Journal of Next Generation Sequencing & Applications

Ye, et al., Next Generat Sequenc & Applic 2016, S1:007
DOI: 10.4172/2469-9853.S1-007

Review Article Open Access

Structural Variation Detection from Next Generation Sequencing

Kai Ye[1*], George Hall[2,3] and Zemin Ning[2*]

[1]*McDonnell Genome Institute, Washington University School of Medicine, Forest Park Avenue, Saint Louis, Missouri, USA*

[2]*The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK*

[3]*Department of Computer Science, University of York, Heslington, York, YO10 5GH, UK*

[*]***Corresponding authors:*** *Kai Ye, McDonnell Genome Institute, Washington University School of Medicine, 4444 Forest Park Avenue, Saint Louis, Missouri 63108, USA, Tel: 1-314-813-0879; E-mail: kye@genome.wustl.edu*

Zemin Ning, The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK, Tel: 44-1223494705; E-mail: zn1@sanger.ac.uk

Rec date: Nov 26, 2015; **Acc date:** Feb 10, 2016; **Pub date:** Feb 15, 2016

Abstract

Structural variations (SVs) are the genetic variations in the structure of chromosome with different types of rearrangements. They comprise millions of nucleotides of heterogeneity within every genome, and are likely to make an important contribution to genetic diversity and disease susceptibility. In the genomics community, substantial efforts have been devoted to improving understanding of the roles of SVs in genome functions relating to diseases and researchers are working actively to develop effective algorithms to reliably identify various types of SVs such as deletions, insertions, duplications and inversions. Structural variant detection using Next-generation sequencing (NGS) data is difficult, and identification of large and complex structural variations is extremely challenging. In this short review, we mainly discuss various algorithms and computational tools for identifying SVs of different types and sizes with a brief introduction to complex SVs. At the end, we highlight the impact and potential applications of the 3rd generation sequencing data, generated from PacBio and Oxford Nanopore long read sequencing platforms.

Keywords: Structural variations; Bioinformatics; Algorithms; Next generation sequencing; Single molecule sequencing

Introduction

The complete catalogue of genetic variants contains substitutions, small insertions and deletions (indels) and large complex structural variants. The most abundant genomic variations are substitutions of single (single nucleotide variant, SNV or SNP) or multiple consecutive nucleic bases (multiple nucleotide variant or MNV). Due to having a higher density than other types of variants, SNPs are often used in genome-wide association studies (GWAS) to mark genome fragments related to diseases or certain traits. In addition to substitutions and mutations, insertions or deletions of bases in the DNA are another form of genomic variations between individuals. At low complexity regions, especially homopolymer runs or microsatellite tracks, substantially higher mutation rates are observed. Structural variation (SV) was originally defined as variation other than substitutions and small indels. Currently, SV includes large insertions, duplications, deletions, inversions and translocations. With substantial advances in sequencing technologies and analysis strategies, the definition of SV has been widened to include variants as small as 50 bp in length [1-3]. While substitutions and short indels could be visualized from short read data, SV was largely detected from indirect evidence of disturbance of read mapping around the variation. There are extensive studies on structural variations reported in the literature, in terms of both method development and data collection [1-4]. A number review papers have also been published on general methods [5,6], the analysis of human genome [7-9], mouse [10] and more recently pig [11]. However, algorithms designed for detection and archived datasets are predominantly for Illumina pair-end sequencing [4]. With recent advancing in single molecule sequencing with long read platforms such as PacBio and Oxford Nanopore, substantial impact and changes can be expected in the genomics community. In this review, our focus will be on the current algorithms used in the identification of SVs, a type of variant which are difficult to detect, but with huge medical and health implications. At the end, short discussions are presented about characteristics of the long read data and how it might impact on SV detection.

Methods of Structural Variation Detection

Over the past years, sequencing technology has fuelled genomic studies and many computational tools have been developed to deepen our understanding of structural variation and its role in genome functions. Rapid advancing in reducing NGS costs also makes genome-wide SV detection possible, even at population scale [1-3]. Detection of structural variation is an active research area and different methods are being explored with different applications. It is widely agreed that the method can be classified into four different algorithms: read-pair (RP), split-read (SR), read-depth (RD) and assembly. Figure 1 shows the four general methods one by one, where SVs are identified after reads are aligned against a given reference sequence. Table 1 is a list of structural variation tools published in the literature. It is not possible to compile a complete list as new tools are coming out every year. It should be noted that some recent tools combine more than one algorithm in order to increase specificity and sensitivity. Our aim is to provide an overall introduction to the methods generally used in structural variation detection and no effort is made to compare software tools based on the metrics of performance, specificity or sensitivity. Below, we briefly introduce these four algorithms and discuss strength and weakness associated with each method.

Program Name	Mutation Signals	SV Type	Data Input	Citations*	Authors and reference	URL
Breakdancer	RP	INS;DEL; INV;TRANS	WGS BAM	602	Chen, et al. [12]	http://breakdancer.sourceforge.net/
BreaKmer	Assembly	INS;DEL; INV;TAN; TRANS	WGS BAM	16	Abo, et al. [36]	https://github.com/a-bioinformatician/BreaKmer
Breakpointer	SR	INS;DEL	WGS BAM	11	Sun, et al. [27]	http://www.bioinfo.org/wiki/index.php/Breakpointer
BreakSeek	SR	INS;DEL	WGS SAM	1	Zhao, et al. [11]	http://sourceforge.net/projects/breakseek/
cn.MOPS	RD	CNV	WGS BAM	91	Klambauer, et al. [31]	http://www.bioinf.jku.at/software/cnmops/
CNVnator	RD	CNV	WGS BAM	306	Abyzov, et al. [30]	https://github.com/abyzovlab/CNVnator
CREST	SR	INS;DEL; INV; TRANS	WGS BAM	210	Wang, et al. [26]	http://www.stjuderesearch.org/site/lab/zhang
DELLY	RP+SR	DEL;INV; TAN;TRANS	WGS BAM	165	Rausch, et al. [14]	https://github.com/tobiasrausch/delly
GASVPro	RP+RD	DEL; INV	WGS BAM	66	Sindi, et al. [15]	http://compbio.cs.brown.edu/projects/gasv/
GenomeSTRiP	RD	DEL; CNV	WGS BAM	25	Handsaker, et al. [32]	http://www.broadinstitute.org/software/genomestrip/
HYDRA	Assembly	INS; DEL;INV; TAN	WGS BAM	158	Quinlan, et al. [37]	https://github.com/arq5x/Hydra
inGAP-sv	RD	INS;DEL; INV; TRANS	WGS SAM	49	Qi, et al. [33]	http://ingap.sourceforge.net/
LUMPY	RP+SR	INS;DEL; INV;TAN; CNV; TRANS	WGS BAM	52	Layer, et al. [16]	https://github.com/arq5x/lumpy-sv
MultiBreak-SV	RS	INS;DEL; INV;TRANS	WGS BLASR	7	Ritz, et al. [29]	https://github.com/raphael-group/multibreak-sv
NovelSeq	Assembly	INS	WGS FASTA/ FASTQ	81	Hajirasouliha, et al. [39]	http://novelseq.sourceforge.net/Home
Pindel	SR	INS;DEL; INV;TAN; TRANS	WGS BAM	602	Ye, et al. [23]	https://github.com/genome/pindel
PRISM	RP+RS	INS;DEL; INV;TAN	WGS SAM	50	Jiang, et al. [19]	http://compbio.cs.toronto.edu/prism/
RDXplorer	RD	CNV	WGS BAM	329	Yoon, et al. [34]	http://rdxplorer.sourceforge.net/
ReadDepth	RD	CNV	WGS bed	81	Miller, et al. [35]	https://github.com/chrisamiller/readDepth
SOAPindel	Assembly	INS;DEL	WGS SOAP/SAM	47	Li, et al. [38]	http://soap.genomics.org.cn/soapindel.html
SoftSearch	RP+SR	INS;DEL; INV;TRANS	WGS BAM	17	Hart, et al. [17]	https://code.google.com/p/softsearch/
SplitRead	RS	INS;DEL	RNA-seq FASTQ	71	Karakoc, et al. [28]	http://splitread.sourceforge.net/
SVdetect	RP or RD	INS;DEL; INV;TAN; CNV	WGS BAM	104	Zeitouni, et al. [18]	http://svdetect.sourceforge.net/Site/Home.html
SVseq2	SR	INS;DEL	WGS BAM	24	Zhang, et al. [24]	http://www.engr.uconn.edu/~jiz08001/svseq.html
TIGRA-SV	Assembly	INS;DEL; INV;TRANS	WGS BAM	30	Chen, et al. [12]	http://bioinformatics.mdanderson.org/main/

Table 1: List of computational tools for structural variation detection.

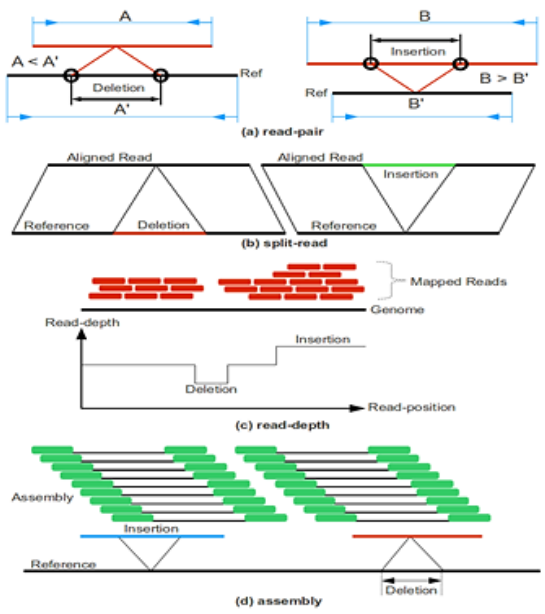

Figure 1: Methods of structural variation detection. (a) Read-Pair (RP) method is to estimate the likelihood of expected value of insert size variation associated with deletion and insertion; (b) Read-S (RS) method is capable of pinpointing exact breakpoints at base level accuracy; (c) Read-Depth based algorithm reports exact number of sequence copies in the genome; (d) local assemblies from extracted reads lead to accurate detection of deletions and insertions.

Read-Pair

The structural variant will disturb the normal read mapping pattern when using the standard reference genome. For example, mapping distance between two reads in a pair will be different if there are deletions or insertions between the two mapped reads, shown in Figure 1a. For inversion or translocations (where a short DNA sequence is reversed locally within a chromosome or chromosome abnormality caused by rearrangement of parts between non-homologous chromosomes), the relative mapping orientation will also differ. In principle, read-pair methods are capable of detecting most kinds of structural variants. As mapping distance or insert size between two reads in a pair normally follow a statistical distribution, the variation or tightness of the insert size distribution determines the minimum size of deletions and insertions. However, inversions and translocations do not have such a dependency. Break Dancer [12] is one of the first software packages implemented using the read-pair method, capable of detecting multiple variant types. It first gathers reads with abnormal insert size or orientation, and classifies them into normal, deletion, insertion, inversion, or translocation. It uses the insert size from each sequencing library, and every abnormal region is evaluated independently using a Poisson distribution. For the reasons stated above, this method does not capture small variants or very large insertions due to the limitation of insert size based methods. In general, Break Dancer is fast, computationally efficient, and capable of detecting multiple types of variation. However, as there are large numbers of repetitive sequences in the human genome, the two reads in a pair are often incorrectly mapped to remote regions of the genome. This means that many abnormal read-pair signals are misleading and are therefore the source of false positives. Currently, researchers often apply local assembly tools like TIGRA-SV [13] to further filter BreakDancer calls. In Table 1, the read-pair method is also implemented in other packages, such as Delly [14], GASVPro [15], LUMPY [16], Soft Search [17], SV Detect [18] and Prism [19]. These tools share similarities in the general read pair method, but differ in noise filtering and statistical estimation on insert size distributions. For example, Delly [14] is designed to combine both short insert read pairs and long range mate pairs to reduce false positives, while SVDetect uses a sliding-window strategy to identify all groups of pairs sharing a similar genomic location to improve its performance.

Split-Read

Besides read-pair signals, a structural variant also leaves a split-read signal in short read data, located at the breakpoints of the variant, Figure 1b. We need to break the short reads into two or more fragments and map them separately to the reference genome. The mapping location and orientation of the split-mapped reads will provide us with the precise location, size and types of variants. The split-read approach is powerful for detecting small and medium-size variants such as insertions, deletions and inversions, as well as translocations. This method provides the precise breakpoints with base accuracy down to single base levels. However, for large variants or those in repetitive regions, the split-read method is limited due to its local mapping procedure. We introduced split-read mapping to the field of next-generation sequence data analysis. In late 2008, the read length of NGS platforms was only 36 bp, and paired-end sequencing techniques had just been invented. The paired-end data enabled read-pair approaches making full use of read mapping distance and orientation. At that time, standard mapping tools like MAQ [20] and SOAP [21,22] were not able to create a long split-read mapping for indel calling. Due to sequence errors and short read lengths, a split-read approach with 36 bp reads was considered impossible, as 18 bp sequences (half of the read length) have too many genomic locations to map. To reliably split such short reads for indel calling, Pindel [23] was developed. The goal was to identify the breakpoints of large deletions (1 bp-10 kb) and medium size insertions (1-20 bp). Pindel is often referred to as anchored split-read mapping. When reads were short, we used one of the reads as an anchor in order to reduce the search space of the split-mapping of its mate. Pattern growth, originally a sequential pattern mining approach, was used in the package for efficient string matching. Alongside Pindel, there are now at least six other tools based on the method of split-read for SV detection. In SVseq2 [24], a split-read read alignment algorithm is developed using Burrows-Wheeler transform on local regions near the anchor, and Soft Search [17] uses co-localized split-reads alignment. Longer reads (76-100 bp) and partial mappings from the development of reads alignment tools make more SV discovery tools possible. Directly extracting breakpoint signals from cigar strings in the aligned SAM/BAM files, variants are efficiently detected. Implementations of this method can be found in packages of Break Seek [25], CREST [26], break pointer [27], Delly [14] and LUMPY [16]. Other types of data are also used to create split read tools, for example, Split Read [28] uses whole exome sequencing, and multi break-SV [29] uses both Illumina and PacBio reads.

Read-Depth

As the read mapping density (read-depth) is proportional to the number of copies in the reference genome for unique genomic fragments, duplicated or deleted regions will have significantly higher or lower coverage (Figure 1c). Therefore, one can estimate the short read mapping density in a genomic region to estimate the relative ploidy or copy number. In general, read-depth approaches assume a random distribution in mapping depth. Read-depth signals of per-sample and per genomic region with GC correction are used to increase the signal-noise ratio. Read-depth methods are more effective for larger (>1 kb) copy number variants (CNVs). However, it is not able to identify copy number neutral variants like inversions. As simple as counting the number of reads mapped in a given region, read-depth has proven to be enough to determine the copy-number changes on a large scale. However, if the duplicated or deleted fragments are in repetitive regions, the detection sensitivity is greatly reduced in proportion to the level of repetitiveness. For example, if there is only one copy of a given genomic region in a diploid genome, a copy number gain or loss will either increase or decrease the read-depth signal by 50%, respectively. However, if there are already hundreds of copies in the genome, a single copy loss will not affect a noticeable change in the read-depth signal due to random mapping of reads among highly similar genomic fragments. The CNVnator [30] tool is one of many classical read-depth based approaches. It includes many steps, such as data pre-processing, read-depth calculation, adjustment based on GC content, mean-shift approach, clustering binning and merging. It is a robust way to detect copy number changes from whole genome sequence data, and has been applied successfully in the 1000 Genome Project as well as other major sequencing projects. In Table 1, other tools using this method are listed as well, such as cn.MPOS [31], GenomeStrip [32], inGAP-sv [33], RDXplorer [34] and ReadDepth [35].

Assembly

With further advances in Illumina pair-ending sequencing, read length has been improved continuously over the past years from initially 2 × 36 bp to 2 × 150 in a HiSeq platform and to 2 × 300 bp in a MiSeq platform. With the increase in read length, a method using assembly seems plausible in those difficult regions of the genomes. The assembly method discussed in this article is reference-assisted local assembly, rather than a de novo whole genome build. Often, sets of reads are collected after alignment with missing read-pairs, or other cases with difficulties in read mapping. Local assembly is performed on the read sets and variants are called from assembled contigs. TIGRA-SV [13] is a structural variant assembly tool to determine breakpoint sequences. It was initially coupled with Break Dancer at the Genome Institute of Washington University to reduce false positive rate for both Break Dancer's germline and somatic calls. TIGRA-SV first takes predicted breakpoints from other tools and then extracts reads around them. All reads with at least one end mapped in the local region will be extracted. It then attempts to assemble them into longer contigs using a de Brujin graph-based method with variable k-mer lengths. It reports results as fast files and subsequent alignment to the reference genome is required. This method has been applied in the 1000 Genomes Project and provided high quality breakpoint information. In the literature, we have also found other packages adopting the local assembly method, shown in Table 1, such as BreaKmer [36], HYDRA [37], SOAPindel [38] and NovelSeq [39].

Combination of more than one type of signal

There are a number of tools which have been implemented with more than one algorithm aiming for higher specificity and sensitivity. The DELLY method is the first tool in the field to combine read-pair and split-read signals. For each BAM file, DELLY computes the major read-pair orientation and insert size distributions. Then it scans all reads for abnormal mapping reads with either orientation or insert size significantly different than the majority. Based on the read mapping properties, reads may be classified as groups which support deletions, inversions, tandem duplications and translocations. The read-pair signals provide genomic intervals for downstream split-read mapping in order to fine-tune the breakpoints. In some cases, there are huge pile-ups of reads and many SV predictions, related to poor reference genome assembly in the repetitive regions. A threshold of maximum number of reads was set up to reduce memory footprint and to speed up calculation. In the 1000 Genomes Project, thousands of samples were sequenced to provide low coverage data genome-wide. On one hand, low coverage data enables us to include more samples in the study, but it significantly reduces the sensitivity of tools working on individual data. GenomeSTRiP combines data from a population and gathers variant signals of read-pair and read-depth for an accurate detection and genotyping of large deletions. Most structural variation detection tools use either single or complementary signals. However, comprehensive detection all types and size spectra of structural variants demands integration of all variation signals in addition to prior knowledge. Layer et al. [16] implemented LUMPY, a probabilistic framework for comprehensive structural variant discovery. Distinct modules process each variant signal and convert them to probability distributions. This approach is extendible, enabling one to plug in other variant types or output from another structural variant tool.

Complex SV

The distinct variant types in SV described above are standard in the research community. However, researchers recently uncovered substantial levels of complexity in both cancer genome study and population-scale sequencing results. In a recent study from the Cancer Genome Atlas project, an updated Pindel tool was used to scan over 8,000 cancer cases with an aim to identify complex indels in key cancer genes [40]. Complex indels are defined as deletion and insertion of DNA fragments with different sizes at a common genomic location. Complex indels were reported in earlier studies using traditional Sanger sequencing instruments. However, the detection is extremely challenging using sequencing data from the next generation short read sequence data, which results in being under-represented in current genomics studies. Out of 8,000 cancer cases, there were 285 complex indels identified in cancer genes like PIK3R1, TP53, ARID1A, GATA3, and KMT2D in approximately 3.5% of cases analyzed. Many of the newly discovered somatic complex indels were overlooked or mis-called as simple substitution variants.

Emerging technologies of 3rd generation sequencing

Currently there are two platforms which offer single molecule sequencing with long reads. Single Molecule Real Time sequencing (SMRT) from PacBio is a parallelized single molecule DNA sequencing method. A recent study led by Evan Eichler [8] sequenced a haploid human genome (CHM1) using PacBio long reads to identify missing sequence and genetic variation. The sequenced read-depth is 40X and the reads have an average mapped read length of 5.8 kb. Reads were first aligned to the reference assembly and gap regions were extended

using a reiterative map-and-assemble strategy. At each stage, a new consensus sequenced was obtained from SMRT whole-genome sequencing (WGS) reads mapped to each edge of a gap. In turn, the new consensus served as a template for recruiting additional sequence reads for assembly. In the study, they reported the identification of 26,079 euchromatic structural variants at the base-pair level, including inversions, complex insertions and long tracts of tandem repeats. Another study using assembly for SV detection was reported by Pendleton, et al. [41], who used 46X PacBio long reads and 80X long molecule Bionano mapping data. Another emerging single molecule sequencing platform is Oxford Nanopore's MinION, a real time nanopore-based DNA sequencing instrument [42-44]. With reported features of compact, inexpensive, long read length and fast sequencing data production, the MinION device offers great potential for genome analysis from structural variation perspectives. Figure 2a shows read length profiles from two datasets, *E. coli* [45] and yeast [46]. In the future, it would be interesting to know how to use long reads to accurately annotate complex SVs, such as the one shown in Figure 2b, where popular SV calling tools failed using short Illumina reads. From the study reported in [41], there seems to be a satisfactory solution using local assemblies obtained from long reads with high coverage. For cost reasons, however, challenges ahead for bioinformatics scientists are to seek practical solutions with low read coverage, say 5-10 at which assembly is not able to be done.

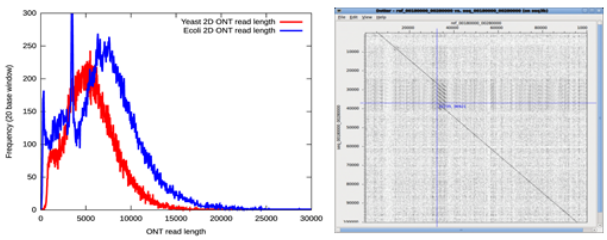

Figure 2: Can data from long read sequencing technology solve complex SVs like shown in (b)? (a) Read length distribution of two datasets recently released from the community. The *E. coli* dataset from Oxford Nanopore MinION [46] contains 295 Mb of bases with an average length of 7,915 bp and the largest read length at 45,588 bp. The yeast MinION dataset [47] contains 373 Mb of DNA bases with an average length of 5,651 bp and the largest read length at 32,826 bp. (b) A structural variant is shown with a global insertion of 6 kb, but also with 6 copies of local variable number of tandem repeats. It was missed by all the tools using Illumina short reads.

Conclusion

A full range of structural variation can be detected from NGS data, including insertions, deletions, inversions, CNVs and translocations. In this short review paper, we discussed bioinformatics algorithms which have been implemented into various tools in identifying structural variations. Each method of detection shows advantages/disadvantages. For example, CNV types of variations are well suited for RD based methods, while SR algorithms can accurately identify SV breakpoints down to single-base resolution. However, it should be noted that there is currently no single informatics method or algorithm which is capable of identifying the full range structural DNA variation. Recently, developers are attempting to incorporate multiple algorithms into a single package for robust variant detection. In practical applications, the users are advised to combine calling results obtained from multiple complementary tools in order to increase sensitivity and specificity. Advances in long read sequencing platforms such as PacBio and Oxford Nanopore, provide exciting prospects in SV detection. The benefit gained from long reading length and drawback of high error rate at base level remain to be fully examined and quantified in details, particularly at low read coverage.

Acknowledgement

George Hall and Zemin Ning are supported by the Wellcome Trust.

References

1. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, et al. (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470: 59-65.
2. Sudmant P, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, et al. (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526: 75-81.
3. Zarrei M, MacDonald JR, Merico D, Scherer SW (2015) A copy number variation map of the human genome. Nature Review Genetics 16:172-183.
4. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Research 42: D986-D992.
5. Lin K, Smit S, Bonnema G, Sanchez-Perez G, de Ridder D (2015) Making the difference: integrating structural variation detection tools. Briefings in Bioinformatics 16: 852-864.
6. Tattini L, D'Aurizio R, Magi A (2015) Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Front Bioeng Biotechnol 3:92.
7. Abel Hj, Duncavage EJ (2014) Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. Cancer Genetics 206:432-440.
8. Alkan C, Bradley PC, Eichler EE (2011) Genome structural variation discovery and genotyping. Nature Reviews Genetics 12: 363-376.
9. Hurles ME, Dermitzakis ET, Tyler-Smith C (2008) The functional impact of structural variation in humans. Trends Genet 24: 238-245.
10. Keane TM, Wong K, Adams DJ, Flint J, Reymond A et al. (2014) Identification of structural variation in mouse genomes. Front. Genet.
11. Zhao P, Li J, Kang H, Wang H, Fan Z, et al. (2016) Structural Variant Detection by Large-scale Sequencing Reveals New Evolutionary Evidence on Breed Divergence between Chinese and European Pigs. Scientific Reports 6: 18501.
12. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, et al. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods 6: 677-681.
13. Chen K, Chen L, Fan X, Wallis J, Ding L, et al. (2014) TIGRA: A targeted iterative graph routing assembler for breakpoint assembly. Genome Research 24: 310-317.
14. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, et al. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28: i333-i339.
15. Sindi SS , Znal S, Peng LC, Wu HT, Raphael BJ (2012) An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biology 13: R22.
16. Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: a probabilistic framework for structural variant discovery. Genome Biology 15: R84.

17. Hart SN, Sarangi V, Moore R, Baheti S, Bhavsar JD, et al. (2013) SoftSearch: Integration of Multiple Sequence Features to Identify Breakpoints of Structural Variations. PloS One 8: e83356.
18. Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, et al. (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics 26:1895-1896.
19. Jiang Y, Wang Y, Brudno M (2012) Prism: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. Bioinformatics 28: 2576-2583.
20. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18: 1851-1858.
21. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24: 713-714.
22. Li R, Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966-1967.
23. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865-2871.
24. Zhang J, Wang J, Wu Y (2012) An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. BMC Bioinformatics 6: S6.
25. Zhao H, Zhao F (2015) BreakSeek: a breakpoint-based algorithm for full spectral range INDEL detection. Nucleic Acids Res 43: 6701-6713.
26. Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, et al. (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nature Methods 8: 652-654.
27. Sun R, Love MI, Zemojtel T, Emde AK, Chung HR et al. (2012) Breakpointer: using local mapping artifacts to support sequence breakpoint discovery from single-end reads. Bioinformatics 28: 1024-1025.
28. Karakoc E, Alkan C, O'Roak BJ, Dennis MY, Vives L, et al. (2011) Detection of structural variants and indels within exome data. Nature Methods 9: 176-178.
29. Ritz A, Bashir A, Sindi S, Hsu D, Hajirasouliha I, et al. (2014) Characterization of structural variants with single molecule and hybrid sequencing approaches. Bioinformatics 30: 3458-3466.
30. Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research 21: 974-984.
31. Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, et al. (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res 40: e69.
32. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, et al. (2015) Large multiallelic copy number variations in humans. Nature Genetics 47: 296-303.
33. Qi J, Zhao F (2011) inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. Nucleic Acids Res 39: W567-W575.
34. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Research 19: 1586-1592.
35. Miller CA, Hampton O, Coarfa C, Milosavljevic A (2011) ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. PLoS One 6: e16327.
36. Abo RP, Ducar M, Garcia EP, Thorner AR, Rojas-Rudilla V, et al. (2015) BreaKmer: detection of structural variation in targeted massively parallel sequencing data using kmers. Nucleic Acids Res 43: e19.
37. Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, et al. (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Research 20: 623-635.
38. Li S, Li R, Li H, Lu J, Li Y, et al. (2013) SOAPindel: Efficient identification of indels from short paired reads. Genome Research 23: 195-200.
39. Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, et al (2010) Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics 26: 1277-1283.
40. Ye K, Wang J, Jayasinghe R, Lameijer EW, McMichael JF, et al. (2015) Systematic Discovery of Complex Indels in Human Cancers, Nature Medicine 22: 97-104.
41. Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, et al. (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nature Methods 12:780-786.
42. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, et al. (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517: 608-611.
43. Quick J, Quinlan A R, Loman NJ (2014) A reference bacterial genome dataset generated on the minion™ portable single-molecule nanopore sequencer. Gigascience 3: 22.
44. Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W (2015) Nanopore sequencing detects structural variants in cancer.
45. Ip CLC, Loose M, Tyson JR, de Cesare M, Brown BL, et al. (2015) MinION Analysis and Reference Consortium: Phase 1 data release and analysis.
46. Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, et al. (2015) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol 33: 296-300.
47. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, et al. (2015) Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Research 26: 1750.

This article was originally published in a special issue, entitled: **"Sequencing Technologies"**, Edited by Jianping Wang