602 (2017).
126. Xu, J. et al. An integrated framework for genome analysis reveals numerous previously unrecognizable structural variants in leukemia patients' samples. *bioRxiv* https://doi.org/10.1101/563270 (2019).
127. Zhou, B. et al. Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res.* **29**, 472–484 (2019).
128. Zhou, B. et al. Haplotype-resolved and integrated genome analysis of the cancer cell line HepG2. *Nucleic Acids Res.* **47**, 3846–3861 (2019).
129. Chan, E. K. F. et al. Optical mapping reveals a higher level of genomic architecture of chained fusions in cancer. *Genome Res.* **28**, 726–738 (2018).
130. English, A. C. et al. Assessing structural variation in a personal genome—towards a human reference diploid genome. *BMC Genom.* **16**, 286 (2015).
**This study is one of the first applications of hybrid assembly for structural variant detection, showing highly increased sensitivity from platform integration**.
131. Ritz, A. et al. Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics* **30**, 3458–3466 (2014).
132. Fan, X., Chaisson, M., Nakhleh, L. & Chen, K. HySA: a Hybrid Structural variant Assembly approach using next-generation and single-molecule sequencing technologies. *Genome Res.* **27**, 793–800 (2017).
133. Weischenfeldt, J. et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2017).
134. McPherson, A. et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLOS Comput. Biol.* **7**, e1001138 (2011).
135. McPherson, A. et al. nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Res.* **22**, 2250–2261 (2012).
136. Yorukoglu, D. et al. Dissect: detection and characterization of novel structural alterations in transcribed sequences. *Bioinformatics* **28**, i179–i187 (2012).
137. Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).
138. Gheldof, N. et al. Structural variation-associated expression changes are paralleled by chromatin architecture modifications. *PLOS ONE* **8**, e79973 (2013).
139. Fudenberg, G. & Pollard, K. S. Chromatin features constrain structural variation across evolutionary timescales. *Proc. Natl Acad. Sci. USA* **116**, 2175–2180 (2019).
140. Quigley, D. A. et al. Genomic hallmarks and structural variation in metastatic prostate cancer. *Cell* **174**, 758–769.e9 (2018).
141. Stranger, B. E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
142. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
143. Merker, J. D. et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet. Med.* **20**, 159–163 (2018).
144. Miao, H. et al. Long-read sequencing identified a causal structural variant in an exome-negative case and enabled preimplantation genetic diagnosis. *Hereditas* **155**, 32 (2018).
145. Roberts, D. S. et al. Linked-read sequencing analysis reveals tumor-specific genome variation landscapes in neurofibromatosis type 2 (NF2) patients. *Otol. Neurotol.* **40**, e150–e159 (2019).
146. Sanchis-Juan, A. *et al*. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med*. **10**, 95 (2018).
147. Cantsilieris, S. et al. Recurrent structural variation, clustered sites of selection, and disease risk for the complement factor H (*CFH*) gene family. *Proc. Natl Acad. Sci. USA* **115**, E4433–E4442 (2018).
148. Nattestad, M. et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res.* **28**, 1126–1135 (2018).
149. Aneichyk, T. et al. Dissecting the causal mechanism of X-linked dystonia–parkinsonism by integrating genome and transcriptome assembly. *Cell* **172**, 897–909.e21 (2018).
150. Sharim, H. et al. Long-read single-molecule maps of the functional methylome. *Genome Res.* **29**, 646–656 (2019).
151. Lee, I. et al. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *bioRxiv* https://doi.org/10.1101/504993 (2019).
152. Beck, C. R. et al. Megabase length hypermutation accompanies human structural variation at 17p11.2. *Cell* **176**, 1310–1324.e10 (2019).
153. Viswanathan, S. R. et al. Structural alterations driving castration-resistant prostate cancer revealed by linked-read genome sequencing. *Cell* **174**, 433–447.e19 (2018).
**This study leverages layered biological information to understand the role of SVs in oncogene amplification for a specific cancer type**.
154. Huynh, L. & Hormozdiari, F. TAD fusion score: discovery and ranking the contribution of deletions to genome structure. *Genome Biol*. **20**, 60 (2019).
155. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
156. Sebat, J. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
157. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
158. McCarroll, S. A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
159. Kidd, J. M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
160. Zhou, B. et al. Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *J. Med. Genet.* **55**, 735–743 (2018).
161. Speicher, M. R. & Carter, N. P. The new cytogenetics: blurring the boundaries with molecular biology. *Nat. Rev. Genet.* **6**, 782–792 (2005).
162. Lee, C., Iafrate, A. J. & Brothman, A. R. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat. Genet.* **39**, S48–S54 (2007).
163. Scherer, S. W. et al. Challenges and standards in integrating surveys of structural variation. *Nat. Genet.* **39**, S7–S15 (2007).
164. Tattini, L., D'Aurizio, R. & Magi, A. Detection of genomic structural variants from next-generation sequencing data. *Front. Bioeng. Biotechnol.* **3**, 92 (2015).
165. Guan, P. & Sung, W.-K. Structural variation detection using next-generation sequencing data. *Methods* **102**, 36–49 (2016).
166. Quinlan, A. R. & Hall, I. M. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet.* **28**, 43–53 (2012).
167. Tan, R. et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum. Mutat.* **35**, 899–907 (2014).
168. Hehir-Kwa, J. Y., Tops, B. B. J. & Kemmeren, P. The clinical implementation of copy number detection in the age of next-generation sequencing. *Expert. Rev. Mol. Diagn.* **18**, 907–915 (2018).
169. Hehir-Kwa, J. Y., Pfundt, R. & Veltman, J. A. Exome sequencing and whole genome sequencing for the detection of copy number variation. *Expert. Rev. Mol. Diagn.* **15**, 1023–1032 (2015).
170. Pang, A. W. et al. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol*. **11**, R52 (2010).
171. Park, H. et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.* **42**, 400–405 (2010).
172. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
173. Anderson-Trocmé, L. et al. Legacy data confounds genomics studies. *Mol. Biol. Evol*. https://doi.org/10.1093/molbev/msz201 (2019).
174. Lappalainen, I. et al. dbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.* **41**, D936–D941 (2012).
175. Demaerel, W. et al. The 22q11 low copy repeats are characterized by unprecedented size and structure variability. *Genome Res.* **29**, 1389–1401 (2019).
176. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
177. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
178. Jiang, T., Liu, B., Li, J. & Wang, Y. rMETL: sensitive mobile element insertion detection with long read realignment. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz106 (2019).
179. Meng, G. et al. TSD: a computational tool to study the complex structural variants using PacBio targeted sequencing data. *G3* **9**, 1371–1376 (2019).
180. Frith, M. C. & Khan, S. A survey of localized sequence rearrangements in human DNA. *Nucleic Acids Res.* **46**, 1661–1673 (2018).
181. Greer, S. U. & Ji, H. P. Structural variant analysis for linked-read sequencing data with gemtools. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz239 (2019).
182. Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V. & Bafna, V. Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res.* **28**, 1709–1719 (2018).
183. Ummat, A. & Bashir, A. Resolving complex tandem repeats with long reads. *Bioinformatics* **30**, 3491–3498 (2014).

184. Liu, Q., Zhang, P., Wang, D., Gu, W. & Wang, K. Interrogating the “unsequenceable” genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med.* **9**, 65 (2017).
185. Shao, H. et al. npInv: accurate detection and genotyping of inversions using long read sub-alignment. *BMC Bioinform.* **19**, 261 (2018).
186. Mitsuhashi, S. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol.* **20**, 58 (2019).
187. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **7**, 233–245 (2007).
188. Zhang, Q. et al. Clinical application of single-molecule optical mapping to a multigeneration FSHD1 pedigree. *Mol. Genet. Genom. Med.* **7**, e565 (2019).
189. Norris, A. L., Workman, R. E., Fan, Y., Eshleman, J. R. & Timp, W. Nanopore sequencing detects structural variants in cancer. *Cancer Biol. Ther.* **17**, 246–253 (2016).
190. Euskirchen, P. et al. Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathol.* **134**, 691–703 (2017).
191. Jacobson, E. C. et al. Hi-C detects novel structural variants in HL-60 and HL-60/S4 cell lines. *Genomics* https://doi.org/10.1016/j.ygeno.2019.05.009 (2019).
192. Greer, S. U. et al. Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome Med.* **9**, 57 (2017).
193. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
194. Marshall, C. R. et al. Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
195. Sullivan, P. F. & Geschwind, D. H. Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. *Cell* **177**, 162–183 (2019).
196. Yuen, R. K. et al. Genome-wide characteristics of de novo mutations in autism. *Npj Genomic Med.* **1**, 160271–1602710 (2016).
197. Brand, H. et al. Paired-duplication signatures mark cryptic inversions and other complex structural variation. *Am. J. Hum. Genet.* **97**, 170–176 (2015).
198. Turner, T. N. et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am. J. Hum. Genet.* **98**, 58–74 (2016).
199. Brandler, W. M. et al. Paternally inherited *cis*-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018).
200. Turner, T. N. et al. Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710–722.e12 (2017).
201. Mizuguchi, T. et al. Detecting a long insertion variant in SAMD12 by SMRT sequencing: implications of long-read whole-genome sequencing for repeat expansion diseases. *J. Hum. Genet.* **64**, 191–197 (2019).
202. Mizuguchi, T. et al. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. *J. Hum. Genet.* **64**, 359–368 (2019).
203. Barseghyan, H. et al. Next-generation mapping: a novel approach for detection of pathogenic structural variants with a potential utility in clinical diagnosis. *Genome Med.* **9**, 90 (2017).
204. Collins, R. L. et al. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.* **18**, 36 (2017).
205. Eisfeldt, J. et al. Comprehensive structural variation genome map of individuals carrying complex chromosomal rearrangements. *PLOS Genet.* **15**, e1007858 (2019).
206. Dutta, U. R. et al. Breakpoint mapping of a novel de novo translocation t(X;20)(q11.1;p13) by positional cloning and long read sequencing. *Genomics* **111**, 1108–1114 (2019).
207. Sherman, R. M. et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).
208. Zhou, B. et al. Extensive and deep sequencing of the Venter/HuRef genome for developing and benchmarking genome analysis tools. *Sci. Data* **5**, 180261 (2018).
209. Levy, S. et al. The diploid genome sequence of an individual human. *PLOS Biol.* **5**, e254 (2007).
210. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *bioRxiv* https://doi.org/10.1101/735928 (2019).
211. Wang, Y.-C. et al. High-coverage, long-read sequencing of Han Chinese trio reference samples. *Sci. Data* **6**, 91 (2019).
212. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).

**Acknowledgements**
The authors thank Y. Wang, W. Zhou, A. Weber and B. Zhou for their valuable comments and help with proofreading the manuscript. S.S.H. was supported through the Michigan Predoctoral Training in Genetics grant (T32 GM007544). A.E.U. acknowledges funding by the National Institutes of Health (NIH) and the Simons Foundation, and is a Tashia and John Morgridge Faculty Scholar of the Stanford Child Health Research Institute.

**Author contributions**
S.S.H. and R.E.M researched the literature and wrote the article. All authors provided substantial contributions to discussions of the content, and reviewed and/or edited the manuscript before submission.

**Competing interests**
The authors declare no competing interests.

**Peer review information**
*Nature Reviews Genetics* thanks C. Alkan, F. Sedlazeck and M. Talkowski for their contribution to the peer review of this work.

**Publisher’s note**


**Supplementary information**
Supplementary information is available for this paper at https://doi.org/10.1038/s41576-019-0180-9.