

A decade of structural variants: description, history and methods to detect structural variation

Geòrgia Escaramís, Elisa Docampo and Raquel Rabionet

Corresponding author. Raquel Rabionet, Genomics and Disease Group, Bioinformatics and Genomics, Centre for Genomic Regulation, Dr. Aiguader, 88, 08003 Barcelona, Spain. Tel.: +34 933160233; Fax: +34 933160099; E-mail: Kelly.Rabionet@crg.eu

Abstract

In the past decade, the view on genomic structural variation (SV) has been changed completely. SVs, previously considered rare events, are now recognized as the largest source of interindividual genetic variation affecting more bases than single nucleotide polymorphisms, variable number of tandem repeats and other small genetic variants. They have also been shown to play a role in phenotypic variation and in disease. In this review, the authors will provide an introduction to SV; a short historical perspective on the research of this source of genomic variation; a description of the types of structural variants, and on how they may have arisen; and an overview on methods of detecting structural variants, focusing on the analysis of high-throughput sequencing data.

Key words: structural variation; copy number variants

Structural variation (SV) is generally defined as a region of DNA that shows a change in copy number (deletions, insertions and duplications), orientation (inversions) or chromosomal location (translocations) between individuals (Figure 1a). SV can be balanced, with no loss or gain of genetic material, such as inversions of a genetic fragment or translocations of a stretch of DNA within or between chromosomes, or unbalanced, where a part of the genome is lost or duplicated, which is termed copy number variation (CNV). Chromosomal aneuploidies would be, in fact, extreme cases of unbalanced SV.

History of SV

The existence of genomic SV in the human genome has been known for a long time, since the use of karyotyping and the identification of chromosomal aneuploidies causing various syndromes, such as Klinefelter's [1] or Down's [2]. Other large structural variants that were visible at the cytogenetic level have also been described for several years [3–6], as well as some

gene duplication events, such as the alpha-globin gene [7]. In fact, a gene duplication in *Drosophila* was detected as early as 1936 [8]. These microscopically visible types of variation (~3 Mb or more in size) are relatively uncommon and usually associated with disease. Thus, until the middle of the past decade, it was thought that most of the common human genetic variation was due to single nucleotide polymorphisms (SNPs) and small insertion/deletion polymorphisms (indels), repetitive elements such as mini- and microsatellites, and small insertions or deletions (<1 Kb) that were detectable by traditional molecular biology techniques, while larger structural variants represented only rare events.

About 10 years ago, thanks to development of technologies such as BAC array-comparative genomic hybridization (array-CGH) [9] and ROMA [10], it was finally possible to perform finer analysis of genome SV. In 2004, two studies, by Sebat *et al.* [11] and Iafrate *et al.* [12], indicated that intermediate size structural variants, not visible at the karyotype level, represented a substantial amount of genetic variation, and had yet been

Geòrgia Escaramís holds a PhD in Statistics. She is a postdoctoral fellow at the Centre for Genomic Regulation, where she applies and develops statistical tools to identify the genetic causes of complex disorders.

Elisa Docampo is an MD, PhD. She is a postdoctoral fellow at the unit of Animal Genomics, GIGA_R & Faculty of Veterinary Medicine (University of Liège). Her current research focuses on the genetics of Crohn's disease.

Raquel Rabionet holds a PhD in Biology. Her research at the Centre for Genomic Regulation focuses on the identification of genetic causes of complex diseases.

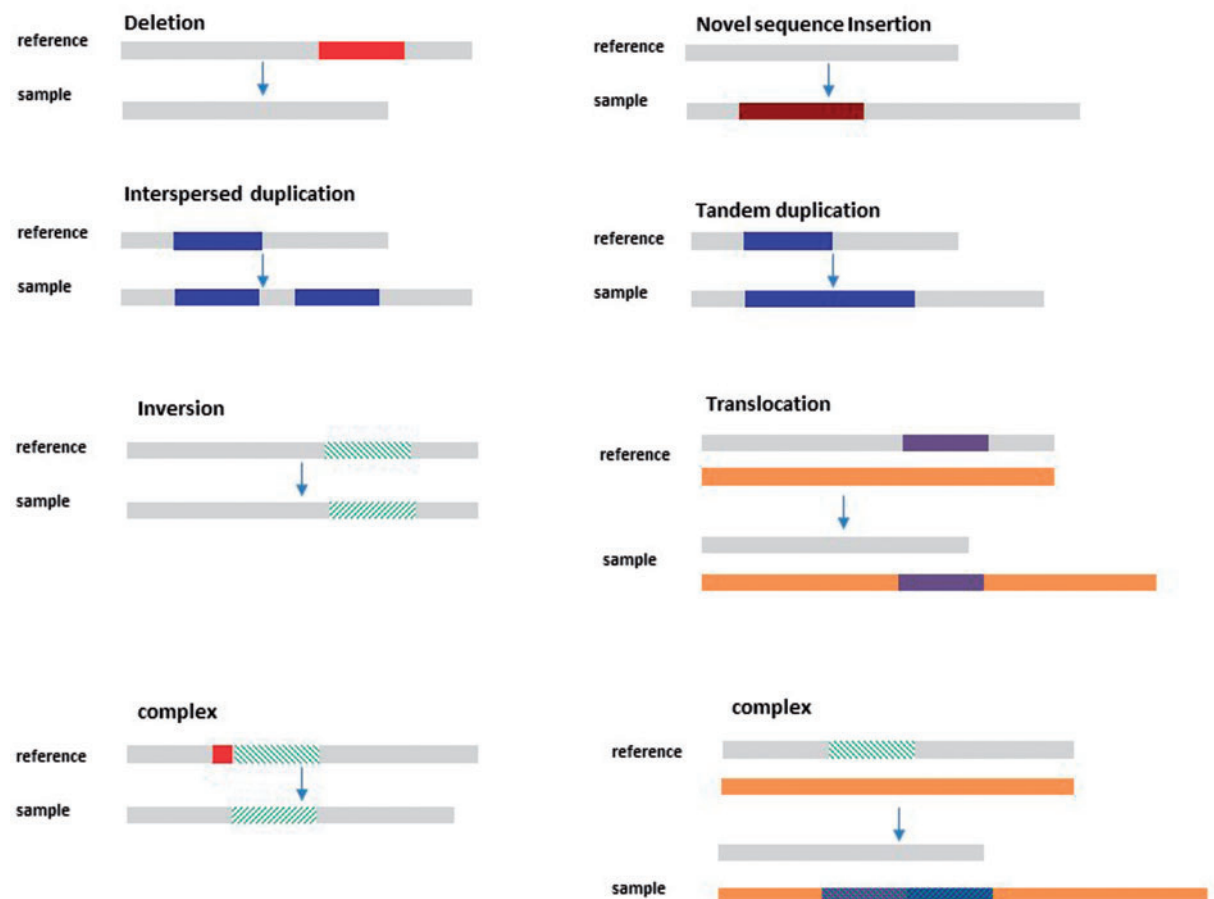


Figure 1. Types of structural variants. Unbalanced SVs represented in the top two rows include deletion, insertion of novel sequence and duplication (tandem duplication and interspersed duplication). Balanced SVs are represented in the third row and include inversions and translocations. Examples of complex SVs are presented in the bottom row. (A colour version of this figure is available online at: <http://bfg.oxfordjournals.org>)

unexplored. Both papers described the presence of common large CNVs in phenotypically normal individuals, opening a new field in genetic analysis. Several other studies followed in their steps, describing hundreds of insertion, deletion and inversion polymorphisms [13–22] in both phenotypically normal and abnormal individuals. By 2006, it was already speculated that these types of variations contributed at least as much as SNPs and variable number of tandem repeats (both mini- and microsatellites), if not more, to interindividual genetic variability [23]. It was also clear that these novel type of variants, especially CNVs, had the potential to affect gene dosage, and thus could be related to genetic disease, as was the case for the several variants involved in neurodevelopmental disorders [14, 18, 24, 25].

In fact, SV in human genomes has often been related to disease. In some cases, involving relatively rare SVs with large effect in rare disorders; in others, relatively common SVs have been shown to have smaller effects in complex diseases. SVs associated to complex disease include the involvement of a defensin gene cluster CNV [26] and another near IRGM [27] in predisposition to Crohn's disease, the relationship between a CNV in the FCGR gene region and various autoimmune disorders [28–30], the association of a CNV at the LCE gene cluster and risk of psoriasis [31] or the influence in susceptibility to HIV infection of a CNV of CCL3L1 [32]. On the other hand, the duplication of SNCA in Parkinson's disease [33] or that of APP in Alzheimer's disease [34] would be examples of CNVs with large

effect. In addition, it has been described that multiple rare, large CNVs are individually associated with neurodevelopmental phenotypes, such as intellectual disability [14, 18], autism [24, 25] or schizophrenia [35, 36], among others. In other instances, CNVs may affect a phenotype not related to disease, such as the correlation of CNVs affecting the amylase gene cluster and the amount of α -amylase in saliva and its relation to diet and starch digestion [37]. Finally, although most studies encompassing SV involve unbalanced variation, it has also been reported that balanced SV affect disease risk when their breakpoints affect relevant genes [38–40], and they have also been shown to increase risk of further, disease-causing, rearrangements [40, 41]. Nevertheless, not all structural variants are related to damaging phenotypes, by far. On the contrary, it is a small proportion of the total number of known structural variants that has so far been related to disease. In this regard, SV failed to reach the expectations that were raised as a possible answer to the missing heritability of common disorders [42]. For a more detailed account of the relation of structural variants to disease, refer to the accompanying reviews by Usher and McCarroll [43], Tubio [44], Iyer and Girirajan [45] and Puig et al. [46].

In the past few years, the definition of structural variants and their breakpoints has been greatly improved with the use of next-generation sequencing (NGS) technologies and the development of algorithms to detect structural variants more accurately in whole genome and exome data [47]. This has prompted large consortia to tackle the systematic analysis of SV. Thus, in

a similar way to what has happened with single variants analysis, the 1000 Genomes project (www.1000genomes.org), launched in 2008, aims to include structural variants in its catalog of human genetic variation. Currently, high-throughput sequencing data for 1092 samples from 14 populations have been published, including both low-coverage whole-genome sequencing and high-coverage targeted exome sequencing, and these have been analyzed for SNPs, indels and large deletions [48, 49]. Phase 3 data will include >2500 samples from 26 populations, some of which is already available. The Genome of the Netherlands project (<http://www.nlgenome.nl/>) has also characterized structural variants (deletions) in 250 Dutch trios [50, 51], including 30–500-bp events, which were previously poorly characterized, and is working on the characterization of other types of SV. The PanCancer Analysis of Whole Genomes, a joint International Cancer Genome Consortium – The Cancer Genome Atlas ongoing initiative, is also interested in defining structural variants, and although it is focused on cancer genomes, it will also provide relevant insight into the landscape of SV in the germline of a large number of genomes.

Several initiatives have attempted to generate comprehensive catalogs of the identified SVs. The Database of Genomic Variants (<http://dgv.tcag.ca/>) has been collecting data on SV since 2004. It now includes data from more than 55 publications, compiling benign variants (i.e. not overlapping with DECIPHER phenotypes; <https://decipher.sanger.ac.uk/index>) that are at least 50 bp and less than 3 MB (for CNVs) or less than 10 Mb (for inversions), that map to the autosomes or sex chromosomes, and that do not span gaps in the reference genome. It comprises >2.5 million entries in more than 22 300 individuals [52]. Most of these SVs range in size from 100 bp to 50 Kb and, globally, they overlap with approximately 60% of the genome, taking into account that the boundaries of these variants are not well-defined, and thus there could be an over-estimation in the actual size of the variants. The InvFEST project (<http://invfestdb.uab.cat/>; [53]) has collected predictions on human polymorphic inversions, and provides a database of non-redundant inversions, including data on experimental validation, population frequency, functional effects and evolutionary history, where available. Currently, 1092 predicted inversions have been included in the database. Of these, most have a ‘predicted’ (48%) or ‘unreliable prediction’ (39%) status, 85 (8%) have a validated status and 51 (5%) are considered false predictions.

The realization that structural genomic variation was common among humans has led to the study of CNVs in other species, mostly through microarray technology. Lack of, or incomplete, reference genomes has hampered these studies. Nevertheless, several studies have identified SVs in other organisms: Graubert et al. (2007) [54] characterized 80 CNVs across the genomes of 21 inbred strains of mice, and Yalcin et al. (2011) [55] performed a deeper characterization of structural variants in 17 mouse strains. Guryev et al. (2008) [56] characterized CNVs in several rat inbred and outbred strains, while maps of SV in cattle [57, 58] sheep [59], pig [60] and horse [61, 62] genomes have been recently published. Indeed, some of these structural variants have been shown to have a phenotypic effect [63, 64].

Types of SV

Main types of structural variants considered here involve: *translocations*, where a segment of DNA changes its position, intra- or interchromosomally, without gain or loss of genetic material; *inversions*, defined as segments of DNA that are reversed in orientation from the rest of the chromosome, termed *pericentric* if

they involve the centromere or *paracentric* otherwise; *insertions* of novel sequence with respect to a reference genome, including mostly mobile element insertions (MEIs); and CNVs, where a segment of DNA is present in a variable number of copies when compared to a reference genome. The latter include *deletions*, when there is a loss of genetic material with respect to the reference, and *duplications* of a given sequence, which can be inserted contiguously (in *tandem*) or elsewhere in the genome. Other types of variations that could also be considered structural variants are isochromosomes, ring chromosomes, fragile sites, segmental uniparental disomies and double minutes. In addition, complex rearrangements might involve inversions or translocations flanked by CNV, or other combinations of events. A particular type of complex rearrangements are those formed by chromothripsis [65], a process whereby multiple classes of structural rearrangements occur simultaneously, generating multiple clustered deletion, translocation and inversion events.

Generation of structural variants

Several mutational mechanisms can lead to the generation of SV, both meiotically and mitotically (as shown by CNVs between identical twins). These include recombination errors, in particular non-allelic homologous recombination (NAHR) [66]; errors generated in DNA break repair, as in non-homologous end joining (NHEJ) [67] and microhomology-mediated end joining (MMEJ) [68, 69]; or errors in replication, such as fork stalling and template switching (FoSTeS) [70] or microhomology-mediated break-induced replication (MMBIR) [71] and serial replication slippage (SRS); and MEI. Each of these methods would generate a particular molecular signature in and around the breakpoints of the SV [72, 73].

NAHR: This type of non-allelic recombination events occur by misalignment of highly identical sequences during mitosis or meiosis. The main substrates of NAHR are segmental duplications, although other long stretches of homology, such as Alu or L1 elements, may also cause NAHR. NAHR events usually occur when homologous sequences are 10 Kb or larger, have more than 95% homology and are separated by 50 Kb to 10 Mb. A recent study by Girirajan et al. (2013) [74] has broadened the NAHR ‘hotspots’ definition to include smaller homology regions, as short as 200 bp, and as close together as 1 Kb. Depending on the location and orientation of the homology regions, NAHR can cause deletions and duplications, when they are in direct orientation within the same chromosome; inversions, when they are in opposite orientation on the same chromosome; and translocations, when they are located on different chromosomes [75–77]. This mechanism of SV formation is more common in recurrent, large SVs, sharing the same breakpoints, which are often flanked by segmental duplications [72].

The majority of non-recurrent events are caused by a variety of different mechanisms:

NHEJ: Non-homologous end joining is the preferred method of double-strand break repair in mammals, especially during G0 and G1 phases. This is a very fast process that fuses the ends of a double-strand DNA break with little or no sequence homology (<4 bp), generating blunt joins or short insertions or deletions at the breakpoint junction. Although it potentially can lead to translocations as well as small deletions and insertions, it is not considered a major source of genomic instability [78].

MMEJ: It is a form of alternative NHEJ. It usually takes place when NHEJ and homologous replication repair mechanisms are

repressed, and it is more error-prone than NHEJ. It is often a cause of translocations, and it leads to deletions at the breakpoints. It is considered a major source of genomic instability [69].

FoSTeS/MMBIR: FoSTeS, later included in the generalized mechanism MMBIR, has been described as a mechanism for the generation of complex rearrangements, as well as inversions, tandem duplications or translocations. In these cases, the proposed mechanism involves stalling of the replication at the fork, and shifting of the polymerase by microhomology to any nearby single-stranded DNA [71]. These rearrangements can occur at different genomic scales, from a few kilobases to several megabases, and the microhomology regions can be as short as 2 bp [73, 79].

SRS: Smaller complex rearrangements can be caused by serial replication slippage [80], which would be explained by multiple rounds of forward and backward replication slippage. In replication slippage, when replicating regions with short direct repeats, there is a transient dissociation of the primer and template strands followed by misaligned reassociation, leading to loss (forward slippage) or duplication (backward slippage) of the region within the repeats.

Some other mechanisms are being proposed for other large complex rearrangements. For instance, a *foldback inversion* mechanism has been proposed to explain the generation of inverted duplications adjacent to terminal deletions. This involves an initial double-strand break, loss of the chromosome end and generation of a 3' overhang, that then folds and pairs with itself at regions of microhomology. The gap is filled by DNA synthesis, and a monocentric foldback chromosome is generated, that may resolve into the inverted duplication and terminal deletion [81].

In the case of *chromothripsis*, a single catastrophic event leads to the shattering of a chromosomal region followed by non-homologous DNA repair, which generates a 'patchwork' of genomic segments. To infer a chromothripsis event, the pattern of detected SV must meet six criteria, detailed in [82]. This occurs in 2–3% of all cancers [65], but it has also been described in the germline, in individuals with multiple congenital abnormalities [83, 84]. Constitutional chromothripsis events tend to be more balanced than those in cancer.

Finally, MEIs have had an important evolutionary role in shaping mammalian genomes. Although most annotated mobile elements in the human genome have lost the capacity of active retrotransposition, there remain a few active retrotransposons, mostly from the *Alu*, *L1* and *SVA* families [85]. In addition, as copies of mobile elements retain high levels of homology, they also play a role in generation of structural variants as the substrate of NAHR or MMBIR.

As we have mentioned, the generation of SV can occur both meiotically and mitotically. Thus, monozygotic twins can carry differences in SV [86], and individuals can be mosaic carriers of structural variants, between tissues and even within tissues [87–89], even in the central nervous system, where it was detected that between 13 and 41% of neurons harbored at least one megabase-scale *de novo* CNV [90].

Detection of SV

The most well-studied types of structural variants are those involving change in copy number of genetic material, as those are easier to detect with current technologies. Detection of balanced structural variants, aside from those large enough to be

visible on a karyotype, generally involves the analysis of paired-end NGS data and the application of complex algorithms.

In 2004, the field of SV analysis was opened thanks to the development of array-CGH, which remains a robust method for discovery of CNV. Array-CGH is based on the analysis of intensity ratios of the hybridization of two differentially dyed DNAs against the same target oligonucleotides. Different algorithms were developed to obtain CNV information from these data, such as circular binary segmentation [91], and Genome Alteration Detection Algorithm [92]. One of the disadvantages of array-CGH is that it detects imbalances between two individual genomes, and thus it cannot provide an absolute copy number. Array-CGH is also blind to balanced SV events. Commercial tools for array-CGH analysis are offered by several companies. CNV analysis through array-CGH has been implemented in clinical diagnostics, mostly in the field of prenatal medicine, and in diagnosis and genetic counseling of intellectual disability, autism spectrum disorder and multiple congenital abnormalities [93]. To avoid CNV regions of clinical uncertainty, it has been preferred to use targeted arrays instead of whole-genome scan methods [94, 95].

A few years later, several algorithms were developed with the aim of extracting CNV information from SNP-array data, leveraging on the large amounts of samples that had been genotyped on SNP-arrays for genome-wide association studies. These algorithms (PennCNV [96], QuantiSNP [discontinued support in 2011] [97], Birdsuite [98], among others) used the SNP intensity information, both the normalized sum of intensities of the two alleles, which is usually referred as *log R ratio*, and the intensity ratio of one allele relative to the other, which is usually referred as *B allele frequency*, to infer allele copy number. Nevertheless, despite the advantage of using 'free' data, the resolution and reliability of the CNV calls are lower than that of array-CGH, and the overlap between the existing software tools is poor [99]. SNP-array analysis offers the advantage that, unlike array-CGH, it is possible to infer the presence of certain balanced structural variants, such as inversions, from SNP-array data, as these variants affect the linkage disequilibrium pattern, leaving a detectable signature [100–102]. In addition to discovery of structural variants, it is also possible to impute certain structural variants from SNP array data, as most common bi-allelic CNVs can be tagged by nearby SNPs (79% of CNVs with >10% minor allele frequency have at least one SNP that tags them with $r^2 > 0.8$) [42].

More recently, the arrival of NGS and the rapid increase in its throughput have enabled the exploration of genomic SVs at an even finer scale. Initially, most NGS were performed with single reads (DNA is fragmented and sequenced from one end only). As technology developed, sequencing of paired-end reads (DNA is fragmented and both ends of the fragments are sequenced, with or without a non-sequenced stretch in between) became more common. Paired-end reads are better aligned to the reference and offer additional information for detection of structural variants. This has allowed the detection of small CNVs, which were missed due to low resolution in array-CGH or SNP arrays, and it has enabled the genome-wide characterization of breakpoints for all classes of SVs. However, the characterization of structural variants from NGS is not straightforward, and many algorithms have been developed to try to gather the most accurate information. These algorithms follow one (or a combination) of four main strategies to identify SVs from NGS data (Figure 2): read depth, paired-end, split and clip reads and *de novo* assembly.

Read depth (RD) analysis mainly identifies one class of SVs, CNVs. RD methods analyze the density of reads mapped to a

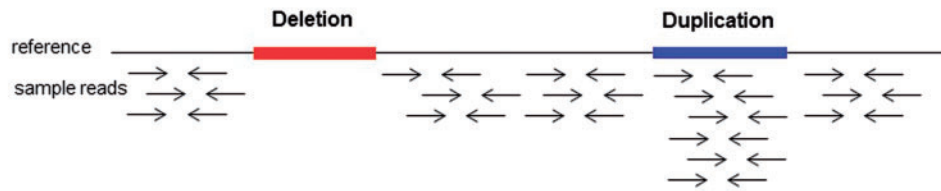
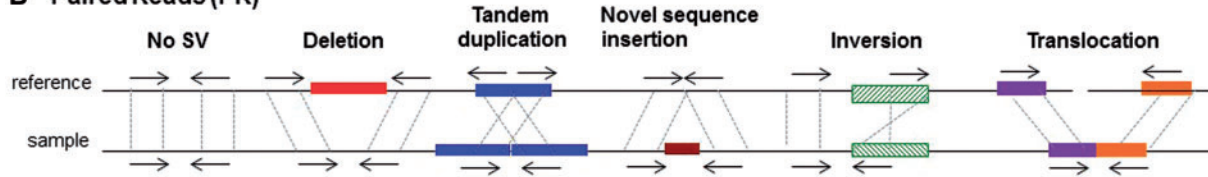
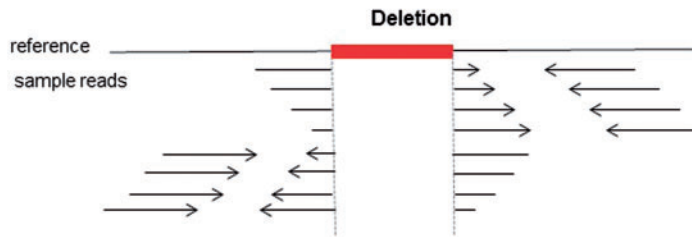
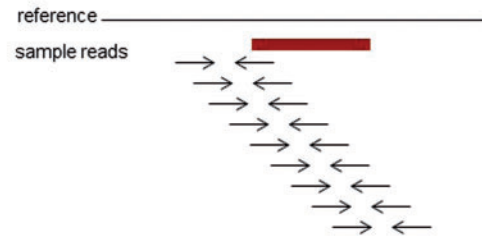
A Read Depth (RD)**B Paired Reads (PR)****C Split Reads (SR)****D. De Novo Assembly (AS)**

Figure 2. Strategies for structural variant (SV) detection. (A) Read depth. Reads are aligned into the reference genome and when compared to diploid regions they show a reduced number of reads in a deleted region or higher read depth in a duplicated region. (B) Paired reads. Pairs of sequence reads are mapped into the reference genome (from left to right): (1) no SV, pairs are aligned into correct order, correct orientation and spanned as expected based on the library's insert size; (2) deletion, the aligned pairs span far apart from that expected based on library insert size; (3) tandem duplication, read pairs are aligned in unexpected order, where expected order means that the leftmost read should be aligned in the forward strand and the rightmost read in the reverse strand; (4) novel sequence insertion, the pairs are aligned closer from that expected based on library insert size; (5) inversion, read pairs are aligned in wrong orientation, both reads align either in forward or reverse strand; and (6) read pairs mapped to different chromosomes. (C) Split reads. Sequenced reads pointing to the same breakpoint are split at the nucleotide where the breakpoint occurs. The corresponding paired read is properly aligned to the reference genome. (D) De novo assembly. Sample reads from novel sequence insertions are assembled without a reference sequenced genome. (A colour version of this figure is available online at: <http://bfg.oxfordjournals.org>)

given interval of the reference genome [103–105]. They may use single-end or paired-end reads. Although RD analysis is the only sequencing-based method to accurately predict absolute copy-numbers [103, 106], the breakpoint resolution is poor, and depends mainly on the sequence coverage and the length of the genomic intervals used. RD is further hampered by polymerase chain reaction (PCR)-induced coverage biases and is unable to detect copy-number neutral variants such as inversions and balanced translocations. An example of an RD-based algorithm is CNVnator [104].

Paired read (PR) approaches involve the identification of clusters of aberrantly mapped read pairs, suggesting the presence of SV breakpoints between the reads [107–111]. These discordantly mapped paired-reads may be a) further apart (or closer together) than expected based on the library's insert size, b) in inverted orientation, c) in incorrect order (pointing apart from each other in the reference genome) or d) mapping in different chromosomes. Each of these possibilities is an indicator of different types of structural variants, e.g. a) could indicate deletions or insertions occurring between the resulting alignment of the paired ends, or b) would indicate translocations. The resolution of the breakpoints in this approach will depend on the library's insert size mean and standard deviation, and on the coverage. Pitfalls

of this approach include reads mapping to repetitive regions or the presence of SNPs or other sequence features mapping close to the breakpoints. It is also limited in the detection of insertions, as it can't detect insertions larger than the average insert size of the library. An example of a tool using this approach is Breakdancer [108]. A particular subset of the paired-end approach is the mate-pair strategy [111]. It generates a fosmid-like library by fragmenting DNA into large pieces (>2 Kb), circularizing it and fragmenting it again and capturing and sequencing both ends of the 'junction' fragment. Analysis is similar to that of paired-end reads, but orientation of the reads is inverted. However, this strategy has not been widely used.

Split reads (SR) [112] and clip reads (CR) [113, 114] analysis are aimed at the direct identification of sequence reads that span the breakpoints of SVs. These approaches provide, by definition, single nucleotide-level resolution, although the existence of micro-homologies at fusion sites reduces the accuracy to 1–10 nucleotides, while large imperfect homologies at the breakpoint may cause even lower accuracy [114]. The SR strategy has had a limited application in the analysis of NGS data due to the difficulty of aligning short reads (resulting from splitting a read), issues when spanning large gaps and the need of a higher coverage depth to obtain sufficient SRs overlapping the

breakpoints to achieve a confident call. A tool based on this approach is Socrates [114].

De novo sequence assembly (AS) enables the fine-scale discovery of SVs, including novel (non-reference) sequence insertions [48, 115], as they do not rely on a reference sequence. AS approaches require higher computational costs, are very-time consuming and are prone to assembly errors. However, as a tool for following evidence from other approaches, targeted sequencing assembly is very valuable. An example of an AS-based tool is TIGRA [116].

Each of these approaches has its own strengths and weaknesses, but none is capable of detecting the full spectrum of structural variants [117, 118].

As each of the approaches by themselves lacks the capacity to detect all types of SVs, recent methods have begun to consider the combination of different approaches. For example, RD and PR signals have been combined to accurately detect CNVs in inGAP-sv [119] or in CNVer [120] and to detect CNVs and inversions in GASVPro [121]. HYDRA [122] and the next-generation VariationHunter [123] offer as well the capability to predict retrotransposition events. PeSV-Fisher [124] combines PR and RD to detect CNVs, inversions, intra- and inter- chromosomal translocations and further provides the colocalization of the SVs in the genome. Pindel [112], PRISM [125] and Delly [126] use PR evidence and incorporate SR proof to further define the appropriate SV.

In general, the SVs detected by these different algorithms do not show a high overlap. In fact, the 1000 Genomes project undertook a thorough evaluation [48] of several SV detection methods and found a wide variability in sensitivity and specificity. They found GenomeSTRIP [127] to perform best, given its low false discovery rate. To improve on the false discovery rate, Michaelson et al. (2014) [128] have developed a machine learning approach, forestSV, that integrates prior knowledge about the characteristics of SVs in addition to a combination of RD and PR approaches leading to improved discovery rates.

All the listed tools are designed for the detection of structural variants from whole-genome sequencing data. However, most NGS studies performed to date have taken advantage of the reduced costs (both computing and economic) of whole-exome sequencing (WES). Exome sequencing focuses on protein-coding regions, which only cover a sparse 1% of the entire genome. This scenario, involving small and discontinuous target regions (exon plus flanking regions), limits the detection capacity for structural variants. So far, only RD approaches have been successfully integrated from whole-genome sequencing to WES data [129]. The 'windows' used for read-depth calculation in whole-genome sequence have to be modified to take into account the exonic/targeted regions. In addition, given the hybridization capture and amplification steps in the preparation of exome sequencing libraries, the expected coverage of genomic regions is even less uniform than in whole-genome sequencing, raising further normalization issues. Several algorithms (CoNIFER [130] and XHMM [131], among others) address this by analyzing a batch of samples simultaneously and applying different levels of normalization procedures.

Targeted detection of CNVs for the analysis of known SV or the validation of variants identified by discovery methods include real-time quantitative PCR, multiplex ligation-dependent probe amplification [132], the use of paralogue ratios from dispersed repeats [133] and multiplex PCR spanning the breakpoints of the CNVs, if known. When the CNV events are large enough (>3 Mb), visualization through fluorescence in-situ hybridization (FISH) can be used. For a higher resolution, FiberFISH

allows the visualization of smaller events (1–400 Kb) [134]. In addition to the size limitation and in contrast with the other validation methods, FISH doesn't allow high-throughput validation. It is worth mentioning that the major challenge is the validation and accurate quantification of multicopy CNV. CNVs that involve more than three copies of genomic material are still difficult to assess with any of the techniques mentioned above.

Concluding remarks

Structural variants represent an important part of human genetic variation, and they play relevant roles in phenotypic variability and disease. Several mechanisms are involved in the generation of structural variants, which can occur meiotically and mitotically. As SVs can arise mitotically and clonally, it would be interesting to explore the role of mosaicism and tissue-specific SV in disease. Current tools are capable of characterizing all types of SV at nucleotide-level resolution, and thanks to the efforts of large consortia, it can be expected that in the next few years, a very detailed catalog of human SV will become available, and will be very useful in the functional interpretation of the genome.

Key points

- Structural variation is a common form of interindividual genetic variation.
- Balanced structural variation poses bigger challenges and has been less studied than unbalanced forms.
- Structural variants are mostly discovered through array-CGH, intensity analysis from SNP arrays and read-pair and RD analysis of high-throughput sequencing data.
- Structural variants can be generated by errors in recombination, replication or by erroneous break repair mechanisms, arising both mitotically and meiotically.

Acknowledgement

We would like to thank Laia Bassaganyas for her input.

Funding

E.D. is supported by a Marie Curie Intraeuropean Fellowship (IEF), R.R. is supported through the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 262055. We acknowledge support of the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017', SEV-2012-0208, the 'Secretaria d'Universitats i Recerca del Departament d'Economia' (2009 SGR 1502) and the 'PN de I+D+I 2008-2011' and ISCIII (SAF2008-00357). The research leading to these results reflects only the author's views and the funding agencies are not liable for any use that may be made of the information contained in this paper.

References

1. Jacobs PA, Strong JA. A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature* 1959; 183:302–3.
2. Jacobs PA, Baikie AG, Court Brown WM, et al. The somatic chromosomes in mongolism. *Lancet* 1959;1:710.

3. Nowell PC, Hungerford DA. Chromosome studies on normal and leukemic human leukocytes. *J Natl Cancer Inst* 1960;25: 85–109.
4. Manolov G, Manolova Y. Marker band in one chromosome 14 from Burkitt lymphomas. *Nature* 1972;237:33–4.
5. Mitelman F, Andersson-Anvret M, Brandt L, et al. Reciprocal 8;14 translocation in EBV-negative B-cell acute lymphocytic leukemia with Burkitt-type cells. *Int J Cancer* 1979;24:27–33.
6. Craig-Holmes AP, Moore FB, Shaw MW. Polymorphism of human C-band heterochromatin. I. Frequency of variants. *Am J Hum Genet* 1973;25:181–92.
7. Goossens M, Dozy AM, Embury SH, et al. Triplicated alpha-globin loci in humans. *Proc Natl Acad Sci USA* 1980;77:518–21.
8. Bridges CB. The bar “Gene” a duplication. *Science* 1936;83: 210–11.
9. Vissers LE, de Vries BB, Osoegawa K, et al. Array-based comparative genomic hybridization for the genome-wide detection of submicroscopic chromosomal abnormalities. *Am J Hum Genet* 2003;73:1261–70.
10. Lucito R, Healy J, Alexander J, et al. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res* 2003;13: 2291–305.
11. Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. *Science* 2004;305: 525–8.
12. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. *Nat Genet* 2004;36:949–51.
13. Tuzun E, Sharp AJ, Bailey JA, et al. Fine-scale structural variation of the human genome. *Nat Genet* 2005;37:727–32.
14. de Vries BB, Pfundt R, Leisink M, et al. Diagnostic genome profiling in mental retardation. *Am J Hum Genet* 2005;77: 606–16.
15. Dhami P, Coffey AJ, Abbs S, et al. Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *Am J Hum Genet* 2005;76: 750–62.
16. Feuk L, MacDonald JR, Tang T, et al. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet* 2005; 1:e56.
17. Sharp AJ, Locke DP, McGrath SD, et al. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* 2005;77:78–88.
18. Tyson C, Harvard C, Locker R, et al. Submicroscopic deletions and duplications in individuals with intellectual disability detected by array-CGH. *Am J Med Genet A* 2005;139: 173–85.
19. Conrad DF, Andrews TD, Carter NP, et al. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 2006;38:75–81.
20. Hinds DA, Kloek AP, Jen M, et al. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* 2006;38:82–5.
21. McCarroll SA, Hadnott TN, Perry GH, et al. Common deletion polymorphisms in the human genome. *Nat Genet* 2006;38: 86–92.
22. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature* 2006;444:444–54.
23. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet* 2006;7:85–97.
24. Sebat J, Lakshmi B, Malhotra D, et al. Strong association of de novo copy number mutations with autism. *Science* 2007;316: 445–9.
25. Marshall CR, Noor A, Vincent JB, et al. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* 2008;82:477–88.
26. Fellermann K, Stange DE, Schaeffeler E, et al. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am J Hum Genet* 2006;79:439–48.
27. McCarroll SA, Huett A, Kuballa P, et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn’s disease. *Nat Genet* 2008;40:1107–12.
28. Niederer HA, Willcocks LC, Rayner TF, et al. Copy number, linkage disequilibrium and disease association in the FCGR locus. *Hum Mol Genet* 2010;19:3282–94.
29. Fanciulli M, Norsworthy PJ, Petretto E, et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet* 2007; 39:721–3.
30. Mamtani M, Anaya JM, He W, et al. Association of copy number variation in the FCGR3B gene with risk of autoimmune diseases. *Genes Immun* 2010;11:155–60.
31. de Cid R, Riveira-Muñoz E, Zeeuwen PL, et al. Deletion of the late cornified envelope (LCE) 3C and 3B genes as a susceptibility factor for psoriasis. *Nat Genet* 2009;41:211–15.
32. Gonzalez E, Kulkarni H, Bolivar H, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 2005;307:1434–40.
33. Singleton AB, Farrer M, Johnson J, et al. alpha-Synuclein locus triplication causes Parkinson’s disease. *Science* 2003; 302:841.
34. Rovelet-Lecrux A, Hannequin D, Raux G, et al. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat Genet* 2006;38:24–6.
35. International Schizophrenia C. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 2008; 455:237–41.
36. Stefansson H, Rujescu D, Cichon S, et al. Large recurrent microdeletions associated with schizophrenia. *Nature* 2008; 455:232–6.
37. Perry GH, Dominy NJ, Claw KG, et al. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 2007;39:1256–60.
38. Talkowski ME, Rosenfeld JA, Blumenthal I, et al. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* 2012; 149:525–37.
39. Antonarakis SE, Rossiter JP, Young M, et al. Factor VIII gene inversions in severe hemophilia A: results of an international consortium study. *Blood* 1995;86:2206–12.
40. Feuk L. Inversion variants in the human genome: role in disease and genome architecture. *Genome Med* 2010; 2:11.
41. Osborne LR, Li M, Pober B, et al. A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat Genet* 2001;29:321–5.
42. Wellcome Trust Case Control C, Craddock N, Hurles ME, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 2010;464:713–20.
43. Usher CL, McCarroll SA. Complex and multi-allelic copy number variations in human disease. *Brief Funct Genomic* 2015;14:329–38.
44. Tubio JMC. Somatic structural variation and cancer. *Brief Funct Genomic* 2015;14:339–51.

45. Iyer J, Girirajan S. Gene discovery and functional assessment of rare copy-number variants in neurodevelopmental disorders. *Brief Funct Genomic* 2015;14:315–28.
46. Puig M, Casillas A, Villatoro S, Cáceres M. Human inversions and their functional consequences. *Brief Funct Genomic* 2015; 14:369–79.
47. Koboldt DC, Larson DE, Chen K, et al. Massively parallel sequencing approaches for characterization of structural variation. *Methods Mol Biol* 2012;838:369–84.
48. Mills RE, Walter K, Stewart C, et al. Mapping copy number variation by population-scale genome sequencing. *Nature* 2011;470:59–65.
49. Genomes Project C, Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061–73.
50. Boomsma DI, Wijmenga C, Slagboom EP, et al. The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* 2014;22:221–7.
51. Genome of the Netherlands C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 2014;46:818–25.
52. MacDonald JR, Ziman R, Yuen RK, et al. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 2014;42:D986–92.
53. Martinez-Fundichely A, Casillas S, Egea R, et al. InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Res* 2014;42: D1027–32.
54. Graubert TA, Cahan P, Edwin D, et al. A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet* 2007;3:e3.
55. Yalcin B, Wong K, Agam A, et al. Sequence-based characterization of structural variation in the mouse genome. *Nature* 2011;477:326–9.
56. Guryev V, Saar K, Adamovic T, et al. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet* 2008;40:538–45.
57. Fadista J, Thomsen B, Holm LE, et al. Copy number variation in the bovine genome. *BMC Genomics* 2010;11:284.
58. Bae JS, Cheong HS, Kim LH, et al. Identification of copy number variations and common deletion polymorphisms in cattle. *BMC Genomics* 2010;11:232.
59. Liu J, Zhang L, Xu L, et al. Analysis of copy number variations in the sheep genome using 50K SNP BeadChip array. *BMC Genomics* 2013;14:229.
60. Esteve-Codina A, Paudel Y, Ferretti L, et al. Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. *BMC Genomics* 2013;14:148.
61. Ghosh S, Qu Z, Das PJ, et al. Copy number variation in the horse genome. *PLoS Genet* 2014;10:e1004712.
62. Wang W, Wang S, Hou C, et al. Genome-wide detection of copy number variations among diverse horse breeds by array CGH. *PLoS One* 2014;9:e86860.
63. Durkin K, Coppieters W, Drogemuller C, et al. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* 2012;482:81–4.
64. Kadri NK, Sahana G, Charlier C, et al. A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. *PLoS Genet* 2014;10:e1004049.
65. Stephens PJ, Greenman CD, Fu B, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 2011;144:27–40.
66. Lupski JR. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* 1998;14:417–22.
67. Moore JK, Haber JE. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in *Saccharomyces cerevisiae*. *Mol Cell Biol* 1996;16:2164–73.
68. Roth DB, Wilson JH. Nonhomologous recombination in mammalian cells: role for short sequence homologies in the joining reaction. *Mol Cell Biol* 1986;6:4295–304.
69. McVey M, Lee SE. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet* 2008;24:529–38.
70. Lee JA, Carvalho CM, Lupski JR. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* 2007;131:1235–47.
71. Hastings PJ, Lupski JR, Rosenberg SM, et al. Mechanisms of change in gene copy number. *Nat Rev Genet* 2009;10: 551–64.
72. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. *Nature* 2010;464:704–12.
73. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. *Pathogenetics* 2008;1:4.
74. Girirajan S, Dennis MY, Baker C, et al. Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am J Hum Genet* 2013;92: 221–37.
75. Inoue K, Lupski JR. Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet* 2002;3:199–242.
76. Sharp AJ, Cheng Z, Eichler EE. Structural variation of the human genome. *Annu Rev Genomics Hum Genet* 2006;7: 407–42.
77. Lupski JR, Stankiewicz P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet* 2005;1:e49.
78. Ferguson DO, Sekiguchi JM, Chang S, et al. The nonhomologous end-joining pathway of DNA repair is required for genomic stability and the suppression of translocations. *Proc Natl Acad Sci USA* 2000;97:6630–3.
79. Zhang F, Carvalho CM, Lupski JR. Complex human chromosomal and genomic rearrangements. *Trends Genet* 2009;25: 298–307.
80. Chen JM, Chuzhanova N, Stenson PD, et al. Complex gene rearrangements caused by serial replication slippage. *Hum Mutat* 2005;26:125–34.
81. Hermetz KE, Newman S, Conneely KN, et al. Large inverted duplications in the human genome form via a fold-back mechanism. *PLoS Genet* 2014;10:e1004139.
82. Korbel JO, Campbell PJ. Criteria for inference of chromothripsis in cancer genomes. *Cell* 2013;152:1226–36.
83. Kloosterman WP, Guryev V, van Roosmalen M, et al. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum Mol Genet* 2011;20:1916–24.
84. Kloosterman WP, Tavakoli-Yaraki M, van Roosmalen MJ, et al. Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *Cell Rep* 2012;1:648–55.
85. Stewart C, Kural D, Stromberg MP, et al. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* 2011;7:e1002236.
86. Bruder CE, Piotrowski A, Gijsbers AA, et al. Phenotypically concordant and discordant monozygotic twins display

- different DNA copy-number-variation profiles. *Am J Hum Genet* 2008;**82**:763–71.
87. Rodriguez-Santiago B, Malats N, Rothman N, et al. Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. *Am J Hum Genet* 2010;**87**: 129–38.
 88. Abyzov A, Mariani J, Palejev D, et al. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* 2012;**492**:438–42.
 89. Laurie CC, Laurie CA, Rice K, et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat Genet* 2012;**44**:642–50.
 90. McConnell MJ, Lindberg MR, Brennand KJ, et al. Mosaic copy number variation in human neurons. *Science* 2013;**342**:632–7.
 91. Olshen AB, Venkatraman ES, Lucito R, et al. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 2004;**5**:557–72.
 92. Pique-Regi R, Monso-Varona J, Ortega A, et al. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* 2008;**24**: 309–18.
 93. Coughlin CR, 2nd, Scharer GH, Shaikh TH. Clinical impact of copy number variation analysis using high-resolution microarray technologies: advantages, limitations and concerns. *Genome Med* 2012;**4**:80.
 94. Le Caignec C, Redon R. Copy number variation goes clinical. *Genome Biol* 2009;**10**:301.
 95. Armengol L, Nevado J, Serra-Juhe C, et al. Clinical utility of chromosomal microarray analysis in invasive prenatal diagnosis. *Hum Genet* 2012;**131**:513–23.
 96. Wang L, Hauser ER, Shah SH, et al. Peakwide mapping on chromosome 3q13 identifies the kalirin gene as a novel candidate gene for coronary artery disease. *Am J Hum Genet* 2007;**80**:650–63.
 97. Colella S, Yau C, Taylor JM, et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 2007;**35**:2013–25.
 98. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 2008;**40**: 1253–60.
 99. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic* 2009;**8**: 353–66.
 100. Bansal V, Bashir A, Bafna V. Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Res* 2007;**17**:219–30.
 101. Antonacci F, Kidd JM, Marques-Bonet T, et al. Characterization of six human disease-associated inversion polymorphisms. *Hum Mol Genet* 2009;**18**:2555–66.
 102. Caceres A, Sindi SS, Raphael BJ, et al. Identification of polymorphic inversions from genotypes. *BMC Bioinformatics* 2012;**13**:28.
 103. Alkan C, Kidd JM, Marques-Bonet T, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* 2009;**41**:1061–7.
 104. Abyzov A, Urban AE, Snyder M, et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 2011;**21**:974–84.
 105. Yoon S, Xuan Z, Makarov V, et al. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* 2009;**19**:1586–92.
 106. Sudmant PH, Kitzman JO, Antonacci F, et al. Diversity of human copy number variation and multicopy genes. *Science* 2010;**330**:641–6.
 107. Korbel JO, Urban AE, Grubert F, et al. Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *Proc Natl Acad Sci USA* 2007;**104**:10110–15.
 108. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 2009;**6**:677–81.
 109. Hormozdiari F, Alkan C, Eichler EE, et al. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* 2009;**19**:1270–8.
 110. Kidd JM, Cooper GM, Donahue WF, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* 2008;**453**:56–64.
 111. Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 2007;**318**:420–6.
 112. Ye K, Schulz MH, Long Q, et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 2009;**25**:2865–71.
 113. Wang J, Mullighan CG, Easton J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* 2011;**8**:652–4.
 114. Schroder J, Hsu A, Boyle SE, et al. Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics* 2014;**30**:1064–72.
 115. Hajirasouliha I, Hormozdiari F, Alkan C, et al. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics* 2010;**26**: 1277–83.
 116. Chen K, Chen L, Fan X, et al. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome Res* 2014;**24**:310–17.
 117. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* 2009;**6**:S13–20.
 118. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011;**12**:363–76.
 119. Qi J, Zhao F. inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Res* 2011;**39**:W567–75.
 120. Medvedev P, Fiume M, Dzamba M, et al. Detecting copy number variation with mated short reads. *Genome Res* 2010;**20**: 1613–22.
 121. Sindi SS, Onal S, Peng LC, et al. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol* 2012;**13**:R22.
 122. Quinlan AR, Clark RA, Sokolova S, et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* 2010;**20**:623–35.
 123. Hormozdiari F, Hajirasouliha I, Dao P, et al. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 2010;**26**:i350–7.
 124. Escaramis G, Tornador C, Bassaganyas L, et al. PeSV-Fisher: identification of somatic and non-somatic structural variants using next generation sequencing data. *PLoS One* 2013;**8**:e63377.
 125. Jiang Y, Wang Y, Brudno M. PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics* 2012;**28**: 2576–83.

126. Rausch T, Zichner T, Schlattl A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 2012;**28**:i333–9.
127. Handsaker RE, Korn JM, Nemesh J, et al. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* 2011;**43**: 269–76.
128. Michaelson JJ, Sebat J. forestSV: structural variant discovery through statistical learning. *Nat Methods* 2012;**9**:819–21.
129. Tan R, Wang Y, Kleinstein SE, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data *Hum Mutat* 2014;**35**:899–907.
130. Krumm N, Sudmant PH, Ko A, et al. Copy number variation detection and genotyping from exome sequence data. *Genome Res* 2012;**22**:1525–32.
131. Fromer M, Moran JL, Chambert K, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* 2012;**91**: 597–607.
132. Schouten JP, McElgunn CJ, Waaijer R, et al. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res* 2002;**30**:e57.
133. Armour JA, Palla R, Zeeuwen PL, et al. Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Res* 2007;**35**:e19.
134. Florijn RJ, Bonden LA, Vrolijk H, et al. High-resolution DNA Fiber-FISH for genomic DNA mapping and colour bar-coding of large genes. *Hum Mol Genet* 1995;**4**:831–6.