Making the difference: integrating structural variation detection tools

Ke Lin, Sandra Smit, Guusje Bonnema, Gabino Sanchez-Perez and
Dick de Ridder

Corresponding author. Dick de Ridder, Bioinformatics Group, Wageningen University, Droevendaalsesteeg 1, 6708PB Wageningen, The Netherlands.
Tel: +31 317 484074. E-mail: dick.deridder@wur.nl

Abstract

From prokaryotes to eukaryotes, phenotypic variation, adaptation and speciation has been associated with structural variation between genomes of individuals within the same species. Many computer algorithms detecting such variations (*callers*) have recently been developed, spurred by the advent of the next-generation sequencing technology. Such callers mainly exploit split-read mapping or paired-end read mapping. However, as different callers are geared towards different types of structural variation, there is still no single caller that can be considered a community standard; instead, increasingly the various callers are combined in integrated pipelines. In this article, we review a wide range of callers, discuss challenges in the integration step and present a survey of pipelines used in population genomics studies. Based on our findings, we provide general recommendations on how to set-up such pipelines. Finally, we present an outlook on future challenges in structural variation detection.

Key words: structural variation detection; integrative pipelines; population genomics; next generation sequencing

Introduction

Structural variation (SV) is an umbrella term used to denote medium-scale differences found between genomes of individuals within a certain species. Usually, small-scale differences, such as single-nucleotide variants (SNVs) and short insertions/deletions (indels, <50 bp), and large-scale differences, such as chromosome duplication, are considered as separate categories. For the purposes of this review, we define SV as genomic variation in the range of 50 bp to 1 Mb, including small (< 500 bp), medium (<5 kb) and large SVs [1]. Over the last decade, it has been realized that SV is quite abundant and can have major phenotypic consequences [2–4]. It is the underlying cause of many forms of diseases, but is also a driver of evolution, resulting in phenotypic variations of a trait, ecological adaptation, and speciation [5–10]. Based on the net loss or gain of genetic material, there are two types of SVs: balanced and imbalanced (Figure 1). A balanced variant involves no net loss or gain of genetic material. Balanced variants include inversions, where part of a chromosome is reversed, and translocations, where a segment of a chromosome is transferred to a new position, either within a chromosome (intrachromosomal) or between chromosomes (interchromosomal). Genomic imbalances, also called copy number variants (CNVs), include deletions, insertions and

Ke Lin is a PhD candidate in bioinformatics at Wageningen University, studying the genetic basis of the morphological variation found in species of *Brassica rapa*.

Sandra Smit is a postdoctoral researcher in the Bioinformatics Group at Wageningen University, specializing in the development and application of algorithms for (comparative) genomics.

Guusje Bonnema is associate professor in the Laboratory of Plant Breeding at Wageningen University. She studies the genetics of *B. rapa* species, with a particular interest in morphology and development.

Gabino Sanchez-Perez is the head of the Applied Bioinformatics cluster at Plant Research International in Wageningen. His interest is in complex genomics and the link between evolution and development.

Dick de Ridder is professor in bioinformatics and heads the Bioinformatics Group at Wageningen University. He works on learning-based algorithms, with applications in the analysis of complex genomes and systems and synthetic biology.

Submitted: 20 August 2014; **Received (in revised form):** 14 November 2014
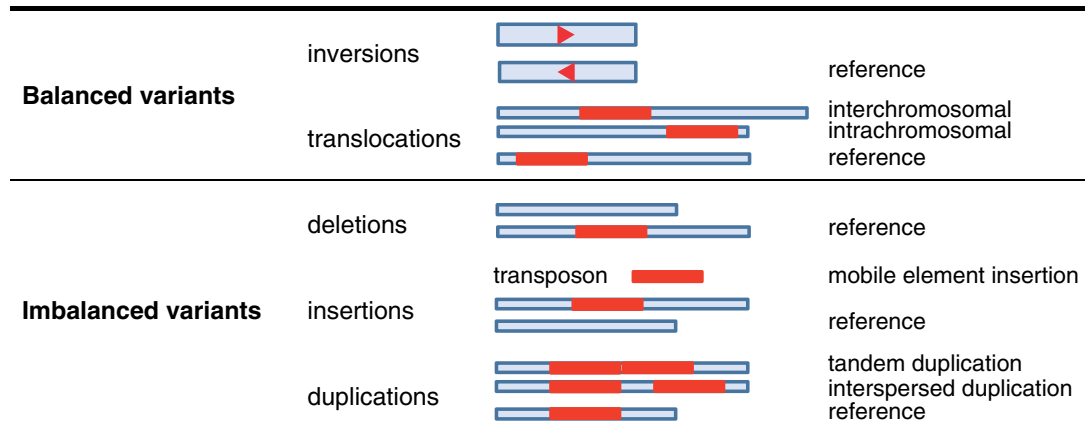
Figure 1. Types of structural variation.

duplications. A specific type of insertion is that of transposons, called mobile element insertion (MEI). Duplications can occur next to each other (in tandem) or further apart (interspersed).

With the advent of the next-generation sequencing (NGS) technology, a number of large-scale community efforts have generated nucleotide-resolution genetic variation maps in different species. Examples are the 1000 Human Genomes Project [11], the 1001 Genomes Project for *Arabidopsis thaliana* [12] and the 100 K Foodborne Pathogen Genome Project. Such variants are reported in databases, of which NCBI's dbVar and EBI's DGVa are the most well known. The majority of SVs contained in these databases are imbalanced variants found in humans and other animals (e.g. only 5842 out of 3 000 000 SVs in dbVar are balanced variants); a similar low rate is found in plant genome studies [13]. One possible biological explanation is that inversions can prevent chromosomal recombination, which is crucial to the proper pairing and segregation during meiosis [14, 15]; an alternative, technical reason is that the detection of balanced variants may be hampered by the misassembly of reference genomes and the presence of flanking inverted duplications [16].

SVs are typically identified by resequencing genomes of individuals, using short-read sequencing technology, and comparing them to a reference genome sequence. The complete SV detection process thus includes read mapping, SV calling and SV annotation (Figure 2). In this article, we focus mainly on the SV calling step, reviewing a number of promising, publically available, non–species-specific SV callers and offering guidelines on the integration of different callers to help confidently detect a wider range of variants across a number of samples.

Structural variation callers

The ideal situation for SV detection is when both the query genome and the reference genome are fully known at base resolution. For the reference genome, this is the case for a number of model organisms, but nearly all recently published draft genomes sequenced using NGS are far from complete. Likewise, the query genome is usually only sequenced at shallow depth and cannot be fully reconstructed before calling SVs. Therefore, most SV callers directly use sequencing data of a query genome. Although most tools are able to handle data obtained using different sequencing technologies [17, 18], over the last years developments have focused on the Illumina platform, as it provides high coverage at low cost and low error rates, with short (but increasing) read lengths. There are four types of information present in such data (henceforth referred to as *signals*) that can be exploited to detect SVs [19]: based on assembly (AS), read depth (RD), read pairs (RPs) and split reads (SRs) (Figure 3) [17, 20]. In general, callers exploiting information present in the paired-end reads can use all four signals, so we will focus on such callers in this review [21].

Except for callers based on AS, the sequenced reads need to be mapped to the reference genome before calling variants. After mapping, a paired-end read is called *concordant* (or properly mapped) if both ends of the read can be mapped in the expected orientation and within an expected distance range. In contrast, a *discordant* pair is aligned with an orientation or distance different from that expected based on the reference genome. RP callers exploit such discordant paired-end reads to detect SVs. For some pairs, one read fully maps to the reference while the other read can be aligned only by allowing a large gap in the mapping (*split read*) or by removing the part that cannot be mapped (*soft clipped read*). These cases are exploited by SR-based callers. RD-based callers use all mapped reads, no matter whether pairs are concordant or discordant. Finally, unmapped reads are not used by RD-, RP- and SR-based callers; only AS-based callers initially employ all sequenced reads, assembling these into contigs and mapping these to the reference genome if possible.

Specifically, the four types of callers work as follows (for more details, see [17, 20]):

1. In **AS-based callers**, short reads are first used to construct longer sequence stretches called contigs before variations are detected, a process called assembly. *De novo* AS-based (AS-D) callers first assemble the reads before mapping the resulting contigs to the reference. As contigs are longer than individual reads, they are easier to map while allowing a few mismatches, thus avoiding confusion by mapping ambiguity and mapping problems near SVs. However, to assemble reads into reasonably sized contigs, a certain minimum RD is required, which may not be possible or affordable. In contrast, local assembly (reference-based assembly, AS-R) relies on first mapping the data to a reference genome to determine the reads involved in the assembly for a certain region. This reduction in the number of reads involved reduces the complexity of the assembly process and thus improves the quality of the assembled contigs. Contig quality can be
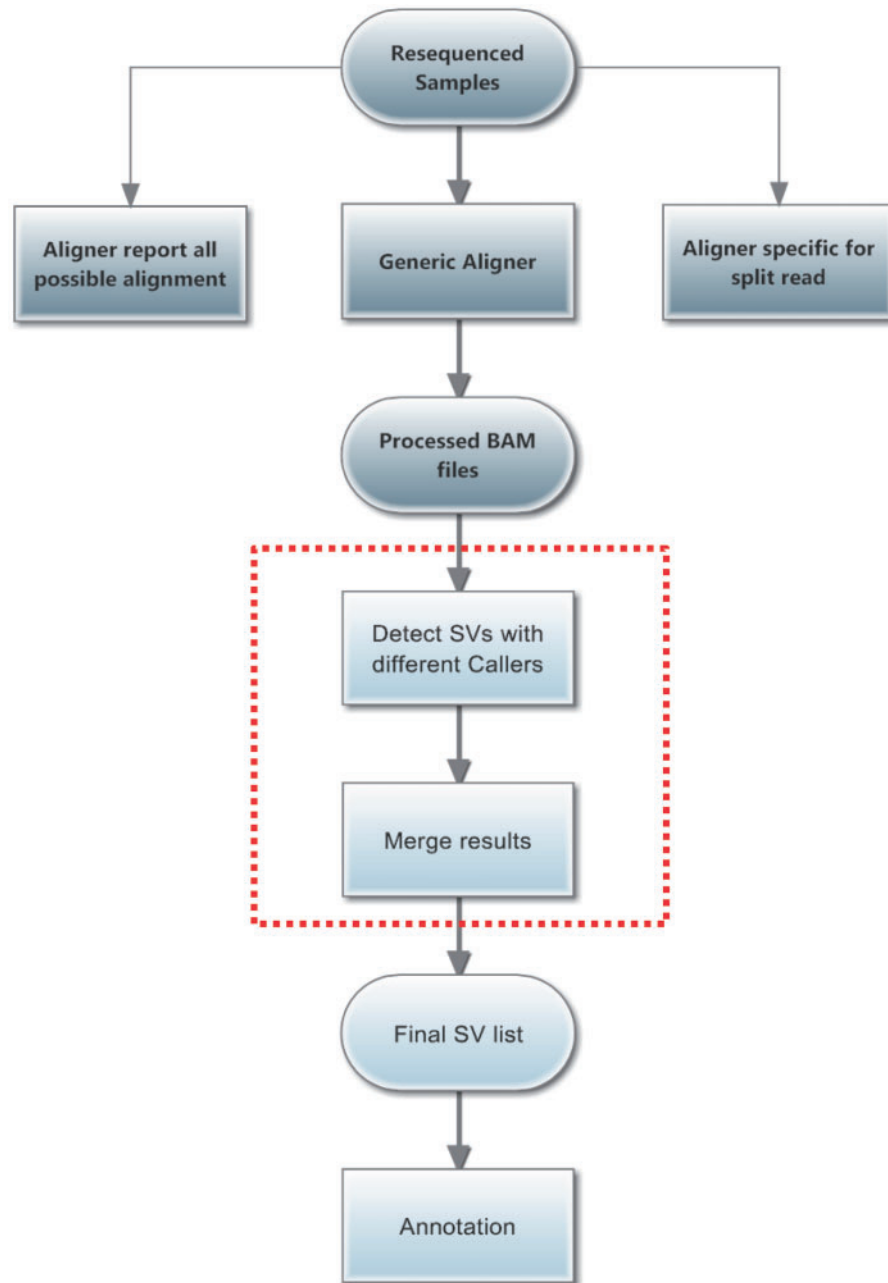
Figure 2. The complete process of SV detection. The two processes in the dotted rectangle are discussed in the article.

further improved by iterative realignment and local assembly. This is useful for detecting more complex loci, as reads can be aligned more precisely after high-confidence SVs are incorporated into the consensus sequence [22, 23].

2. **SR-based callers** can also provide base resolution SV detections, but unlike AS-based callers also work with low-coverage genome sequencing data. Callers using SR normally detect more SVs than callers using soft clipped reads (SR-SC), since reads can be split into multiple parts which can individually be aligned to multiple adjacent locations in the reference genome whereas clipped reads map only at a single location. The drawback of using SR is that short parts of reads will map to many locations on the reference genome. By itself, SR-based analysis is not sufficiently sensitive

to detect larger SVs of certain types, i.e. tandem duplications, inversions, translocations and more complex variants [24, 25].

3. **RP signals** are used by many proposed callers, owing to the relative simplicity of SV detection and the capability to detect all SV types [26–28]. Most RP callers (except Clever [28]) focus only on discordant pairs, based on distance thresholds, distributions or graphs [26]. Threshold-based callers depend on the distribution of the insert size of the sequenced library (characterized by an estimated mean and standard deviation), calling read pairs discordant if their mapping distance significantly deviates from this mean of this distribution, e.g. by more than two times the standard deviation. Small SVs may be ignored using threshold-based

Signal	Detection resolution	Detectable SV types	Used paired end reads	
Assembly (AS)	base pair	all	all reads	
Split read (SR)	base pair	all		soft clipped read one anchored split read
Read pair (RP)	rough	all		discordant pair
Read depth (RD)	very rough	deletion, duplication		all reads

Figure 3. Signals used to detect SVs using paired-end sequencing data. Read pairs are indicated by opposing arrows.

callers because smaller differences in mapping distance are likely considered to variation in library insert sizes. Distribution-based callers try to discover small and medium size SVs by comparing the global insert size distribution with the distribution of distances between locally mapped reads. Graph-based callers are developed especially for large insert size libraries, which normally contain huge deviations between the local and global insert size distributions, by exploiting both concordant and discordant reads.

4. **RD-based callers** can find duplications and deletions using all mapped reads, but only at coarse resolution. RD signals are the only signal to allow detection of copy variation directly, by calculating and comparing Read depths [29–31]. Read depths should be normalized, i.e. corrected for GC content and mappability, to increase detection resolution and lower the false-positive discovery rate. Differences in the local GC content can result in varying genome coverage in different regions and hence give a bias in the RD signal [32].

It should be clear that callers based on individual signals are geared towards detection of specific types of SVs, over a certain range of resolutions. In general, AS- and SR-based algorithms can both provide base resolution SV detection, while RP-based tools typically yield only approximate breakpoint locations and SV lengths (Table 1). AS- and SR-based methods can detect not only small SVs but also SNVs and indels, while RP signals work best for median size SVs and RD is more appropriate for large SVs [25, 33–36]. AS-based callers are able to detect insertions longer than the sequence insert, which is impossible with SR-based callers. Note that callers based on RD can detect only duplications and deletions with poor resolution [37, 38]. Consequently, it makes sense to combine different callers to allow detection of multiple SVs at a wider range of resolutions. To this end, two directions can be taken: (i) integrating different signals in a single caller or (ii) integrating different callers. These approaches will be reviewed below.

Improving detection by data integration

Callers integrating different signals

Several callers have been developed to combine different signals (Table 1). The integration of SR and RP signals resulted in improvements in SV detection with low false discovery rate in DELLY, PRISM, MATE-CLEVER, Tangram and SoftSearch [39–43]. SoftSearch uses soft-clipped reads to determine breakpoints at

Table 1. Scales of SVs detectable using different signals

Used signal(s)	Type	Breakpoint resolution	Size
RP	++	−	++
RD	−	−−	0
SR	++	+	+
AS	+	++	−−
RP + RD	−	0	+
RP + SR	++	+	+
AS + SR	+	++	+
RP + RD + SR	0	+	0
RP + RD + AS	−−	++	−

Note. The scale, from weak to strong, is indicated by −−, −, 0, +, ++. For 'Type', this corresponds to the number of different SV types that can be detected (the stronger, the more types). For 'Size', it corresponds to the range of SV sizes that can be detected (the stronger, the larger the range). Callers using the signal (combinations) are found in Table 3.

base resolution, while the other callers use SR. Pindel was initially designed to use only SR, but now also integrates RP information. DELLY shows good performance on large SVs (>10 kb) with complex rearrangements, which complements the SVs detected by MATE-CLEVER. PRISM can be sensitive to the breakpoints close to other variants in complex regions. SoftSearch is able to process exomes or custom capture experiments, whereas Tangram has been developed especially to detect mobile element insertions.

Both RP- and RD-based callers can detect many SVs, though at a high false discovery rate [44, 45]. Combining RP and RD signals can help reduce this rate and increase the resolution at which breakpoints can be determined. GASVPro integrates both signals, resulting in 50% higher specificity in detecting deletions and inversions over the original RP-based GASV [44]. CNVer is another example of a caller supplementing the RD signal with RP information. For non-human genomes, it requires several additional files to be prepared besides the reference genome, such as a contig break file, a repeat region file and a self-alignment file [46]. Unlike GASVPRo, CNVer can also compute the absolute copy counts in a region. Ingap-SV is a caller including visualization and correction, which can detect deletions, insertions and inversions [47].

The integrative callers discussed above generally first exploit one signal to detect candidate SVs and then refine detected candidates with the second signal. Such a stepwise approach can help increase specificity, but it cannot increase the number of

Table 2. Four publicly released pipelines and the default SV callers they use

Pipeline	RP	RD	SR	AS	SR + RP
SVMerge	BreakDancer	cnD		SECluster	Pindel
HugeSeq	BreakDancer	CNVnator		BreakSeq	Pindel
IntanSV	BreakDancer	CNVnator	SVseq2		Pindel, DELLY
iSVP	BreakDancer				Pindel

true positive SVs. Some efforts have been attempted to overcome this limitation. cnvHitSeq [48] combines SR, RP and RD signals in an hidden Markov model-based approach to detect copy number variation; LUMPY [49] provides a probabilistic framework allowing not only detection based on SR and RP signals, but also on variants found by any other SV detection tool. Another approach is taken by GenomeSTRiP (see section 3.3), which detects deletions at the population scale [50] and subsequently genotypes individual samples using breakpoint assembly.

Pipelines integrating multiple callers

In practice, different callers are found to yield lists of detected SVs that have limited overlap, even when using the same signal. Therefore, it makes sense to combine the output of the various implementations of callers. This avenue has recently been widely researched. Four integrated computational pipelines have been publicly released which employ a number of different SV callers and then merge their results. These pipelines fully automate the process of SV detection, from mapping the reads to the generation of a final merged list of SVs (Table 2). SVMerge [51] is a pipeline combining four 'default' SV callers: BreakDancer (RP-based), Pindel (SR-based with support for RP), cnD (RD-based) and SECluster (internal implementation of the pipeline) using AS on reads of which just the paired read mapped [25, 31, 36]. Its modular set-up allows SVs detected by other callers to be incorporated, allowing newly developed callers to be used when they become available. IntanSV [52] is another pipeline able to integrate predictions of other callers besides its default callers, which include BreakDancer and Pindel, DELLY (using RP and SR signals), CNVnator (RD-based) and SVseq2 (SR-based) [53]. HugeSeq and iSVP are pipelines that allow the use of only supplied default callers [54, 55]. Pindel and BreakDancer are used for SV detection in both HugeSeq and iSVP, with HugeSeq additionally employing BreakSeq (AS-based) and CNVnator [56]. The number of default callers is limited by the high computational cost of running each individual caller.

All four pipelines provide a filtering step for each default caller and a merge step to obtain a final list of SVs, the union of the output of the individual callers. To determine whether SVs detected by different callers can be clustered into a single SV, a certain minimum percentage of overlap between SVs is required. To filter false-positive SV discoveries as much as possible, two different approaches can be taken. One approach (taken in SVMerge) is to filter the output of individual callers by tuning their parameters. However, optimal settings are study dependent and are hard to ascertain without a ground truth set of known, experimentally validated SVs. An alternative approach (used by IntanSV and HugeSeq) is to consider SVs supported by only two or more callers, as these are considered to have higher confidence than SVs detected by just one caller. In both cases, the strength of individual callers is not taken into account: SVs detected by individual callers are deemed equally

correct. If sufficient data are available, a weighted combination could be used, with weights reflecting the relative strengths of the callers.

iSVP only detects and reports deletions, while the other three pipelines can detect all types of SVs and include a number of useful post-processing steps. HugeSeq and IntanSV both annotate the predicted effects of the SVs; HugeSeq also incorporates SNVs. SVMerge uses local assembly to validate the detected SVs and refine the corresponding breakpoints.

Integrating samples in population genomics

Besides the integration of signals and callers, another approach to increase both specificity and sensitivity of SV detection is to integrate multiple samples. SV detection in population-scale data can ascertain more low-frequency variation, which often is functional and hence increases the power for linking variation to phenotype. This leads to a consistently low false discovery rate (1.5%–4.2% reported) with high overall sensitivity (contributed 80% of phase 1 deletions in 1000 Human Genome Project) especially on genomes sequenced at low coverage (<10 fold) [50]. The GenomeSTRiP pipeline and cortex_var algorithm are useful, because they not only combine multiple signals (cortex_var uses the paired end information before the assembly) but also take population information into account.

However, in a large number of published population genomics studies, in-house scripts were used to detect SVs. Mills *et al.* studied variation in 185 human genomes using 19 different callers, including not only callers using single signals but also callers combining RP and RD signals [19]. This is currently the most complex SV detection pipeline reported in terms of the large number of samples, range of sequencing technology used and the number of callers applied to the data. Interestingly, the developers of each caller analysed the data sets themselves and optimized these towards discovering SVs in the 185 genomes, making it a template approach for SV detection at the population scale.

To choose callers, it is important to be aware of each caller's precision in SV breakpoint detection (i.e. start position, end position and content). Precisely determining breakpoints (possible when using SR or AS signals, see section 2) is a prerequisite to allow results from different callers or samples to be merged. It is also essential to predict the origin and functional impact of an SV. Incorrectly predicted SVs can make it hard to uncover true genetic variation by introducing noise, in the form of many false candidate genes. These false discoveries can be largely reduced in population genomics studies on variants with appreciable allele frequency (>1%) shared with other members of the population [57].

In population genomics studies, the merging of results from individual callers is far more complex than can be handled by the four aforementioned published integrative pipelines (section 3.2), because of the large number of samples and callers involved. The key to allow merging of SVs found by a single caller in different samples is the precise determination of unique breakpoints. A solution currently used is to locally assemble reads around breakpoints. While many tools can do this, only few (among which TIGRA-SV) can unify and refine the SV breakpoints of different individuals or from various callers. Some studies circumvent this problem by using a single caller supporting the analysis of multiple samples simultaneously, hence bypassing the merge step [50, 58–61]. Limited by this single caller, such studies usually focus on a certain type or size of SVs. In a study on 16 cultivated and wild soybean genotypes,

deletions were detected using GenomeSTRiP, which is optimal for deletions [62]. However, to find various SV types and sizes, multiple callers need to be employed, requiring the investment of additional computational power at decreasing returns: roughly half of the final list of SVs found in the Mills study are detected by only 5 of the 19 SV callers. Similar large numbers of callers were used in many other population studies [6, 11, 63–65], choosing the best caller for certain type of SV and then merging in results of other callers suited for that type. This approach of choosing 'best' callers is also useful to set-up the merge criteria when conflicts on SV type or breakpoint found by different callers occur. Zichner *et al.* used GenomeSTRiP as the best tool for detection of deletions and included 283 deletions solely found by GenomeSTRiP, while filtering out SVs predicted by a single other caller [64].

Suggested integrated pipelines

Here, we outline what elements an integrative SV detection pipeline should contain. A large number of callers have been developed in recent years. Even when selecting only those that are not species specific and capable of working with paired-end sequencing, the number is too large to use in any single study, given the required computational power to run the callers and the merging step. Therefore, callers need to be prioritized. Below, we recommend a number of callers for detecting specific SV types, focusing on those useful for population studies, and discuss methods of merging the sets of detected SVs. We surveyed 51 non–species-specific callers that work with paired-end reads, mainly by consulting OMICtools and SEQwiki (Table 3) [66, 67]. We classified each caller by the signal(s) it uses, the SV types that it can detect, its support for population studies, information on benchmarking in the original publication (indicating whether it is well evaluated), the date of the last update (indicating whether it is maintained) and the number of citations (indicating whether it is widely used). To learn about their application in practice, we also surveyed population genomics studies of a large number of species, reporting the number of the samples, the average genome coverage, SV types detected, SV callers used and the merge strategy taken (Table 4).

Filtering and merging SVs detected in a range of samples is a complex process, which can generate many false-positive results, mainly due to the presence of SNVs [116]. It is therefore recommended to use SV callers that can also perform genotyping; this is, in fact, essential when studying heterozygous genomes such as in studies on animals and wild plants. Platypus [68] seems a good choice for a caller, as its sensitivity in detecting SNVs is similar to that of the widely used GATK Unified Genotyper and Haplotype Caller utilities, at even lower false discovery rates. As Platypus depends on AS-R, it can easily detect deletions up to a few kilobases in size, but its detection of insertions is limited to a few hundred base pairs. Reference-based assembly is largely dependent on the mappability of the reads in the resequencing data; unmapped reads are generally discarded. Hence, it is not the best choice for studies with large genomic divergence within the population, such as maize, tomato etc. In contrast to Platypus, cortex_var employs AS-D, exploiting all reads, and is theoretically capable of detecting SNVs as well as deletions and insertions of arbitrary size. A combination of Platypus and cortex_var was recently used in a chimpanzee study [101].

DELLY and Pindel, also supporting multiple samples and genotyping, are widely used in population studies (Table 4) because of their ease-of-use and relatively low computational requirements. These two callers complement each other, with DELLY being more sensitive in detecting large SVs and Pindel more sensitive to small SVs (Supplementary Table S1). Genome-STRiP genotypes samples by combining information on RD, RP and breakpoint-spanning reads, which makes it a good choice for detecting medium and large size SVs (though, thus far, only deletions). Similar to Genome-STRiP, cnvHitSeq also uses multiple signals to accurately detect deletions and duplications. The lack of a single, streamlined script has limited the practical application of cnvHitSeq. Even though it is not based on AS but on SR/RP signals, MATE-CLEVER is able to accurately call and genotype indels in the range 30–120 bp and is therefore a good choice to include in a pipeline in addition to AS-based callers. Although BreakDancer does not support genotyping, it is still used in many population studies as it is one of the earliest developed callers, easy-to-use with low computational requirements. Finally, if a sufficient number of experimentally validated SVs is available, forestSV is useful as it is reported to yield low false-positive and false-negative rates [77].

Some callers should be considered for specific studies. For example, if repetitive regions are of interest, specific algorithms may be preferred that allow mapping reads to multiple locations on the reference genome. A range of callers based on different signals has been developed to detect repeats, including VariationHunter (RP-based), mrCaNaVaR (RD-based) and NovelSeq (AS-based) [38, 79, 78]. These are all based on a specific alignment tool called mrFAST, which can generate all possible mapping positions of the reads [117]. As another example, the detection of mobile element insertions requires specific methods such as Tangram, SPANNER and RetroSeq [73, 86]. Tangram integrates SR and RP signals, while SPANNER and RetroSeq use only RP signals.

After the various callers have been applied, the detected SVs should be merged. Existing pipelines such as SVMerge, iSVP and HugeSeq use BEDtools [118] to merge SVs based on overlap. Although this strategy is easy to implement, it is prone to errors caused by accidental overlaps, in particular for large SVs. To reduce this problem, some studies use a stricter 50% reciprocal overlap requirement. A better merge strategy employs confidence intervals around the detected breakpoints, rather than overlap between SVs. Different callers will yield different confidence intervals; for instance, RD-based callers will yield larger confidence intervals than those using other signals, as RD only allows for limited breakpoint resolution. To improve accuracy, a 50% reciprocal overlap strategy could precede a confidence interval strategy [19].

In practice, a good pipeline for studies with multiple samples should start with a pilot study using a few representative callers to detect each desire type of SVs. A number of such callers are suggested in Figure 4, based on the arguments given above. We suggest to integrate the output of more than one caller for each type whenever possible, to reduce the false discovery rate. The SVs detected can be merged using the confidence interval strategy, after filtering SVs having too little overlap or SVs supported by only one or a few callers. Based on experience, one may want to put more trust in the SVs detected by certain callers. While there are no large-scale benchmarks to support this, such a choice can be based on the frequency of use (Table 4) or on comparative evaluations published thus far (Supplementary Table S1). Additional callers can then be added if the biological questions are not yet fully answered by the pilot study, or if the experimental validation demonstrates poor performance.

Table 3. An overview of generic SV callers[a]

Tool	Signals used	SV types detected	Multiple samples	Genotyping	Benchmarking data type	Used samples	Validated by	SV type benchmarking	SV size benchmarking	Coverage benchmarking	Last update	Cited by	Reference
Platypus	AS-R	Del, Ins	Yes	Yes	Real biological data	CEU trio	Fosmid	No[b]	No	No	15-9-2014	1	[68]
Delly	RP, SR	Del, TD, Inv, ITX, CTX	Yes	Yes	Real biological data	1000GP	PCR	Yes	No	Yes	3-9-2014	68	[39]
CNVseq	RD	Del, Dup	No	No	Real biological data	Human[c]	Standards	No	No	No	12-8-2014	194	[30]
Gustaf	SR, RP	Del, Ins, Inv, Dup, ITX, CTX	No	No	Simulation	NA	NA	Yes	No	Yes	1-8-2014	0	[1]
Control-FREEC	RD	Del, Dup	No	No	Real biological data	Human	SNP-array	No	No	No	11-6-2014	41	[29]
softsearch	SR, RP	Del, Ins, Inv, CTX	Yes	No	Real biological data	2 human	Standards	No	No	Yes	9-4-2014	4	[41]
Genome STRiP	RP, RD, AS-R	Del	Yes	Yes	Real biological data	1000GP	Standards	No	No	Yes[d]	13-2-2014	110	[50]
IMR-DENOM	AS-R	Del, Ins	Yes	No	Real biological data	Col/Ler	PCR	No	No	No	12-2-2014	159*	[33]
tangram	RP and SR	MEI	Yes	Yes	Real biological data	CEU trio	Standards	Yes	No	No	9-2-2014	0	[69]
TakeABreak	de Bruijn graph	Inv	No	No	Simulation	NA	NA	No	No	No	1-2-2014	1	[70]
Cloudbreak	RP	Del, Ins	No	No	Real biological data	NA18507	Standards	Yes	Yes	No	1-2-2014	2	[71]
Meerkat	SR-SC, RP, AS-R	Del, Ins, Inv, TD, ITX	Yes	No	Real biological data	Yoruba&CEU	Standards	Yes	No	No	1-1-2014	24	[35]
PRISM	SR, RP	Del, Ins, Inv, TD, CTX	No	No	Real biological data	Yoruba&CEU	Standards	Yes	Yes	No	29-12-2013	22	[40]
cortex_var	AS-D	Del, Ins, Inv	Yes	Yes	Real biological data	NA12878	Standards	No	No	No	13-11-2013	97	[34]
lumpy	SR, RP, Others	Del, Dup, Inv	No	No	Real biological data	NA12878	Standards	Yes	No	Yes	18-10-2013	2	[49]
GASVPro	RP, RD	Del, Inv	Yes	Yes	Real biological data	Yoruba&CEU	Standards	Yes	No	No	1-10-2013	30	[44]
SVMiner	RP	Del, Inv	Yes	Yes	Real biological data	NA18507	Standards	No	No	No	19-9-2013	5	[72]
MrCaNaVaR	RD	Del, Dup	Yes	No	Real biological data	NA18507	Microarray	No	No	No	4-9-2013	328	[38]
RetroSeq	RP	MEI	Yes	Yes	Real biological data	CEU trio	Standards	Yes	No	No	21-7-2013	11	[73]
MATE-CLEVER	SR, RP	Del, Ins	Yes	Yes	Simulation	NA	NA	No	Yes	Yes	17-10-2013	16	[43]
CLEVER	RP	Del, Ins	Yes	Yes	Real biological data	NA18507	Standards	Yes	Yes	No	2-7-2013	21	[28]
cn.MOPS	RD	Del, Dup	Yes	Yes	SVs inserted in real data	NA20755	NA	Yes	No	No	14-5-2013	42	[74]
inGAP-sv	RP, RD	Del, Ins, Inv, CTX	Yes	No	Real biological data	NA12878	Standards	Yes	No	No	29-4-2013	17	[47]
cnvHitSeq	SR, RP, RD	Del, Dup	Yes	Yes	Real biological data	CEU trio	Standards	Yes	No	No	26-4-2013	4	[48]
Pindel	SR	Del, Ins, Inv, TD	Yes	Yes	Real biological data	NA18507	Standards	No	No	No	25-4-2013	337	[25]
SVseq2	SR, SR-SC	Del, Ins	Yes	No	Real biological data	YRI	Standards	No	No	No	22-4-2013	13	[53]
ERDS	RD, RP, SR-SC	Del, Dup	No	No	Real biological data	NA12878	Standards	No	Yes	No	18-4-2013	14	[75]
Breakway	RP	Del, Ins, ITX	No	No	Real biological data	Human	PCR	No	No	No	11-4-2013	120*	[76]
BreakDancer	RP	Del, Ins, Inv, ITX, CTX	Yes	No	Real biological data	CEU&YRI	Standards	No	No	No	24-3-2013	380	[36]
forestSV	NA	Del, Dup	Yes	No	Real biological data	1 KG	Standards	Yes	No	No	15-3-2013	7	[77]
SVDetect	RP or RD	Del, Ins, Inv, Dup, ITX, CTX	Yes	No	Real biological data	Yeast	PCR	No	No	No	22-1-2013	65	[27]
NovelSeq	AS-D	Ins	No	No	Real biological data	NA18507	Standards	No	No	No	12-11-2012	64	[78]
CNVnator	RD	Del, Dup	Yes	Yes	Real biological data	Yoruba	Standards	No	No	No	27-9-2012	127	[37]
VariationHunter	RP	Del, Ins, Inv, MEI	Yes	No	Real biological data	NA18507	Standards	No	No	No	15-7-2012	184	[79]
Readdepth	RD	Del, Dup	No	No	Simulation	NA	NA	Yes	No	No	31-5-2012	42	[80]
SOAPSV	AS-D	Ins, Del, Inv	No	No	Real biological data	NA18507	PCR	Yes	No	No	17-2-2012	60	[81]
cnD	RD	Del, Dup	Yes	No	Real biological data	Mouse	array CGH	No	No	No	16-2-2012	30	[31]
ClipCrop	SR-SC	Del, Ins, INV, DUP, CTX	Yes	No	Simulation	NA	NA	Yes	No	No	27-1-2012	8	[45]
EWT	RD	Del, Dup	No	No	Real biological data	5 human	Standards	No	No	No	13-1-2012	242	[82]
MUMmer	AS-D		No	No	Real biological data	NA	NA	No	No	No	17-12-2011	1,179	[83]

(continued)

Table 3. (continued)

Tool	Signals used	SV types detected	Multiple samples	Genotyping	Benchmarking data type	Used samples	Validated by	SV type benchmarking	SV size benchmarking	Coverage benchmarking	Last update	Cited by	Reference
CREST	SR-SC, AS	TD, Inv, Ins, Del, ITX, CTX	Yes	No	SVs inserted in real data	NA12878	Standards	Yes	No	No	8-11-2011	120	[84]
AGE	AS-D	Del, Ins, Inv, ITX, CTX	No	No	Real biological data	Human	NA	No	No	No	12-9-2011	31	[85]
SPANNER	RP	Del, Inv, ITX, TD, MEI	No	No	Real biological data	1000GP	Standards	No	No	Yes^d	21-8-2011	NA	[86]
CNVer	RP, RD	Del, Dup	No	No	Real biological data	NA18507	Standards	Yes	No	No	11-7-2011	73	[46]
splazerS	SR	Del, Ins	No	No	Real biological data	NA12878	Standards	Yes	Yes	No	28-4-2011	21	[24]
SLOPE	SR	Del, Ins, ITX, CTX	No	No	Simulation	NA	NA	Yes	Yes	No	3-11-2010	25	[87]
JointSLM	RD	Del, Dup	No	Yes	Real biological data	CEU trio	Standards	No	No	No	23-8-2010	23	[88]
hydra-sv	RP	NA	No	No	Real biological data	Mouse	SR mapping	No	No	No	20-8-2010	119*	[89]
CnvHMM	RD	Del, Dup	No	No	Real biological data	Human	NA	No	No	No	4-6-2009	532*	[90]
PEMer	RP	Del, Ins, Inv	Yes	No	Real biological data	NA18505	Standards	No	No	No	25-2-2009	133	[26]
SVM2	RP, RD	Del, Ins	NA	NA	Real biological data	Kidd	Standards	No	No	No	NA	8	[91]

Notes. [a]An overview of generic SV callers, including the signal(s) used (AS-R, AS-D, RD, RP, SR and SR-SC), the SV types detected (Del: deletion, Ins: insertion, Inv: inversion, Dup: duplication; TD: tandem duplication; ITX, CTX: intra- resp. intrachromosomal translocation), support for population studies and genotyping (i.e. reporting presence or absence of variation for a number of genomes), the date of the last update and the number of citations (* means the tool was reported as part of an experimental paper), the type of data used in benchmarking in the original publication, samples used (CEU trio: child NA12878, parents NA12891 and NA12892; 1000 GP: 921 illumina sequenced samples from the 1000 GP Phase 1 [57]; two human: NA12878 at high (40×) and NA18507 at low coverage (4×); Col/Ler: accession Col and Ler of *Arabidopsis thaliana*; Youban&CEU: HapMap individual NA18507 and NA12878; YRI: NA19311, NA19312, NA19313, NA19316, NA19317; 1 KG: NA12878, NA12891, NA12892, NA19240, NA19239; five human: NA12878, NA12891, NA12892, NA18507, and YH; Kidd: data used in the Kidd study [92]), validation used for benchmarking (Gold Standards: Cappillary Read Data, array-CGH and Fosmid Sequencing [93] used by the 1000 Genome Structural Variant discovery study[19]), different categories benchmarkied (SV type, SV size, genome coverage), last update, the number of citations (found using Google Scholar on 3 October 2014) and the literature reference of each caller.
[b]The publication of Platypus only reports the benchmarking of SNVs and short indels.
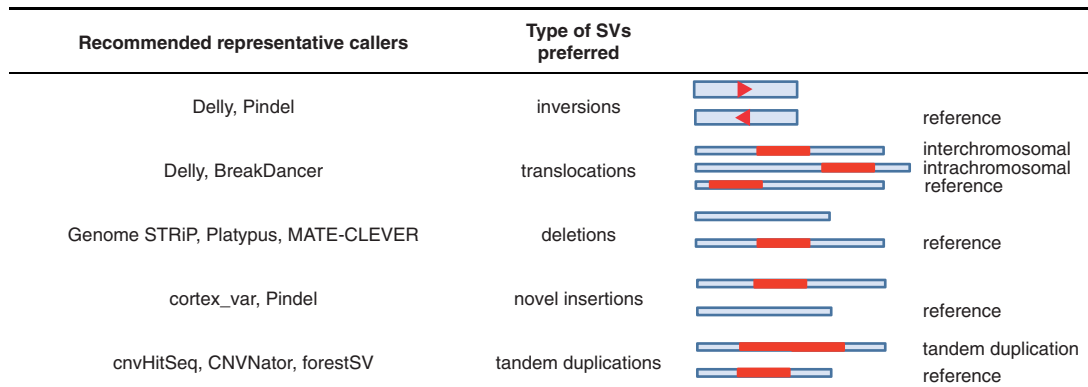[c]The genomes of two individuals , Dr Craig J. Venter [94] and Dr J. Watson [95].
[d]The benchmarking was actually published in the Mills study [19].

Table 4. A list of published population genomics studies on different species and the SV detection procedure involved

Organism	Number of samples	Coverage	Types of SVs detected	SV callers	Merge strategy	Key reference
Mammals						
Homo sapiens	185	3.6×	Del, Dup, Ins	Various[a]	Confidence Interval	[19, 96]
Rattus rattus	27	20×	Del, Ins, Inv, ITX, CTX	BreakDancer	NA	[97]
Mus musculus	18	27.6×	Del, Ins, Inv	BreakDancer, Pindel, cnD	Overlap	[98]
Bos grunniens	6	5.5×	Del, Ins, Inv, ITX, CTX	BreakDancer	NA	[99]
Sus scrofa	48	10×	Deletion, Duplication	CNVnator, CNVseq	Overlap	[100]
Pan troglodytes verus	9	35×	Del, Ins, Inv	Platypus, cortex_var	Overlap	[101]
Great ape	80	23×	MEI	variationHunter	NA	
Canis lupus	60	0.5×	Del, Ins, ITX, CTX	BreakDancer [107] / CNVnator [108], CNVseq	Overlap [108]	[102, 103]
Rhesus macaque, Pongo abelii, Macaca mulatta	15	17.5×	Del, Dup	DELLY, CNVnator, GenomeSTRiP	Confidence Interval	[65]
Birds						
Corvus corone	60	12.2×	Inv	DELLY	NA	[104]
Fish						
Gasterosteus aculeatus	6	10×	Dup, Del, Inv	CNVnator, BreakDancer, Pindel	Overlap	[63]
Insects						
Drosophila melanogaster	39, 205	27×	Del, TD	DELLY, Pindel, CNV-nator, Genome STRiP	Confidence Interval	[64, 105]
Heliconius erato	45	25×	Del, Ins, Inv, ITX, CTX	BreakDancer	NA	[106]
Plants						
Glycine max, Glycine soja	10 max, 6 soja	17×	Deletion	Genome-STRiP	NA	[62]
Prunus persica	84	3.5×	Ins, Del, TD, Inv, CTX	SOAPSV, DELLY	Overlap	[10]
Brachypodium distachyon	7	47×	Ins, Del, Inv, ITX, CTX	IMR-denom, Pindel, BreakDancer	Overlap	[108]
Sesamum indicum	29	13×	Ins, Del, Inv, CTX, ITX	BreakDancer	NA	[109]
Arabidopsis thaliana	180	39×	Ins, Del, Inv, CTX, ITX	BreakDancer, Pindel, SOAPSV	Confidence Interval	[6]
Chlamydomonas reinhardtii	4	39×	Ins, Del, Inv	BreakDancer, Pindel	Overlap	[110]
Zea mays	278	27×	Del, Dup	EWT	NA	[111]
Solanum lycopersicum	8	11.2×	Del, Dup	cn.Mops	NA	[112]
Microorganisms						
Saccharomyces cerevisiae	4	150×	Del, TD, Inv	DELLY, Pindel	Overlap	[113]
Staphylococcus aureus	131	94.3×	Del, Ins, Inv	Cortex	NA	[114]
Cryptococcus neoformans	3	30×	Del, Ins, Inv	BreakDancer, CREST	Overlap	[115]

Notes. Besides the callers used and the SV types detected, the table lists the merge strategy used to combine the results from different callers (O: merge SV events if they overlap; CI: merge SV events if the confidence intervals around breakpoints of SV events overlap), the number of samples studied and the average genome coverage of sequencing are also reported. When multiple studies have been performed on the same species, references indicate the various choices in these studies. Note that there are too many population studies in human to report here; we chose one of the first and a most recent study using the most callers as representatives, with the footnote listing all callers used in all studies.
[a]Callers involved in this study are: EWT, CNVnator, Spanner, VariationHunter, BreakDancer, PEMer, Pindel, Cortex_var, NovelSeq, Genome-STRiP.

Recommended representative callers	Type of SVs preferred	
Delly, Pindel	inversions	reference
Delly, BreakDancer	translocations	interchromosomal intrachromosomal reference
Genome STRiP, Platypus, MATE-CLEVER	deletions	reference
cortex_var, Pindel	novel insertions	reference
cnvHitSeq, CNVNator, forestSV	tandem duplications	tandem duplication reference

Figure 4. Suggested callers for each type of SV.

Future directions

To allow reliable detection of the complete spectrum of SVs over entire genomes, improvements are needed in the methodology, on the computational side and in the data generation process. Methodologically, the lack of reliable 'gold standard' data hampers the evaluation and benchmarking of the plethora of available tools, which makes it hard to make clear-cut recommendations for the use of specific callers in specific circumstances. In this light, the efforts of the TCGA Dream Challenge on somatic variant calling are encouraging; such challenges should be extended to other model species if possible, to avoid biasing results too much towards human studies. Computationally, one of the main issues for most currently available SV detection algorithms is their high false discovery rate. The integration of several signals for SV detection as discussed in this review is promising, but risks introducing additional false negatives. This problem can be tackled by taking a machine learning approach to the merge step, i.e. combining information obtained from individual callers into a final prediction and optimizing the parameters of this process ('training') by learning from example data. A number of tools using such a supervised learning approach have been developed, but these are currently limited to detecting deletions and duplications [77, 91, 119]. The data used to train the combination algorithm are the most important factor and hard to come by, requiring experimentally validated SVs. On the data generation side, large-scale experimental detection of SVs over all sizes has been achieved by an optical mapping involving nanoconfinement of individual DNA strands [120]. However, the combined computational analysis of sequencing data and the resulting restriction map in many ways are as complex and challenging as the assembly of shotgun sequence reads. These problems can be reduced if the DNA is sequenced with longer read lengths, such as those provided currently by PacBio technology [121].

Improvements in sequencing and mapping technology will ultimately allow us to reconstruct complete genomes, negating the need for complex SV detection tools. Until that time, integrative pipelines intelligently combining the output of a large number of appropriate callers are the best option.

Key Points

- Algorithms for structural variation detection on next-generation sequencing data (callers) can use different signals, suitable to find specific variant types or sizes.
- To reliably detect variants over a wide range of types and sizes, callers can combine various signals, or callers optimal for specific signals can be combined in integrative pipelines.
- When analysing multiple samples, it makes sense to use callers or pipelines specifically developed to exploit these, the challenge being to integrate the detected variants over all genomes.
- Ideally, a structural variation detection pipeline integrates the output of a (small) number of callers for specific SV types, adding other callers when allowed by computational resources or required by the results obtained.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

References

1. Trappe K, Emde AK, Ehrlich HC, *et al*. Gustaf: detecting and correctly classifying SVs in the NGS twilight zone. *Bioinformatics* 2014.
2. Maydan JS, Lorch A, Edgley ML, *et al*. Copy number variation in the genomes of twelve natural isolates of Caenorhabditis elegans. *BMC Genomics* 2010;**11**:62.
3. Gazave E, Darre F, Morcillo-Suarez C, *et al*. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Res* 2011;**21**:1626–39.
4. Lu P, Han X, Qi J, *et al*. Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing *Landsberg erecta* and all four products of a single meiosis. *Genome Res* 2012;**22**:508–18.
5. Kirkpatrick M. How and why chromosome inversions evolve. *PLoS Biol* 2010;**8**:e1000501.
6. Long Q, Rabanal FA, Meng D, *et al*. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat Genet* 2013;**45**:884–90.
7. Magwire MM, Bayer F, Webster CL, *et al*. Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a duplication. *PLoS Genet* 2011;**7**:e1002337.
8. Pearce S, Saville R, Vaughan SP, *et al*. Molecular characterization of Rht-1 dwarfing genes in hexaploid wheat. *Plant Physiol* 2011;**157**:1820–31.
9. Vlad D, Rappaport F, Simon M, *et al*. Gene transposition causing natural variation for growth in Arabidopsis thaliana. *PLoS Genet* 2010;**6**:e1000945.
10. Weischenfeldt J, Symmons O, Spitz F, *et al*. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* 2013;**14**:125–38.
11. Abecasis GR, Auton A, Brooks LD, *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**:56–5.
12. Cao J, Schneeberger K, Ossowski S, *et al*. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet* 2011;**43**:956–63.
13. Lappalainen I, Lopez J, Skipper L, *et al*. DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res* 2013;**41**:D936–41.
14. Stevison LS, Hoehn KB, Noor MA. Effects of inversions on within- and between-species recombination and divergence. *Genome Biol Evol* 2011;**3**:830–41.
15. Santos M. Recombination load in a chromosomal inversion polymorphism of Drosophila subobscura. *Genetics* 2009;**181**:803–9.
16. Feuk L. Inversion variants in the human genome: role in disease and genome architecture. *Genome Med* 2010;**2**:11.
17. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* 2009;**6**:S13–20.

18. Koboldt DC, Larson DE, Chen K, *et al*. Massively parallel sequencing approaches for characterization of structural variation. *Methods Mol Biol* 2012;**838**:369–84.

19. Mills RE, Walter K, Stewart C, *et al*. Mapping copy number variation by population-scale genome sequencing. *Nature* 2011;**470**:59–65.

20. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011;**12**:363–76.

21. Korbel JO, Urban AE, Affourtit JP, *et al*. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 2007;**318**:420–6.

22. Chen K, Chen L, Fan X, *et al*. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome Res* 2014;**24**:310–17.

23. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol* 2010;**11**:R41.

24. Emde AK, Schulz MH, Weese D, *et al*. Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. *Bioinformatics* 2012;**28**:619–27.

25. 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**:2865–2871.

26. Korbel JO, Abyzov A, Mu XJ, *et al*. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 2009;**10**:R23.

27. Zeitouni B, Boeva V, Janoueix-Lerosey I, *et al*. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics* 2010;**26**:1895–96.

28. Marschall T, Costa IG, Canzar S, *et al*. CLEVER: clique-enumerating variant finder. *Bioinformatics* 2012;**28**:2875–82.

29. Boeva V, Popova T, Bleakley K, *et al*. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* 2012;**28**:423–5.

30. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 2009;**10**:80.

31. Simpson JT, McIntyre RE, Adams DJ, *et al*. Copy number variant detection in inbred strains from short read sequence data. *Bioinformatics* 2010;**26**:565–7.

32. Bentley DR, Balasubramanian S, Swerdlow HP, *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;**456**:53–9.

33. Gan X, Stegle O, Behr J, *et al*. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* 2011;**477**:419–23.

34. Iqbal Z, Caccamo M, Turner I, *et al*. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* 2012;**44**:226–32.

35. Yang L, Luquette LJ, Gehlenborg N, *et al*. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* 2013;**153**:919–29.

36. 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.

37. Abyzov A, Urban AE, Snyder M, *et al*. CNVnator: an approach to discover, genotype: characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 2011;**21**:974–84.

38. Alkan C, Kidd JM, Marques-Bonet T, *et al*. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* 2009;**41**:1061–1067.

39. Rausch T, Zichner T, Schlattl A, *et al*. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 2012;**28**:i333–39.

40. 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.

41. Hart SN, Sarangi V, Moore R, *et al*. SoftSearch: integration of multiple sequence features to identify breakpoints of structural variations. *PLoS One* 2013;**8**:e83356.

42. Jiantao Wu W-PL. *Fast Structural Variation Detection Toolbox*. https://github.com/jiantao/Tangram (4 January 2014 date last accessed).

43. Marschall T, Hajirasouliha I, Schonhuth A. MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics* 2013;**29**:3143–50.

44. 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.

45. Suzuki S, Yasuda T, Shiraishi Y, *et al*. ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics* 2011;**12**(Suppl 14):S7.

46. Medvedev P, Fiume M, Dzamba M, *et al*. Detecting copy number variation with mated short reads. *Genome Res* 2010;**20**:1613–22.

47. 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.

48. Bellos E, Johnson MR, Coin LJM. cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. *Genome Biol* 2012;**13**:R120.

49. Layer RM, Chiang C, Quinlan AR, *et al*. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* 2014;**15**:R84.

50. 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.

51. Wong K, Keane TM, Stalker J, *et al*. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol* 2010;**11**:R128.

52. Yao W. intansv: Integrative analysis of structural variations, R package version 1.2.0 2014.

53. Zhang J, Wang J, Wu Y. An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. *BMC Bioinformatics* 2012;**13**(Suppl 6):S6.

54. Lam HY, Pan C, Clark MJ, *et al*. Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat Biotechnol* 2012;**30**:226–9.

55. Mimori T, Nariai N, Kojima K, *et al*. iSVP: an integrated structural variant calling pipeline from high-throughput sequencing data. *BMC Syst Biol* 2013;**7**(Suppl 6):S8.

56. Lam HY, Mu XJ, Stutz AM, *et al*. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* 2010;**28**:47–55.

57. Genomes Project C, Abecasis GR, Altshuler D, *et al*. A map of human genome variation from population-scale sequencing. *Nature* 2010;**467**:1061–1073.

58. Montgomery SB, Goode DL, Kvikstad E, *et al*. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* 2013;**23**:749–61.

59. Stewart C, Kural D, Stromberg MP, *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* 2011;**7**:e1002236.

60. Flagel LE, Willis JH, Vision TJ. The standing pool of genomic structural variation in a natural population of *Mimulus guttatus*. *Genome Biol Evol* 2014;**6**:53–64.

61. Ventura M, Catacchio CR, Alkan C, *et al.* Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res* 2011;**21**:1640–9.

62. Chung WH, Jeong N, Kim J, *et al.* Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. *DNA Res* 2014;**21**:153–67.

63. Feulner PG, Chain FJ, Panchal M, *et al.* Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. *Mol Ecol* 2013;**22**:635–49.

64. Zichner T, Garfield DA, Rausch T, *et al.* Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing. *Genome Res* 2013;**23**:568–79.

65. Gokcumen O, Tischler V, Tica J, *et al.* Primate genome architecture influences structural variation mechanisms and functional consequences. *Proc Natl Acad Sci USA* 2013;**110**:15764–9.

66. Arnaud Desfeux BG, Henry V, Pépin A-S. OMICtools - a manually curated metadatabase for multi-omic data analysis. http://omictools.com/sv-calling/ (4 February 2014).

67. Li JW, Robison K, Martin M, *et al.* The SEQanswers wiki: a wiki database of tools for high-throughput sequencing analysis. *Nucleic Acids Res* 2012;**40**:D1313–17.

68. Rimmer A, Phan H, Mathieson I, *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet* 2014;**46**:912–18.

69. Wu J, Lee WP, Ward A, *et al.* Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC Genomics* 2014;**15**:795.

70. Lemaitre C, Ciortuz L, Peterlongo P. Mapping-free and assembly-free discovery of inversion breakpoints from raw NGS reads. In: *Proceedings International Conference on Algorithms for Computational Biology (AlCoB 2014). Springer Lect Notes Comput Sci* 2014;**8542**: 119–30.

71. Whelan CW, Tyner J, L'Abbate A, *et al.* Cloudbreak: accurate and scalable genomic structural variation detection in the cloud with MapReduce. *arXiv*:1307.2331 2013.

72. Hayes M, Pyon YS, Li J. A model-based clustering method for genomic structural variant prediction and genotyping using paired-end sequencing data. *PLoS One* 2012;**7**:e52881.

73. Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* 2013;**29**:389–90.

74. Klambauer G, Schwarzbauer K, Mayr A, *et al.* cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res* 2012;**40**:e69.

75. Zhu M, Need AC, Han Y, *et al.* Using ERDS to infer copy-number variants in high-coverage genomes. *Am J Hum Genet* 2012;**91**:408–21.

76. Clark MJ, Homer N, O'Connor BD, *et al.* U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet* 2010;**6**:e1000832.

77. Michaelson JJ, Sebat J. forestSV: structural variant discovery through statistical learning. *Nat Methods* 2012;**9**:819–21.

78. 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.

79. Hormozdiari F, Hajirasouliha I, Dao P, *et al.* Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 2010;**26**:i350–7.

80. Miller CA, Hampton O, Coarfa C, *et al.* ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One* 2011;**6**:e16327.

81. Li Y, Zheng H, Luo R, *et al.* Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat Biotechnol* 2011;**29**:723–30.

82. 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.

83. Kurtz S, Phillippy A, Delcher AL, *et al.* Versatile and open software for comparing large genomes. *Genome Biol* 2004;**5**:R12.

84. 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.

85. Abyzov A, Gerstein M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* 2011;**27**:595–603.

86. Stewart C. Spanner. https://github.com/chipstewart/Spanner (4 February 2011).

87. Abel HJ, Duncavage EJ, Becker N, *et al.* SLOPE: a quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data. *Bioinformatics* 2010;**26**:2684–8.

88. Magi A, Benelli M, Yoon S, *et al.* Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic Acids Res* 2011;**39**:e65.

89. 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.

90. Ding L, Ellis MJ, Li S, *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* 2010;**464**:999–1005.

91. Chiara M, Pesole G, Horner DS. SVM(2): an improved paired-end-based tool for the detection of small genomic structural variations using high-throughput single-genome resequencing data. *Nucleic Acids Res* 2012;**40**:e145.

92. Kidd JM, Cooper GM, Donahue WF, *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* 2008;**453**:56–64.

93. Kidd JM, Sampas N, Antonacci F, *et al.* Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nat Methods* 2010;**7**:365–71.

94. Levy S, Sutton G, Ng PC, *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* 2007;**5**:e254.

95. Wheeler DA, Srinivasan M, Egholm M, *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008;**452**:872–6.

96. Genome of the Netherlands C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 2014;**46**:818–25.

97. Atanur SS, Diaz AG, Maratou K, *et al.* Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat. *Cell* 2013;**154**:691–703.

98. Nellaker C, Keane TM, Yalcin B, *et al.* The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biol* 2012;**13**:R45.

99. Wang K, Hu Q, Ma H, *et al.* Genome-wide variation within and between wild and domestic yak. *Mol Ecol Res* 2014;**14**:794–801.

100. Rubin CJ, Megens HJ, Martinez Barrio A, *et al*. Strong signatures of selection in the domestic pig genome. *Proc Natl Acad Sci USA* 2012;**109**:19529–36.

101. Venn O, Turner I, Mathieson I, *et al*. Nonhuman genetics. Strong male bias drives germline mutation in chimpanzees. *Science* 2014;**344**:1272–5.

102. Kim HM, Cho YS, Kim H, *et al*. Whole genome comparison of donor and cloned dogs. *Nat Sci Rep* 2013;**3**:2998.

103. Axelsson E, Ratnakumar A, Arendt ML, *et al*. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* 2013;**495**:360–4.

104. Poelstra JW, Vijay N, Bossu CM, *et al*. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* 2014;**344**:1410–14.

105. Huang W, Massouras A, Inoue Y, *et al*. Natural variation in genome architecture among 205 *Drosophila melanogaster* genetic reference panel lines. *Genome Res* 2014;**24**:1193–208.

106. Supple MA, Hines HM, Dasmahapatra KK, *et al*. Genomic architecture of adaptive color pattern divergence and convergence in Heliconius butterflies. *Genome Res* 2013;**23**:1248–57.

107. Cao K, Zheng Z, Wang L, *et al*. Comparative population genomics reveals the domestication history of the peach, *Prunus persica*, and human influences on perennial fruit crops. *Genome Biol* 2014;**15**:415.

108. Gordon SP, Priest H, Des Marais DL, *et al*. Genome diversity in *Brachypodium distachyon*: deep sequencing of highly diverse inbred lines. *Plant J* 2014;**79**:361–74.

109. Wang L, Han X, Zhang Y, *et al*. Deep resequencing reveals allelic variation in *Sesamum indicum*. *BMC Plant Biol* 2014;**14**:225.

110. Blaby IK, Glaesener AG, Mettler T, *et al*. Systems-level analysis of nitrogen starvation-induced modifications of carbon metabolism in a *Chlamydomonas reinhardtii* starchless mutant. *Plant Cell* 2013;**25**:4305–23.

111. Jiao Y, Zhao H, Ren L, *et al*. Genome-wide genetic changes during modern breeding of maize. *Nat Genet* 2012;**44**:812–15.

112. Causse M, Desplat N, Pascual L, *et al*. Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genomics* 2013;**14**:791.

113. Li Y, Zhang W, Zheng D, *et al*. Genomic evolution of *Saccharomyces cerevisiae* under Chinese rice wine fermentation. *Genome Biol Evol* 2014;**6**:2516–26.

114. Golubchik T, Batty EM, Miller RR, *et al*. Within-host evolution of *Staphylococcus aureus* during asymptomatic carriage. *PLoS One* 2013;**8**:e61319.

115. Janbon G, Ormerod KL, Paulet D, *et al*. Analysis of the genome and transcriptome of *Cryptococcus neoformans* var. grubii reveals complex RNA expression and microevolution leading to virulence attenuation. *PLoS Genet* 2014;**10**:e1004261.

116. Hormozdiari F, Hajirasouliha I, McPherson A, *et al*. Simultaneous structural variation discovery among multiple paired-end sequenced genomes. *Genome Res* 2011;**21**:2203–12.

117. Hach F, Hormozdiari F, Alkan C, *et al*. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* 2010;**7**:576–577.

118. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics* 2014;**47**:11.12.11-11.12.34.

119. Grimm D, Hagmann J, Koenig D, *et al*. Accurate indel prediction using paired-end short reads. *BMC Genomics* 2013;**14**:132.

120. Teague B, Waterman MS, Goldstein S, *et al*. High-resolution human genome structure by single-molecule analysis. *Proc Natl Acad Sci USA* 2010;**107**:10848–53.

121. Carneiro MO, Russ C, Ross MG, *et al*. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics* 2012;**13**:375.