Computational methods for discovering structural variation with next-generation sequencing

Paul Medvedev[1], Monica Stanciu[1] & Michael Brudno[1,2]

In the last several years, a number of studies have described large-scale structural variation in several genomes. Traditionally, such methods have used whole-genome array comparative genome hybridization or single-nucleotide polymorphism arrays to detect large regions subject to copy-number variation. Later techniques have been based on paired-end mapping of Sanger sequencing data, providing better resolution and accuracy. With the advent of next-generation sequencing, a new generation of methods is being developed to tackle the challenges of short reads, while taking advantage of the high coverage the new sequencing technologies provide. In this survey, we describe these methods, including their strengths and their limitations, and future research directions.

The discovery of structural variation[1–3] as an important contributor to heterozygosity has revolutionized our understanding of the landscape of human genotypes (for an introduction, see ref. 4). Recognition of the extensive presence of structural variants in the human genome[5–11] has had broad ramifications for several areas of biology, including association studies[12], cancer genomics[13] and molecular evolution[14]. For instance, genetic variants associated with cancer often result from rearrangements and alterations in proto-oncogenes or tumor suppressor genes, and many chromosomal aberrations in tumor genomes have been found to join separate gene sequences into fusion genes[13,15,16]. Other diseases, such as autism[17] and Parkinson's disease[18], have also been associated with changes in gene dosage resulting from alterations in copy number. Besides biomedical applications, analysis of structural variants has led to a better understanding of how the genome has been shaped throughout evolutionary history. Recent studies found that although a surge in *Alu* element activity 40 million years ago fueled a high rate of nonallelic homologous recombination, the prominence of this mechanism has since diminished in favor of nonhomologous end joining[3,19].

These applications have driven the development of methods for the discovery of structural variants in the human population. In addition to specificity and sensitivity, a method's quality is judged by its ability to accurately predict the location of the breakpoints, the variant's size, and the change in copy count. Although the term 'structural variant' has in the past been used for events >1,000 base pairs (1 kbp) in size, this is a rather arbitrary cutoff, and we use structural variants to designate polymorphisms that change the structure of the genome, including all insertions, deletions and inversions. Structural variants are generally categorized on the basis of whether they affect the copy count of any genomic region. Events such as insertions and deletions (indels) are referred to as copy-number variants (CNVs), to distinguish them from events that are copy-count invariant, such as inversions.

The earliest methods for discovering structural variants are based on whole-genome array comparative genome hybridization (aCGH), which tests the relative frequencies of probe DNA segments between two genomes[20–22] (for a comprehensive review, see ref. 23). Alternative approaches take advantage of the extensive data available from the HapMap project[24] and use

[1]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [2]Banting and Best Department of Medical Research, and Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada. Correspondence should be addressed to M.B. (brudno@cs.toronto.edu).

PUBLISHED ONLINE 15 OCTOBER 2009; DOI:10.1038/NMETH.1374

BOX 1 MATE PAIRS/PAIRED ENDS

Sequencing technologies can generate two reads at an approximately known distance in the genome using two disparate sequencing strategies to generate reads from both sides of a segment of DNA (the insert). Mate pairs are created when genomic DNA is fragmented and size-selected inserts are circularized and linked by means of an internal adaptor. The circularized fragment is then randomly sheared, and segments containing the adaptor are purified. Finally, the mate pairs are generated by sequencing around the adaptor. Paired-end reads, by contrast, are generated by the fragmentation of genomic DNA into short (<300 bp, owing to space constraints on the slide) segments, followed by sequencing of both ends of the segment. Paired-end reads provide tighter insert-size distributions, and thus higher resolution, whereas mate pairs give the advantage of larger insert sizes. Although the wet-lab techniques used to generate these two types of data are very different, from a computational perspective the distinction between mate pairs and paired ends is not crucial. All of the methods described in this survey work for both paired-end and mate-pair data, and we will use the term mate pair throughout.

single-nucleotide polymorphism (SNP) arrays to measure the intensity of probe signals at known SNP loci. By also considering allelic ratios at heterozygous sites, they are able to detect CNVs[6,7,25–27], as well as, to a smaller extent, inversions[28]. However, although computational methods based on array data have been successfully used to identify CNVs, their power is limited. The size and breakpoint resolution of any prediction is correlated with the density of the probes on the array, which is limited by either the density of the array itself (for aCGH) or by the density of known SNP loci (for SNP arrays).

More recently, sequencing-based methods have used mate-pair or paired-end reads for structural variant discovery[2,3,5,9] (**Box 1**). In this approach, two paired reads are generated at an approximately known distance in the donor genome. The reads are mapped to a reference genome, and pairs mapping at a distance that is substantially different from the expected length, or with anomalous orientation, suggest structural variants. Paired-end mapping (PEM) techniques, which are based on the mining of such mate pairs, have been successfully used to discover structural variants, including copy-invariant events, at a much higher resolution than array-based methods.

Whereas earlier PEM-based methods use low-coverage Sanger-style sequencing, the last few years have seen the emergence of several high-throughput sequencing platforms, such as Roche's 454, Illumina's Genome Analyzer and ABI's SOLiD, which are able to sequence millions of reads simultaneously, thus substantially accelerating data acquisition. In these next-generation sequencing (NGS) platforms, clonal amplification is performed by PCR-based methods instead of bacterial transformation. PCR 'colonies' are tethered to an array and sequenced in parallel, using either polymerases or ligases for primer elongation during each cycle. Such parallel sequencing of the colonies not only markedly increases the speed of data generation, but also reduces the amounts of reagents needed. (See refs. 29–31 for reviews.) Finally, all three NGS platforms are capable of generating mate-pair or paired-end data, enabling their use for structural variant discovery using PEM techniques.

The advent of NGS has opened many opportunities for structural variant discovery. Though the various NGS platforms differ in certain aspects, they all provide a many-fold improvement over Sanger sequencing in throughput and cost per base pair. This has enabled initiatives such as the 1000 Genomes project[32], an ambitious undertaking that aims to sequence the genomes of 1,000 individuals using NGS platforms in order to expand databases of human variation and to help understand the genetic factors behind human diseases. Nevertheless, these technologies are also limited by the read lengths, which at present range from 35–400 bp, and the raw accuracy of base calls, making direct assembly of the whole genome from raw read data difficult.

A CATALOG OF SIGNATURES

Detecting structural variant differences between two individuals would be a trivial task if their genomes were already assembled. Because this is as yet prohibitive for humans, current methods use only one assembled genome (the 'reference') and another sequenced genome (the 'donor'). Thus, they are unable to compare the sequences directly, and instead rely on detecting variation through 'signatures'—patterns of PEMs that are created by structural variation. Most current methods distinguish themselves by the signatures they can detect, and we therefore first discuss these signatures and their intrinsic limitations.

Signatures based on PEM

Basic insertion, deletion and inversion. Two of the easiest and most commonly detected signatures are the 'basic insertion' and 'basic deletion'[2,15] (**Fig. 1a,b**). A mate pair that spans an isolated deletion event maps to the corresponding regions of the reference, but the mapped distance is greater than the insert size. Conversely, if the event is an insertion, then the distance is smaller. Another variant that leaves a clear signature is an inversion. A mate pair that spans either (but not both) of its breakpoints will map to the reference with the orientation of the read, lying within the inversion, flipped. Two such mate pairs, respectively spanning each of the two breakpoints, form the 'basic inversion' signature[2,8] (**Fig. 1c**).

Note that the basic insertion signature does not appear when the size of the insertion is greater than the insert size of the sequenced fragment, and it does not indicate the inserted sequence itself. However, if the inserted segment is present elsewhere in the genome, a different, linking, signature can be used to identify the connection between the location of insertion and the inserted sequence.

Linking: linked insertion and everted duplication. Consider two distant regions of the reference genome that are adjacent in the donor. A mate pair spanning the donor's breakpoint will map with a distance much greater than the insert size. The two spanning mate pairs that are closest to the breakpoint from the 3′ and 5′ ends, respectively, form a 'linking' signature (**Fig. 1d**). This signature is not associated with any particular type of event; it can be caused by any rearrangement that creates a donor adjacency that is not present in the reference. For example, a basic deletion signature is a type of linking signature, indicating that two segments that were not adjacent in the reference have become adjacent in the donor. Other linking signatures can connect regions that are arbitrarily distant or even on different chromosomes. For example, a fusion gene in a cancer cell can be identified by detecting a linking signature between two distant genes[33,34].

An insertion event where the inserted sequence is present elsewhere in the genome can create a 'linked insertion' signature[3],

which is composed of two linking signatures where the linked regions are close to each other (**Fig. 1e**). Unlike the basic insertion, the linked insertion signature can be used to identify the region that has been inserted. However, if the size of the insertion is large, then the confidence that the two linking signatures are associated with the same insertion decreases, and thus this signature becomes weak for very large insertions.

Another type of linking signature is created by a region of the reference that has been tandemly duplicated in the donor. Cooper *et al.*[7] first observed that a mate pair that has an end in each of the two copies will have an 'everted' mapping: the order of the mates is reversed while the orientation stays the same (**Fig. 1f**). We call this an 'everted duplication' signature. This signature can only be used to detect a novel tandem duplication—for example, it cannot detect a tandemly repeated region whose copy count changes from two to three.

All of the methods outlined above, although able to identify approximate locations of breakpoints, cannot indicate the exact locations. The methods below describe signatures that address this shortcoming.

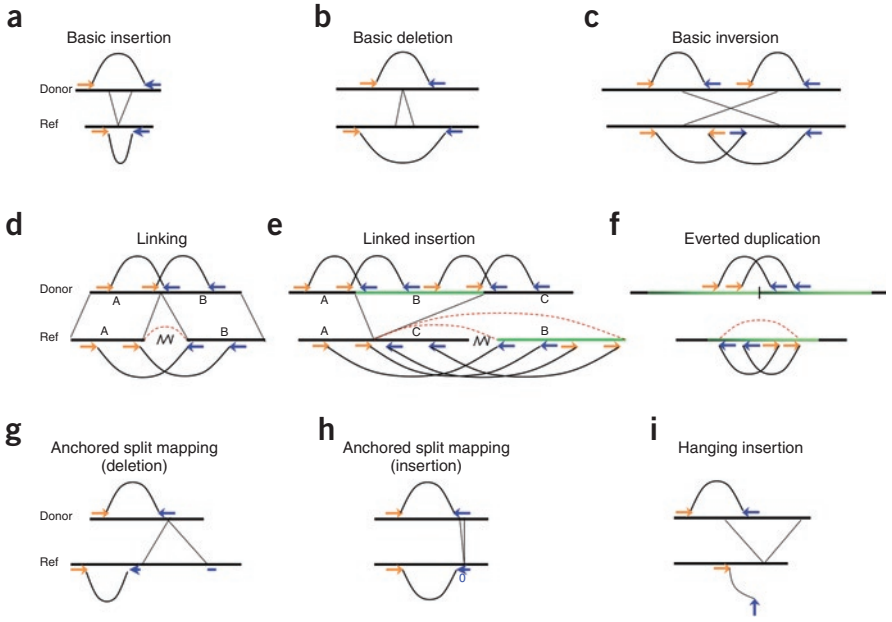

Figure 1 | Illustrations of PEM signatures. Mate pairs are sampled from the donor, where they are ordered with opposite orientation (the blue mate follows the orange), and are mapped to the reference (ref). Basic signatures include (**a**) insertions and (**b**) deletions, where the mapped distance is different from the insert size, as well as (**c**) inversions, where the order of the two mates is preserved but one of them changes orientation. (**d**) The linking signature has several discordant mate pairs with similar mapped distances identifying adjacency in the donor (dashed orange arrows) of two distal segments of the reference. The orientation and order of the mapped mate pairs depends on the orientation and order of the two segments in the reference; here, these are unchanged. (**e**) A linked insertion signature is composed of two linking signatures and arises when the inserted sequence (green) is copied from another location in the genome. (**f**) A tandem duplication will create an everted duplication linking signature, with mates mapping out of order but with proper orientations. These mate pairs link the end of the duplicated region to its beginning. (**g**,**h**) In the anchored split mapping signature, one mate has a good mapping, whereas the other has a split mapping. For a deletion (**g**) the prefix and suffix surround the deletion, whereas for an insertion (**h**) the split read has the prefix and suffix mapped to adjacent locations, while a middle part does not map. (**i**) When a novel genomic segment is inserted, a hanging insertion signature is created, in which only one of the mates has a good mapping.

Breakpoint identification: split mapping and hanging insertion. A read sampled across a deletion breakpoint will leave a 'split mapping' signature in the reference, with a prefix and suffix of the read mapping to different locations. Whereas this signature is detectable with longer reads[5,35], there are too many such spurious mappings of short read halves, and hence too many spurious signatures, with short read data. Nevertheless, Ye *et al.*[36] showed that if one uses the fact that the mate of a split read must map nearby, then the search space for the split mapping of the hanging read can be much reduced. Thus we have the 'anchored split mapping' signature, in which one of the mates maps to the reference and the other has a split mapping with one of its parts about 1 insert size away (**Fig. 1g**). A similar situation occurs when there is an insertion of a few base pairs. This will leave behind a similar signature, except that the split read will have a prefix and suffix mapping to adjacent locations, and there will be a middle part of the read (the bases inserted) that will not be part of either the prefix or suffix mapping (**Fig. 1h**).

The anchored split mapping signature has the advantage that it can pinpoint the breakpoint of the event with base-pair precision. However, if the deletion is too large, then there will be too many spurious hits for the farther part of the split mapping. Similarly, the size of the insertion detectable with this signature is only a few base pairs, as every inserted base reduces the fraction of the read that matches the genome.

To identify insertions that contain a novel genomic segment, it is possible to use mate pairs spanning either of the breakpoints, where one read maps while the other one does not. Such pairs form a 'hanging insertion' signature[5] (**Fig. 1i**). *De novo* assembly of such hanging reads can be used to reconstruct a small inserted segment, although if it is substantially larger than the insert size, hanging reads will not cover the entire insertion.

Signatures based on depth of coverage

The high coverage of NGS makes it possible to identify a completely different type of signature, based on the depth of coverage (DOC). Assuming the sequencing process is uniform, the number of reads mapping to a region follows a Poisson distribution and is expected to be proportional to the number of times the region appears in the donor. Thus, a region that has been deleted (duplicated) will have less (more) reads mapping to it. Although earlier work used DOC to identify recent segmental duplications in the human genome[37] and compare segmental duplications between human and chimp[38], Campbell *et al.*[34] were the first to use these 'gain/loss' signatures to detect CNVs between tumor and healthy samples of the same individuals (**Fig. 2**). Unlike the PEM insertion signatures, the gain signature does not indicate where an insertion occurred, but rather

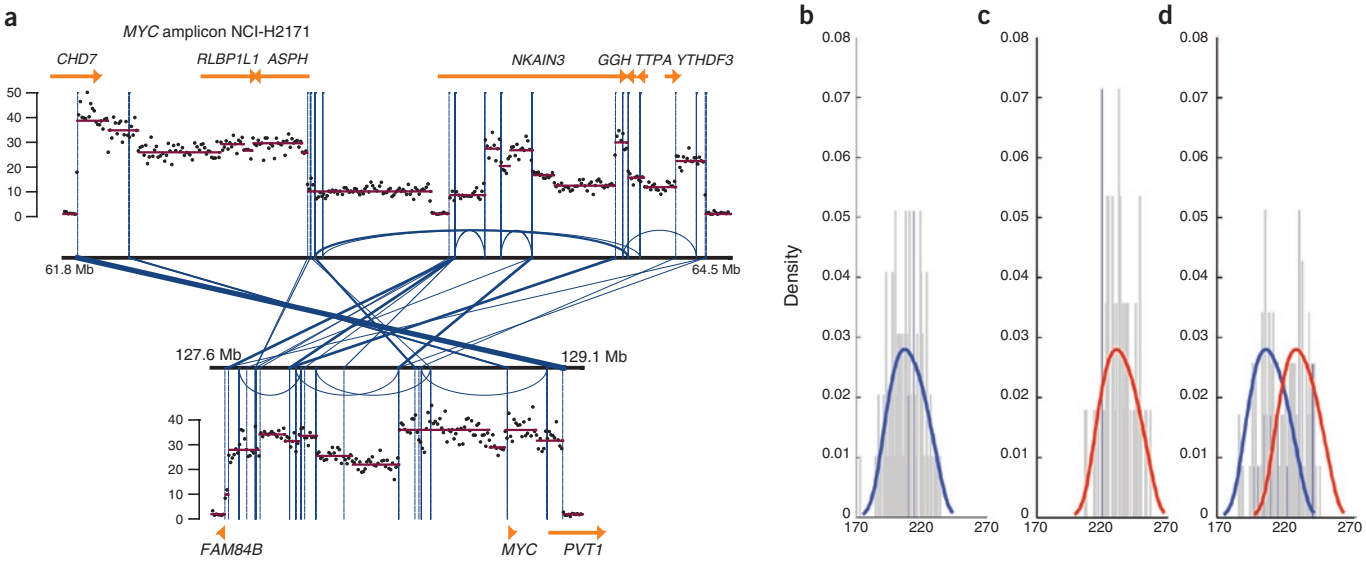

Figure 2 | Depth-of-coverage signature and distribution-based clustering. (**a**) Depth-of-coverage signature, and its relation to the linking signature (Campbell *et al.*[34]). Two distant genomic regions where the plotted dots, representing coverage depth, are grouped together into windows using horizontal bars to show the average depth. The vertical bars denote the borders between windows, where there is a sudden change in the coverage depth. The blue lines connect the two regions of a linking signature cluster. Observe that the ends of blue linking lines correspond closely to the location of the vertical bars. In other words, the loci of coverage depth changes correspond to the loci of linking signatures. (**b**–**d**) Distribution-based clustering (Lee *et al.*[11]). Each graph shows both the empirical (gray bars) and fitted (solid lines) distributions of the mapped distances of all concordant and discordant mate pairs spanning a given point. (**b**) Area where there is no variation. The mean of the distribution is at about the insert size (208 bp). (**c**) Homozygous deletion of size 24 bp. The mean is at ~232 bp. (**d**) Hemizygous deletion of size 22 bp, which results in a mixture of two separate distributions with means of 208 bp and 230 bp.

what duplicate sequence has been inserted; thus, it is not able to detect insertions of novel sequence.

Owing to its statistical nature, the strength of a gain/loss signature is directly related to the coverage of the dataset and to the size of the CNV. In contrast to most PEM signatures, DOC signatures can be used to detect very large events—in fact, the larger the event, the stronger the signature. However, they are not able to identify smaller events that PEM signatures, even with low coverage, are able to detect; they are also much poorer at localizing breakpoints.

METHODS FOR DETECTING SIGNATURES

The basic framework of all structural variant discovery methods is to detect signatures and then call the underlying variants. Different events leave behind different types of signatures, and therefore any method that aims to achieve high sensitivity across a wide range of events needs to detect a combination of signature types. In **Table 1**, we describe several existing methods in terms of the types they can detect.

Moreover, the observation of a single signature is usually insufficient to identify the underlying structural variant with high specificity, owing to the noise in the signature signal. Sequencing errors and chimeric reads may result in a read being incorrectly mapped to the reference genome, whereas chimeric clones will result in incorrect information about the distance and orientation between two reads. The PEM signatures are also dependent on the insert size, which in practice follows a distribution rather than being known exactly. Depending on the tightness of this distribution, it can be difficult to distinguish a true PEM signature caused by a small indel from a mate pair with an insert size from the tail of the distribution. The DOC signatures, by contrast, are affected by the sequencing bias of the current technologies, which cause certain regions of the genome to be over- or under-sampled. Though the bias in G+C rich regions and in homopolymeric stretches of DNA has been well documented[39,40], other sources of bias remain unaccounted for.

To overcome noise, current methods group signatures that support the same variation together into either clusters (for PEM approaches) or windows (for DOC approaches). Along with the types of detectable signatures, the techniques for clustering or windowing serve as the two main distinguishing characteristics of structural variant discovery methods. Having already discussed the types of signatures, we now turn our attention to aspects of clustering or windowing.

Paired-end mapping clusters

Clustering not only helps to improve confidence in the predictions but also to increase the precision of the predicted breakpoints and event size (**Fig. 3**). Clustering can have an even stronger impact with hanging insertion signatures, for which Kidd *et al.*[5] showed that, by feeding all the hanging reads within one cluster into an assembler, one can partially determine the novel inserted sequence.

The most common clustering strategy, introduced by Tuzun *et al.*[2], only considers mate pairs that do not have a 'concordant' mapping—one with the correct orientation and with a mapped distance within 2–4 s.d. of the mean insert size. Such mate pairs are called 'discordant'. Moreover, the strategy ignores any mate pair that has more than one good mapping. A cluster is then formed if there is at least some minimum number (usually two) of signatures of the same type and with similar size and location. This so-called standard clustering strategy was later used by the methods of Korbel *et al.* [3], Bentley *et al.*[9], Korbel *et al.*[41], Ye *et al.*[36], McKernan *et al.*[42] and Chen *et al.*[43].

Table 1 | Description of current methods for structural variant prediction with NGS

			Signatures detected									
Refs.	Name	Availability	Basic deletion	Basic insertion	Basic inversion	Linking	Linked insertion	Hanging insertion	Anchored split mapping	Everted duplication	Gain/loss	Clustering and/or windowing strategies
3, 41	PEMer	Downloadable	•	•	•	•	•					Standard
34		None				•					•	Binary circular segmentation
44	SegSeq	Downloadable									•	Local change-point analysis
9		In the future	•	•				•			•	Standard
10	VariationHunter	Downloadable	•	•	•					•		Soft
11	MoDIL	Downloadable	•	•								Soft, distribution-based
36	Pindel	Downloadable							•			Standard
43	BreakDancer	Downloadable	•	•	•			•				Standard, distribution-based
42	ABI Tools	Downloadable	•	•	•						•	Standard, distribution-based, binary circular segmentation

The two parameters that define this strategy—the minimum number of mate pairs required for a cluster and the number of standard deviations after which a mate pair is considered discordant—are interdependent and related to the coverage. That is, with increasing coverage one can decrease the number of mate pairs and/or standard deviations while achieving the same specificity. Similarly, one can decrease the number of standard deviations if one increases the number of mate pairs, and vice versa. Recently, Korbel *et al.*[41] and Bashir *et al.*[33] quantified these dependencies using both theoretical and simulation analyses.

One of the weaknesses of the standard clustering strategy is that, in ignoring mate pairs that have multiple good mappings, it does not allow the detection of signatures within repetitive regions of the genome. Given the strong association between segmental duplications and copy-number variation[14,19], these regions are actually among the most interesting to study. Several approaches have tried to address this problem by considering all good mate-pair mappings, to improve the sensitivity within duplicate regions[8,10,11]. Such 'soft' clustering approaches face the challenge of maintaining high specificity given the many spurious signatures that are created in this way; this is done by using various optimization procedures to assign each mate pair to a cluster where it will have the most support from other mate pairs. Thus, even though all good mappings are considered, a mate pair is allowed to be part of only one cluster.

Another limitation of the standard clustering strategy is that it uses a fixed cutoff for the number of standard deviations after which a mapped distance is considered to be discordant. This implies that even if there are multiple basic deletion signatures spanning a common point, and they all have a mapped distance of, for example, 1 s.d. away from the mean, then if the discordance threshold is set at 2 s.d. no cluster will be formed. This limitation is addressed by the method of Lee *et al.*[11], which forms basic indel clusters by looking at the distribution of all the mappings spanning a given location on the genome. If this distribution matches the typical insert size distribution but with a shift, then a cluster is formed. This allows the detection of much smaller indels than is possible with the standard clustering strategy. If, however, the distribution resembles a mixture of two separate insert size distributions with different means, then a hemizygous event can be detected by forming two separate clusters (**Fig. 2b**). This addresses another limitation of most current methods—namely, that they are not able to reliably distinguish between homozygous and heterozygous variants. Similar 'distribution-based' clustering approaches have been adopted by McKernan *et al.*[42] and Chen *et al.*[43].

Depth-of-coverage windows

Methods that use DOC signatures must partition the reference into windows so that the coverage depth is consistent within a window but has a sharp difference between adjacent windows. Then each window will correspond to a single loss, gain, or no event. Building upon similar techniques from aCGH, the methods of Campbell *et al.*[34] and Chiang *et al.*[44] use binary circular segmentation and local change-point analysis techniques to find the proper breakpoints for the windows. They use theoretical analysis and simulations to determine the minimum size of the windows so that they are large enough to have a statistically strong signal from the read distribution. This results in a minimum window size, and hence of a prediction, of at least 30 kbp. They help to mitigate the sequencing bias by making relative calls between two sequenced genomes (a healthy and a tumor line), instead of making absolute calls with respect to a reference.

Tools

Several methods have been packaged into algorithms and are publicly available, including SegSeq[44], PEMer[41], VariationHunter[10], MoDIL[11], Pindel[36], BreakDancer[43] and ABI SOLiD Software Tools[42]. Each one can be characterized in terms of two distinguishing factors: the signatures they detect and the way they cluster or window these signatures. These characterizations are shown in **Table 1**, and the experimental results of these studies on concrete data sets in **Table 2**. These tables can be used to guide a user's decision on which method is most applicable.

In addition to the methods already mentioned in the previous section, there have been more recent ones that have combined previously developed methodologies into a single framework[9,42,43]. For example, BreakDancer combines the standard clustering paradigm (BreakDancerMax) with the distribution-based approach proposed in MoDIL (BreakDancerMini), albeit without hemizygous event detection. ABI SOLiD Software Tools combine the standard clustering paradigm with a different distribution-based approach to indel identification and the binary circular segmentation algorithm

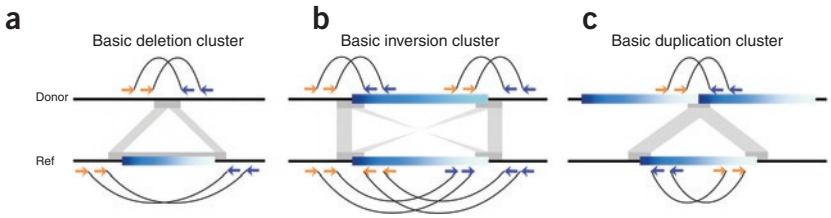

Figure 3 | The power of clustering for improving resolution. Blue gradient, the variant region; gray boxes, the possible location of the breakpoint. (**a**) With a basic deletion cluster, the possible location of the event can be narrowed down much more than with a single signature (**Fig. 1b**). (**b**) For a basic inversion, clustering helps to localize the breakpoints of the inverted segment. (**c**) An everted duplication cluster can more precisely localize the borders of the tandemly duplicated segment.

to identify regions of gain/loss. Another prominent tool is PEMer, a highly modularized framework for detecting structural variants that is specifically tailored to easy modification and development by the user. Some of the PEMer modules include read mapping, filtering of low-quality reads, signature detection, and clustering. Such a modularized framework has the potential to facilitate future algorithmic development by allowing improvements to particular modules without the need for implementing a whole structural variant discovery pipeline. However, there is still work to be done to create full-fledged, user-friendly tools for biologists.

STRENGTHS, LIMITATIONS, AND THE ROAD AHEAD

Despite current limitations, emerging technologies[45] promise to increase the read length to thousands of base pairs, so that the full assembly of human genomes would become economically feasible for structural variant discovery. Until then, however, Sanger sequencing continues to provide the longest reads, which undoubtedly offer many advantages over NGS, such as the ability to detect split mapping signatures. But clever techniques such as that of Ye *et al.*[36], who introduced anchored split mapping, can do much to bridge the gap. Moreover, the overriding advantage of NGS over Sanger sequencing continues to be the much lower cost and higher throughput.

Array CGH and other array-based technologies are considerably cheaper than NGS and will certainly play a role in genotyping individuals for previously known variants. However, for the discovery of new variants, NGS provides many advantages that have justified its higher cost. It is able to detect copy-invariant structural variants, such as inversions. It is better than aCGH at detecting smaller events and at determining the exact location of variation breakpoints. It does not suffer from oversaturation at high copy counts, allowing depth-of-coverage methods to be more accurate than aCGH at determining very high copy counts[34,41,44]. Moreover, it allows one to improve breakpoint resolution, copy-number accuracy, specificity and sensitivity by simply increasing coverage.

Despite these advantages of NGS, it is still not a one-stop solution for structural variation discovery, as some types of structural variants are more difficult to detect with NGS than with other approaches. For instance, current NGS-based methods have low sensitivity for detecting variation in repeating regions. However, these are among the regions that show the most variation: current estimates predict a strong enrichment of CNVs within segmental duplications[1,2,14,19,21,22,46–48]. Though DOC methods can detect such variation, their resolution is relatively poor. For most

Table 2 | The experimental results of existing structural-variant discovery studies using NGS

Refs.	Technology	Individual or cell line	Read length	Mean insert size	Coverage	Detectable events	Mean breakpoint resolution	Range of calls
3	454	NA15510 NA18505	109 bp	~3,000 bp	×2.1 ×4.3[a,b]	Ins, del, inv	644 bp	>3 kbp
34	Illumina	NCI-H2171 NCI-H1770	29–36 bp	~400 bp ~90 bp	2.4 Gb 1.8 Gb[a,c]	Ins, del	500 bp	>30 kbp
44	Illumina	HCC1954 HCC1143 HCI-H2347	32–36 bp	Unpaired	637 Mb 541 Mb 503 Mb[d]	Ins, del	440 bp	10–500 kbp
9						Ins, del	Not available	50 bp–35 kbp (del) 60–160 bp (ins)
10						Ins, del, inv	Not available	<500 kbp (del) <137 bp (ins) <10 Mb (inv)
11	Illumina	NA18507	~36 bp	~200 bp	~×42[e]	Ins, del	<100 bp	>20 bp (del) 20–120 bp (ins)
36						Ins, del	1 bp	<10 kbp (del) <20 bp (ins)
43						Ins, del, inv	Not available	>10 bp (del) 10–130 bp (ins)
42	ABI SOLiD	NA18507	25–50 bp	600–3,500 bp	~×15[e]	Ins, del, inv	Not available	>80 bp (del) 30–1,300 bp (ins)

Ins, insertion; del, deletion; inv, inversion.
[a]Total sequence generated by reads that were part of a mate pair that had a mapping that was not rejected by the algorithm. [b]With respect to the diploid genome. [c]Clone coverage. [d]Total sequence generated that had a high-quality alignment. [e]Total sequence generated with respect to the haploid genome.

PEM-based methods, conversely, the difficulty lies in the reliance on unique mappings. The introduction of soft clustering, which allows the use of mate pairs with multiple good mappings, is a step in the right direction; however, more work is needed before we can reliably predict variation in these regions.

Another important factor in the power of a study is the insert size of its library. Long insert sizes offer the advantage of allowing the detection of larger events. For instance, the strongest signature from an insertion is the basic insertion, and it is only present when the size of the insertion is less than the insert size. By contrast, shorter insert sizes increase the sensitivity for smaller events. Bashir *et al.*[33] performed a quantitative study, concluding that larger sizes are better for detection, whereas smaller ones are better for localization. Bentley *et al.*[9] also observed that, when using two libraries with different sizes, most of their predictions were unique to one data set. Thus, future studies may need to use multiple libraries with varying insert sizes to discover the whole size range of structural variants[9,34,41].

Many of the methods we have described here pioneered a new clustering strategy or a new type of signature when they were first introduced. These studies were intended to show the feasibility of a new technique and were not necessarily intended to provide a comprehensive method that incorporated the whole state of the art. As the field matures, newer methods will begin to combine previous approaches to improve their predictions. In particular, we believe one fruitful direction is the use of different types of signature to support any one event. For example, consider an insertion of a novel sequence, which might result in both basic and hanging insertion signatures. Current methods will consider support from one or the other, but not both, in making a call. However, by recognizing that the evidence from both types of signature supports the same event, one could detect it with less coverage and in the presence of more noise than current methods. Such approaches should also provide better resolution of sizes and breakpoints, areas where there is still much room for improvement.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
2. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
3. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
One of the first studies to use NGS data to detect structural variants, including using the linking signature for detecting insertions larger than the insert size.
4. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
5. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
6. McCarroll, S.A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
7. Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E. & Nickerson, D.A. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat. Genet.* **40**, 1199–1203 (2008).
8. Lee, S., Cheran, E. & Brudno, M. A robust framework for detecting structural variations in a genome. *Bioinformatics* **24**, i59–i67 (2008).
9. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
The first high coverage NGS dataset of an individual. This data set has been used in many subsequent studies.
10. Hormozdiari, F., Alkan, C., Eichler, E.E. & Sahinalp, S.C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* **19**, 1270–1278 (2009).
One of the first comprehensive tools for structural variant detection; supports most basic signatures and uses soft clustering.
11. Lee, S., Hormozdiari, F., Alkan, C. & Brudno, M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods* **6**, 473–474 (2009).
The first method to use a distribution-based clustering approach, allowing the detection of smaller indels, and explicitly modeling heterozygosity.
12. McCarroll, S.A. & Altshuler, D.M. Copy-number variation and association studies of human disease. *Nat. Genet.* **39**, S37–S42 (2007).
13. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
14. Cooper, G.M., Nickerson, D.E. & Eichler, E.E. Mutational and selective effects on copy-number variants in the human genome. *Nat. Genet.* **39**, S22–S29 (2007).
15. Volik, S. *et al.* End-sequence profiling: Sequence-based analysis of aberrant genomes. *Proc. Natl. Acad. Sci. USA* **100**, 7696–7701 (2003).
16. Raphael, B.J., Volik, S., Collins, C. & Pevzner, P.A. Reconstructing tumor genome architectures. *Bioinformatics* **19** (suppl. 2), 162–171 (2003).
17. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
18. Singleton, A.B. *et al.* Alpha-synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).
19. Kim, P.M. *et al.* Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res.* **18**, 1865–1874 (2008).
20. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20**, 207–211 (1998).
21. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
22. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
23. Carter, N.P. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* **39**, S16–S21 (2007).
24. International HapMap Consortium. The International HapMap Project. *Nature* **437**, 1299–1320 (2005).
25. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
26. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38**, 82–85 (2006).
27. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
28. Sindi, S. & Raphael, B. Identification and frequency estimation of inversion polymorphisms from haplotype data. in *Research in Computational Molecular Biology: Proc. RECOMB 2009* vol. 5541 (ed. Batzoglou, S.) 418–433 (Springer, Berlin, 2009).
29. Rusk, N. & Kiermer, V. Primer: Sequencing—the next generation. *Nat. Methods* **5**, 15 (2008).
30. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
31. Ansorge, W.J. Next-generation sequencing techniques. *New Biotechnol.* **25**, 195–203 (2009).
32. Siva, N. 1000 Genomes project. *Nat. Biotechnol.* **26**, 256 (2008).
33. Bashir, A., Volik, S., Collins, C., Bafna, V. & Raphael, B.J. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLOS Comput. Biol.* **4**, e1000051 (2008).
34. Campbell, P.J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
The first study to use the DOC signatures in NGS data, detecting CNVs in tumor samples.
35. Mills, R.E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190 (2006).
36. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect breakpoints of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* published online, doi:10.1093/bioinformatics/btp394 (26 June 2009).
A method that is able to detect indels with base-pair breakpoint

resolution using NGS data, on the basis of the anchored split mapping signature.
37. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
38. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88–93 (2005).
39. Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).
40. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10**, R32 (2009).
41. Korbel, J.O. *et al.* PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* **10**, R23 (2009).
42. McKernan, K.J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541; doi:10.1101/gr.091868.109 (22 June 2009).
43. Chen, K. *et al.* BreakDancer: An algorithm for high resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681; doi:10.1038/nmeth.1363 (9 August 2009).
44. Chiang, D.Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6**, 99–103 (2009).
45. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
46. Wong, K.K. *et al.* A comprehensive analysis of common copy-number variations in the human genome. *Am. J. Hum. Genet.* **80**, 91–104 (2007).
47. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
48. Locke, D.P. *et al.* Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. *Genome Biol.* **4**, R50 (2003).