
Mining for Structural Variations in Next-Generation Sequencing Data

Minja Zorc, Jernej Ogorevc and Peter Dovč

Additional information is available at the end of the chapter

<http://dx.doi.org/10.5772/intechopen.76568>

Abstract

Genomic structural variations (SVs) are genetic alterations that result in duplications, insertions, deletions, inversions, and translocations of segments of DNA covering 50 or more base pairs. By changing the organization of DNA, SVs can contribute to phenotypic variation or cause pathological consequences as neurobehavioral disorders, autoimmune diseases, obesity, and cancers. SVs were first examined using classic cytogenetic methods, revealing changes down to 3 Mb. Later techniques for SV detection were based on array comparative genome hybridization (aCGH) and single-nucleotide polymorphism (SNP) arrays. Next-generation sequencing (NGS) approaches enabled precise characterization of breakpoints of SVs of various types and sizes at a genome-wide scale. Dissecting SVs from NGS presents substantial challenge due to the relatively short sequence reads and the large volume of the data. Benign variants and reference errors in the genome present another dimension of problem complexity. Even though a wide range of tools is available, the usage of SV callers in routine molecular diagnostic is still limited. SV detection algorithms rely on different properties of the underlying data and vary in accuracy and sensitivity; therefore, SV detection process usually utilizes multiple variant callers. This chapter summarizes strengths and limitations of different tools in effective NGS SV calling.

Keywords: bioinformatics, genome organization, next-generation sequencing, structural variation, variant calling

1. Introduction

First, efforts in exploring genetic variations were focused on single-nucleotide polymorphisms (SNPs) which were initially considered the main source of genetic and phenotypic human variation [1], while larger variations were thought to be rare events. However, in 2004

two studies [2, 3] revealed an unexpectedly large amount of large-scale variations (several kb to hundreds of kb) in the human genome. The evidence for the prevalence of structural variants (SVs), such as deletions, duplications, and inversions, began to accumulate. By changing the organization of the DNA, SVs can contribute to the phenotypic differences among healthy individuals or cause severe phenotypic consequences. SVs are involved in a wide range of diseases and conditions, such as autism spectrum disorders [4–6], schizophrenia [7], Crohn's disease [8], rheumatoid arthritis [9], lupus erythematosus [10], psoriasis [11], obesity [12], and cancers [13, 14]. Among the different classes of genetic variations, SVs have remained the most challenging to detect and characterize. SVs were examined since the identification of chromosomal abnormalities using classic cytogenetic methods, revealing changes down to 3 Mb. Later techniques for SVs detection are based on array comparative genome hybridization (aCGH) and single-nucleotide polymorphism arrays. Next-generation sequencing (NGS) has enabled methods for precise definition of breakpoints of SVs of different sizes and types. Characterization of SVs from high-throughput sequencing data presents complex task due to the volume of the data and short sequence reads.

2. Structural variations

Genomic structural variations (SVs) are genetic alterations that result in amplifications, losses, inversions, and translocations of segments of DNA greater than 50 bp. SVs are a normal part of genomic variation but can also cause disorders. Standard detection methods include chromosome banding, fluorescent in situ hybridization (FISH), and array comparative genome hybridization (aCGH) that is very useful to detect copy number variations (CNVs) but cannot detect copy-neutral SVs (inversions, balanced translocations) [15]. Recent methods include employment of NGS to identify SVs, which are not detectable by cytogenetic methods.

Chromosomal rearrangements can occur on a single chromosome (interchromosomal SVs) or can involve exchange of genomic DNA between chromosomes (intrachromosomal SVs). Intrachromosomal SVs are a product of one or more double-strand breaks, which may result in deletions, inversions, and duplications. Deletions and duplications are copy number variations and are easily detected by employing NGS data (read coverage method), whereas inversions are copy number-neutral. Intrachromosomal translocation is the exchange of genetic material between two non-homologous chromosomes. In a reciprocal translocation, two broken-off pieces of two non-homologous chromosomes are exchanged, usually producing two balanced derivative chromosomes. Unless breakpoints disrupt important developmental genes, balanced translocations do not affect phenotype [15]. However, during gamete formation such chromosomes may segregate in unbalanced manner or unbalanced translocations may occur *de novo* and lead to monosomy and trisomy of different chromosome segments [16], which account for approximately 1% of developmental delay and intellectual disability cases in human [17–19]. Robertsonian translocations are a type of SVs resulting from chromosome end breaks near centromeric regions of two acrocentric chromosomes and their reciprocal exchange, which results in one large metacentric chromosome and one very small

chromosome that is usually lost without phenotype effect. In case three or more chromosomal breakpoints are involved, we speak of complex chromosome rearrangements, which may result in balanced or unbalanced state [20].

3. Next-generation sequencing

The first commercially available next-generation sequencing platform was released in 2005 [21]. The technology has been continuously upgraded and has fundamentally changed the field of genetics studies. Next-generation sequencing (NGS), also known as high-throughput sequencing, parallelizes the sequencing process and produces millions of short reads (50–400 bp each) in a single experimental run. It has contributed to rapid progress in single-nucleotide polymorphisms detection. Due to the nature of the NGS short-read sequences, the category of longer variants remained poorly characterized. Variants in range 10–100 kb are small for detection by cytogenetic methods [22] but too large for reliable detection with short-read sequencing. SVs affect more bases than single-nucleotide polymorphisms [23] and present an important class of genetic variation. Moreover, many SVs have been shown to play relevant roles in phenotypic variability and disease [24].

3.1. NGS data analysis pipeline

Once the samples are sequenced, the NGS data analysis becomes the task in bioinformatics field. The computational analysis and interpretation of the data generated remains one of the major bottlenecks in NGS projects. The basic steps for analyzing NGS data are quality assessment, reads alignment (mapping) to a reference sequence, and variant identification. The second stage of analysis comprises variant analysis, visualization, and interpretation of the variants in relation to phenotypes. Commercial packages such as CLCBio Genomic Workbench, CASAVA, and SeqNext often provide all-in-one solutions, while academic pipelines typically consist of sequential tools for specific steps in the analysis.

The output from the sequencing machines are reads, which are usually stored in text-based FASTQ files. The data obtained from NGS are compromised by sequence artifacts, including read errors, poor-quality reads, and primer contamination [25]. To avoid erroneous conclusions, the artifacts should be removed. A number of bioinformatics tools for sequencing quality assessment, such as FastQC, FASTX-Toolkit, PRINSEQ [26], TagDust [27], and NGS QC Toolkit [28] are designed. Next step in NGS data analysis is alignment of short reads to corresponding positions on a reference sequence. A variety of algorithms have been developed for this task. Representative read mappers are Bowtie2 [29], BWA [30], and Novoalign. The typical output from the read mapper is BAM file which contains information about qualities and positions of aligned sequences. Variant analysis consists of genotyping, variant calling, annotation, and prioritization. Genomic variants, such as SNPs and short-scale insertions and deletions are identified by variant callers. Widely used tools for variant calling are Genome Analysis Tool Kit HaplotypeCaller (GATK-HC) [14], Samtools mpileup [31], Freebayes and Torrent Variant Caller (TVC). Variant callers take in a BAM file and return a list of variants. To annotate

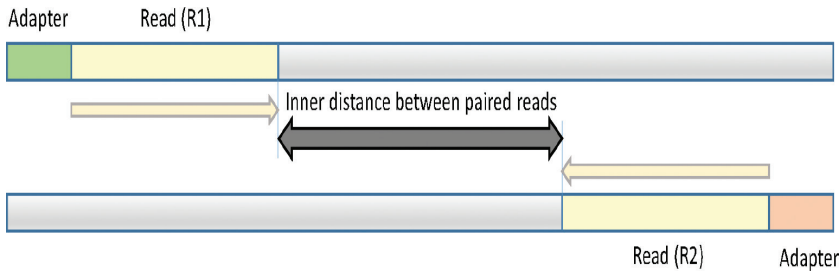


Figure 1. Paired-end sequencing; the inner distance between paired reads (R1 and R2) is known.

variants, SnpEff [32], VariantAnnotator from the GATK [33], and ANOVAR [34] tools are used. To systematically filter, evaluate, and prioritize thousands of variants VAAST 2.0 [35], VarSifer [36], KGGseq [37], and commercial software Ingenuity Variant Analysis are available.

3.2. Single-read and paired-end sequencing

Initially, NGS technologies produced extremely short reads (25–36 bp), sequenced from only one end of the DNA (single-read sequencing) [38]. As technology developed, read lengths consistently increased and sequencers have been improved to sequence both ends of a fragment with or without a non-sequenced stretch in between (paired-end sequencing). This not only has the benefit from doubling the number of reads but also improves accuracy and offers additional information for structural variants detection.

The reads obtained from paired-end sequencing (R1 and R2) come from the same fragment of DNA. The length of the fragment is usually longer than the length of reads ($R1 + R2$), so there is a gap between them (**Figure 1**). Although the sequence of the fragment between reads is not known, the knowledge that R1 and R2 are next to each other on the known distance and have opposite orientation is useful.

4. Overview of the structural variation detection algorithms

Using NGS technologies, large volume of sequence data at an unprecedented speed and constantly reducing cost is produced. Consequently, the computational tools for analysis of massive amounts of genomic data are in demand. There is a growing awareness that structural variations represent a significant contribution to genotypic and phenotypic diversity [39]. However, the accurate detection of structural variants from NGS is a daunting task [40]. A number of algorithms have been proposed to address the issue of structural variants calling from NGS data [41]. SV detection algorithms rely on different properties of the underlying data and vary in accuracy and sensitivity. The algorithms follow one or a combination of strategies, which could be classified into categories: (1) read depth (RD), (2) paired-end (PE), (3) split reads (SR), and (4) de novo assembly (AS). The most suitable method for SV detection depends on the size and variant type as well as characteristics of the sequencing data [42]. SV detection process usually utilizes multiple variant callers.

4.1. Algorithms based on read depth

Read depth (RD) algorithms are able to identify CNVs. RD-based algorithms can accurately predict absolute copy-numbers [43] but are unable to detect copy-number neutral variants such as inversions and balanced translocations. The breakpoint identification resolution is low and depends on the sequence coverage.

RD algorithms divide the reference sequence in intervals and calculate the number of reads aligned within them. The read depth per interval should follow a normal distribution centered at the average read depth for the entire reference sequence. When the read depth of contiguous intervals significantly differs from the average observed, the CNV is detected (**Figure 2**). Deleted regions show reduced read depth when compared to entire reference sequence (**Figure 3**).

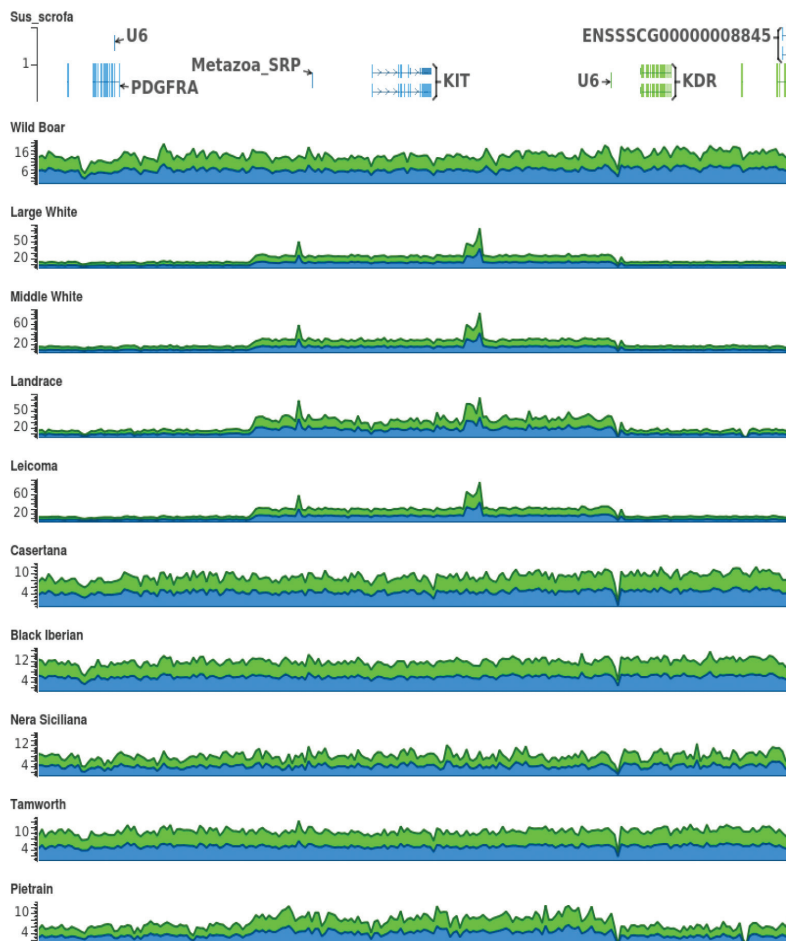


Figure 2. An example of CNV including gene *KIT* with flanking regions in four pig genomes. The read coverage is higher in the region of the CNV. The figure was made using Golden Helix GenomeBrowse.

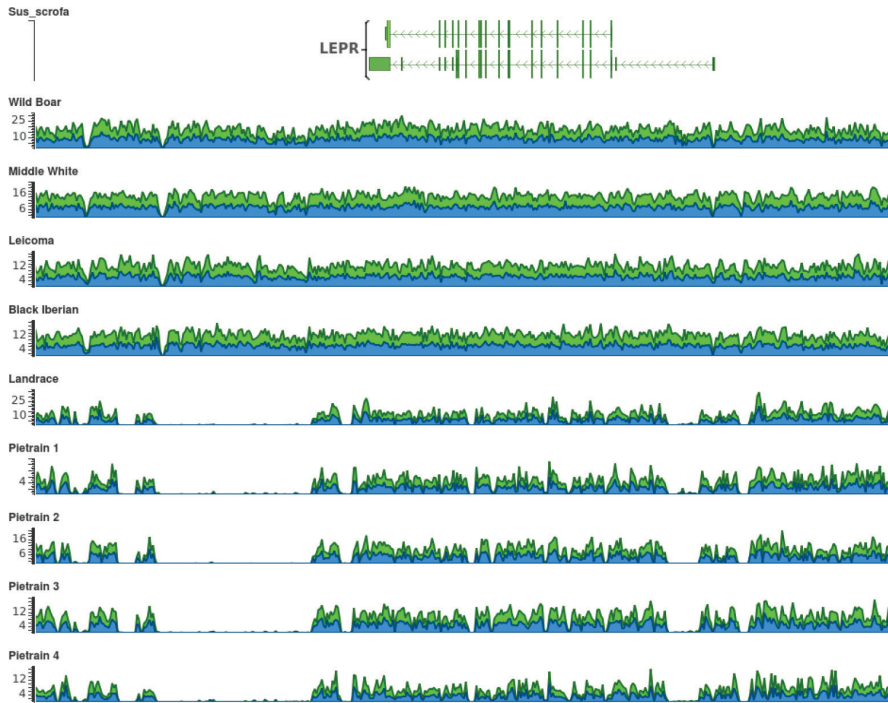


Figure 3. An example of deletion within upstream and downstream regions of *LEPR* locus in five pig genomes. The read coverage is low in the region of deletion. The figure was made using Golden Helix GenomeBrowse.

4.2. Paired-end approaches

Paired-end sequencing data allow detection of many types of SVs. Paired-end (PE) SV calling approaches detect deviations from expected library insert size (donor reads map at inconsistent distances). When a pair of reads does not overlap with any SV, the distance between them is the same as the size of the library insert and reads have correct orientation (concordant pairs). When the read pair overlaps a SV, the mapping distance of paired reads differs from the library insert size and their orientation may be inverted. Discordantly mapped paired-reads can be (1) further apart than expected, (2) closer together than expected, (3) in inverted orientation, (4) in incorrect order, (5) on different chromosomes. Clusters of read pairs aligned to the same genomic regions with the distance shorter than expected can be explained by insertion in the sequenced samples (donor). Larger distances between reads than expected can be explained by deletion in the sample (donor) (**Figure 4**). The resolution of the breakpoints detected by this approach depends on the library's insert size and on the read coverage. Insertions larger than the library insert size cannot be detected.

4.3. Algorithms based on split-reads

Split-read (SR) algorithm can detect SVs with a single base-pair resolution. Split reads contain the breakpoint of the structural variant. Their alignments to the reference genome are split

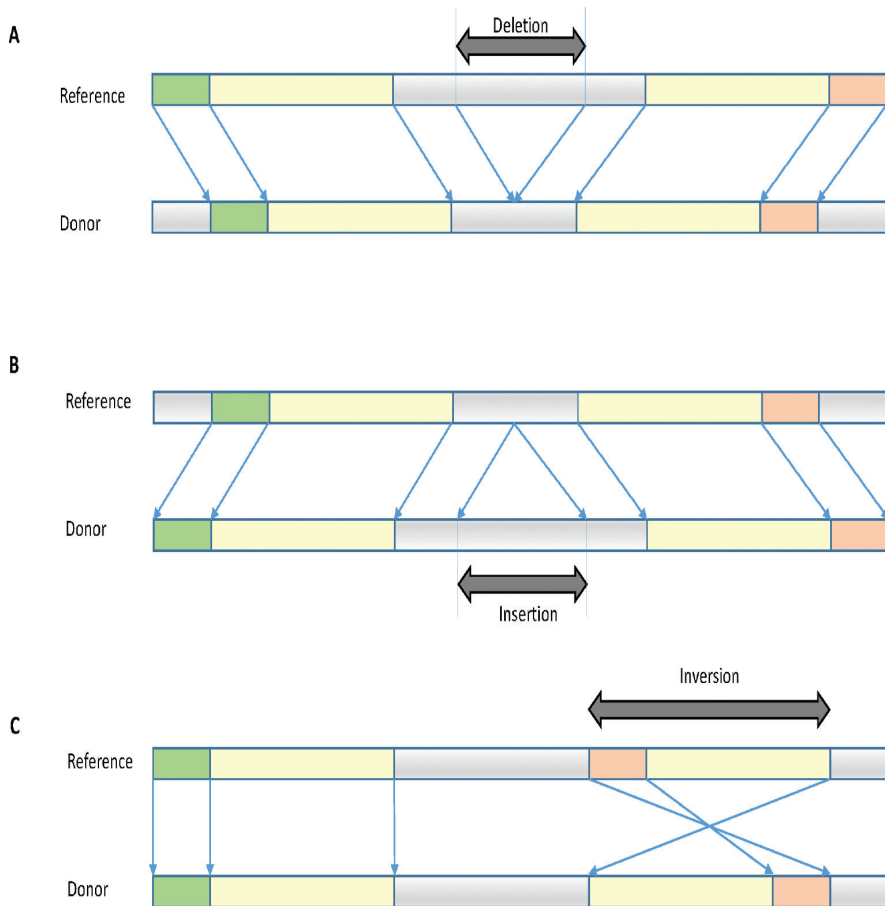


Figure 4. Examples of identification of deletion, insertion, and inversion using paired-end approach: (A) paired-reads are closer together than expected (deletion), (B) paired-reads are further apart than expected (insertion), (C) paired-reads are in inverted orientation (inversion).

into two parts (**Figure 5**). Parts of a read are independently aligned to the reference genome, so the reads should be long enough to be aligned uniquely. Therefore, algorithms based on split-reads are feasible only when the sequencing reads are sufficiently long.

4.4. Algorithms based on de novo assembly

Algorithms based on de novo assembly (AS) are able to detect all forms of structural variation. De novo assembly refers to reassembling the original sequence from which the fragments were sampled. When the sequenced genome is assembled, it is compared to the reference genome to identify SVs. The method enables discovery of novel sequence fragments (insertions). The approach is time-consuming, costly, and prone to assembly errors. In terms of computational efficiency and detection power, targeted SV assembly is more effective. They dissect a problem into a set of local assembly problems that can be more effectively solved.

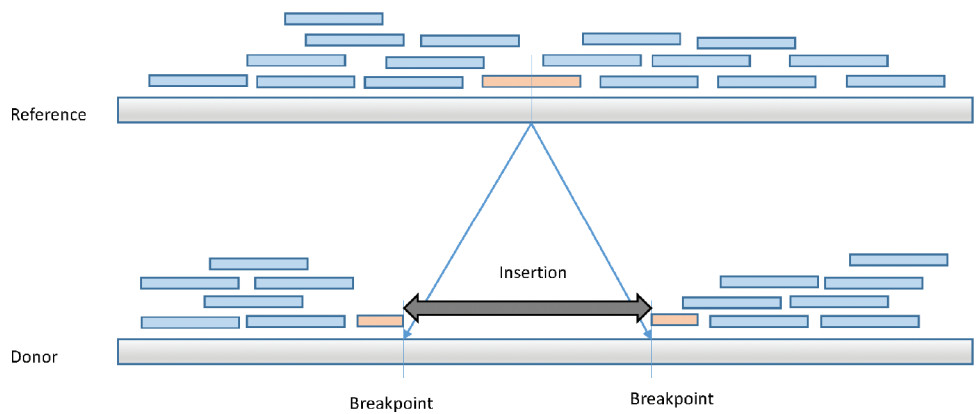


Figure 5. An example of deletion in an individual genome detected by split-read method.

| Tool | SV type | Strategy | Released | Reference |
|------------------|--|------------------------------|----------|-----------|
| PEMer | Indels, inversions | paired-reads | 2009 Feb | [49] |
| VariationHunter | Transposon insertions | paired-reads | 2010 Jun | [50] |
| SegSeq | CNVs | read-depth | 2009 Jan | [51] |
| BreakDancer | Indels, inversions, and translocations | paired-reads | 2009 Jul | [52] |
| Pindel | Breakpoints of large deletions and medium-sized insertions | split-read | 2009 Nov | [53] |
| VariationHunter | Transposon insertions | paired-reads | 2010 Jun | [50] |
| Cortex | simple and complex SVs | de novo assembly | 2011 Apr | [54] |
| CNVnator | CNVs | read-depth | 2011 Jun | [55] |
| GASVPro | Indels, inversions, interchromosomal translocations | read-depth, paired-end | 2012 Mar | [56] |
| SVseq2 | Indels with exact breakpoints | split-read, paired-end | 2012 Apr | [57] |
| Breakpointer | Indels, mobile insertions and non-homologous recombinations | read-depth, split-read, | 2012 Apr | [58] |
| DELLY | Copy-number variable deletions, tandem duplications, inversions, reciprocal translocations | split-read, paired-end | 2012 Sep | [59] |
| SVM ² | Short insertions and deletions | paired-end, machine learning | 2012 Oct | [60] |
| PeSV-Fisher | Deletions, gains, intra- and interchromosomal translocations, and inversions | paired-reads, read-depth | 2013 May | [61] |
| LUMPY | Deletions, inversions, tandem duplications, and interchromosomal translocations | split-read, paired-end | 2014 Jun | [62] |

| Tool | SV type | Strategy | Released | Reference |
|---------|---|---|----------|-----------|
| Gustaf | Deletions, inversions, dispersed duplications and translocations of ≥ 30 bp | split-read | 2014 Dec | [63] |
| MetaSV | Indels, insertions, inversions, translocations, and CNVs | integration of SV callers (BreakSeq, Breakdancer, Pindel, CNVnator), local assembly | 2015 Aug | [64] |
| Manta | Medium-sized indels, large insertions | split-read, paired-end | 2016 Apr | [65] |
| SRBreak | CNV breakpoints | read-depth, split-read | 2016 Sep | [66] |
| Seeksv | Deletion, insertion, inversion and interchromosomal transfer | split-read, paired-end, read-depth fragments with two ends unmapped | 2017 Jan | [67] |
| SVachra | Large insertions-deletions, inversions, inter and intrachromosomal translocations | paired-end | 2017 Oct | [68] |

Table 1. The list of tools for different types of SV calling.

4.5. Hybrid-approaches for SV calling

SV detection algorithms rely on different properties of the underlying data and vary in accuracy and sensitivity. One single method cannot detect complete range of SVs, each is limited to specific type of SVs. Combined approaches can overcome limitations of a single method [44]. Two directions can be taken, combining strategies within one caller or combining SV callers [45]. A class of SV detection methods bases on machine learning. Variations are identified by various methods and are filtered against empirically derived training set data.

4.6. Bioinformatics tools for structural variation calling

A number of algorithms have been proposed to address the issue of structural variants calling from NGS data, but the structural variation calling remains challenging. The complete range of SVs cannot be discovered using one single method. The process of SV calling usually utilizes multiple variant callers to overcome limitations of individual approaches. Knowing advantages and drawbacks of various tools (**Table 1**) is important to make proper decisions when designing NGS data analysis pipelines. Different callers yield lists of identified SVs with limited overlap. Pipelines SVMerge [46], HugeSeq [47], iSVP [48], and IntanSV that integrate different SV callers, such as BreakDancer, CNVnator, SVseq2, Pindel, and DELLY and merge their results were published.

5. Conclusions

Using next-generation sequencing technologies, large volume of sequence data is produced with an unprecedented speed and constantly reducing cost. It allowed rapid progress in

single-nucleotide polymorphisms detection. The awareness that structural variations represent a significant source of genotypic and phenotypic variation is permanently growing. However, the accurate detection of structural variants from NGS data is a daunting task. Relatively short reads, often repetitive character of SV, large amount of data, and large number of benign variants in complex genomes represent a major challenge for bioinformatics analysis of SVs. A number of algorithms have been proposed to address the issue of structural variants calling from NGS data. SV detection algorithms rely on different properties of the underlying data and vary in accuracy and sensitivity. SV detection process usually utilizes multiple variant callers. However, knowing advantages, drawbacks, and properties of different tools is inevitably required for proper decisions when designing NGS data analysis pipelines from publicly available tools. This chapter summarizes basic concepts of bioinformatics analysis of SV and introduces some rules for their assessment.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding P4-0220).

Conflict of interest

We have no conflict of interest to declare.

Author details

Minja Zorc*, Jernej Ogorevc and Peter Dovč

*Address all correspondence to: minja.zorc@bf.uni-lj.si

Biotechnical Faculty, Department of Animal Science, University of Ljubljana, Domzale, Slovenia

References

- [1] Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*. 2001;**409**(6822):928-933
- [2] Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. *Science*. 2004;**305**(5683):525-528
- [3] Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. *Nature Genetics*. 2004;**36**(9):949-951

- [4] Fernandez BA, Roberts W, Chung B, Weksberg R, Meyn S, Szatmari P, et al. Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *Journal of Medical Genetics*. 2010;**47**(3):195-203
- [5] Cho SC, Yim SH, Yoo HK, Kim MY, Jung GY, Shin GW, et al. Copy number variations associated with idiopathic autism identified by whole-genome microarray-based comparative genomic hybridization. *Psychiatric Genetics*. 2009;**19**(4):177-185
- [6] Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, et al. Structural variation of chromosomes in autism spectrum disorder. *American Journal of Human Genetics*. 2008;**82**(2):477-488
- [7] Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, et al. Large recurrent microdeletions associated with schizophrenia. *Nature*. 2008;**455**(7210):232-236
- [8] Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015;**518**(7538):197-206
- [9] Olsson LM, Nerstedt A, Lindqvist AK, Johansson SC, Medstrand P, Olofsson P, et al. Copy number variation of the gene NCF1 is associated with rheumatoid arthritis. *Antioxidants & Redox Signaling*. 2012;**16**(1):71-78
- [10] Molokhia M, Fanciulli M, Petretto E, Patrick AL, McKeigue P, Roberts AL, et al. FCGR3B copy number variation is associated with systemic lupus erythematosus risk in Afro-Caribbeans. *Rheumatology (Oxford, England)*. 2011;**50**(7):1206-1210
- [11] de Cid R, Riveira-Munoz E, Zeeuwen PL, Robarge J, Liao W, Dannhauser EN, et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nature Genetics*. 2009;**41**(2):211-215
- [12] Moon S, Hwang MY, Jang HB, Han S, Kim YJ, Hwang JY, et al. Whole-exome sequencing study reveals common copy number variants in protocadherin genes associated with childhood obesity in Koreans. *International Journal of Obesity*. 2017;**41**(4):660-663
- [13] Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Research*. 2011;**39**(Database issue):D945-D950
- [14] Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics*. 2008;**40**(6):722-729
- [15] Weckselblatt B, Rudd MK. Human structural variation: Mechanisms of chromosome rearrangements. *Trends in Genetics*. 2015;**31**(10):587-599
- [16] Weckselblatt B, Hermetz KE, Rudd MK. Unbalanced translocations arise from diverse mutational mechanisms including chromothripsis. *Genome Research*. 2015;**25**(7):937-947
- [17] Ravnan JB, Tepperberg JH, Papenhausen P, Lamb AN, Hedrick J, Eash D, et al. Subtelomere FISH analysis of 11 688 cases: An evaluation of the frequency and pattern of

- subtelomere rearrangements in individuals with developmental disabilities. *Journal of Medical Genetics*. 2006;**43**(6):478-489
- [18] Shao L, Shaw CA, Lu XY, Sahoo T, Bacino CA, Lalani SR, et al. Identification of chromosome abnormalities in subtelomeric regions by microarray analysis: A study of 5,380 cases. *American Journal of Medical Genetics. Part A*. 2008;**146a**(17):2242-2251
- [19] Ballif BC, Sulpizio SG, Lloyd RM, Minier SL, Theisen A, Bejjani BA, et al. The clinical utility of enhanced subtelomeric coverage in array CGH. *American Journal of Medical Genetics. Part A*. 2007;**143a**(16):1850-1857
- [20] Zhang F, Carvalho CM, Lupski JR. Complex human chromosomal and genomic rearrangements. *Trends in Genetics*. 2009;**25**(7):298-307
- [21] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bembien LA, et al. Genome sequencing in open microfabricated high density Picoliter reactors. *Nature*. 2005;**437**(7057):376-380
- [22] de Ravel TJ, Devriendt K, Fryns JP, Vermeesch JR. What's new in karyotyping? The move towards array comparative genomic hybridisation (CGH). *European Journal of Pediatrics*. 2007;**166**(7):637-643
- [23] Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. *Nature*. 2010;**464**(7289):704-712
- [24] Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annual Review of Medicine*. 2010;**61**:437-455
- [25] Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM, et al. NGSQC: Cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics*. 2010;**11**(Suppl 4):S7
- [26] Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011;**27**(6):863-864
- [27] Lassmann T, Hayashizaki Y, Daub CO. TagDust-A program to eliminate artifacts from next generation sequencing data. *Bioinformatics*. 2009;**25**(21):2839-2840
- [28] Patel RK, Jain M. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS One*. 2012;**7**(2):e30619
- [29] Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nature Methods*. 2012;**9**(4):357-359
- [30] Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*. 2009;**25**(14):1754-1760
- [31] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;**25**(16):2078-2079
- [32] Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;**6**(2):80-92

- [33] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. 2010;**20**(9):1297-1303
- [34] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. 2010;**38**(16):e164
- [35] Hu H, Huff CD, Moore B, Flygare S, Reese MG, Yandell M. VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genetic Epidemiology*. 2013;**37**(6):622-634
- [36] Teer JK, Green ED, Mullikin JC, Biesecker LG. VarSifter: Visualizing and analyzing exome-scale sequence variation data on a desktop computer. *Bioinformatics*. 2012;**28**(4):599-600
- [37] Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Research*. 2012;**40**(7):e53
- [38] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. *Cell*. 2007;**129**(4):823-837
- [39] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*. 2011;**12**(5):363-376
- [40] Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*. 2014;**15**(2):256-278
- [41] Guan P, Sung WK. Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods*. 2016;**102**:36-49
- [42] Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. *Frontiers in Bioengineering and Biotechnology*. 2015;**3**:92
- [43] Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of human copy number variation and multicopy genes. *Science*. 2010;**330**(6004):641-646
- [44] Gao J, Qi F, Guan R, editors. Structural variation discovery with next-generation sequencing. In: 2013 2nd International Symposium on Instrumentation and Measurement, Sensor Network and Automation (IMSNA); December 23-24, 2013. pp. 23-24
- [45] Lin K, Smit S, Bonnema G, Sanchez-Perez G, de Ridder D. Making the difference: Integrating structural variation detection tools. *Briefings in Bioinformatics*. 2015;**16**(5):852-864
- [46] Wong K, Keane TM, Stalker J, Adams DJ. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biology*. 2010;**11**(12):R128
- [47] Lam HYK, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, et al. Detecting and annotating genetic variations using the HugeSeq pipeline. *Nature Biotechnology*. 2012;**30**:226
- [48] Mimori T, Nariiai N, Kojima K, Takahashi M, Ono A, Sato Y, et al. iSVP: An integrated structural variant calling pipeline from high-throughput sequencing data. *BMC Systems Biology*. 2013;**7**(Suppl 6):S8

- [49] Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, et al. PEMer: A computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*. 2009;**10**(2):R23
- [50] Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, et al. Next-generation VariationHunter: Combinatorial algorithms for transposon insertion discovery. *Bioinformatics*. 2010;**26**(12):i350-i3i7
- [51] Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods*. 2009;**6**(1):99-103
- [52] Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*. 2009;**6**:677
- [53] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;**25**(21):2865-2871
- [54] Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*. 2012;**44**:226
- [55] Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*. 2011;**21**(6):974-984
- [56] Sindi SS, Onal S, Peng LC, Wu HT, Raphael BJ. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biology*. 2012;**13**(3):R22
- [57] 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
- [58] Sun R, Love MI, Zemojtel T, Emde AK, Chung HR, Vingron M, et al. Breakpointer: Using local mapping artifacts to support sequence breakpoint discovery from single-end reads. *Bioinformatics*. 2012;**28**(7):1024-1025
- [59] Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;**28**(18):i333-i3i9
- [60] 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 Research*. 2012;**40**(18):e145-e14e
- [61] Escaramis G, Tornador C, Bassaganyas L, Rabionet R, Tubio JM, Martinez-Fundichely A, et al. PeSV-fisher: Identification of somatic and non-somatic structural variants using next generation sequencing data. *PLoS One*. 2013;**8**(5):e63377

- [62] Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*. 2014;**15**(6):R84
- [63] Trappe K, Emde AK, Ehrlich HC, Reinert K. Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. *Bioinformatics*. 2014;**30**(24):3484-3490
- [64] Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, et al. MetaSV: An accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics*. 2015;**31**(16):2741-2744
- [65] Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;**32**(8):1220-1222
- [66] Nguyen HT, Boockock J, Merriman TR, Black MA. SRBreak: A read-depth and split-read framework to identify breakpoints of different events inside simple copy-number variable regions. *Frontiers in Genetics*. 2016;**7**:160
- [67] Liang Y, Qiu K, Liao B, Zhu W, Huang X, Li L, et al. Seeksv: An accurate tool for somatic structural variation and virus integration detection. *Bioinformatics*. 2017;**33**(2):184-191
- [68] Hampton OA, English AC, Wang M, Salerno WJ, Liu Y, Muzny DM, et al. SVachra: A tool to identify genomic structural variation in mate pair sequencing data containing inward and outward facing reads. *BMC Genomics*. 2017;**18**(Suppl 6):691

