Computational Methods for Detecting Large-Scale Structural Rearrangements in Chromosomes

Muneeba Jilani[1] • Nurit Haspel[2]

[1]Computational Sciences PhD program, University of Massachusetts Boston, Boston, MA, USA; [2]Department of Computer Science, University of Massachusetts Boston, Boston, MA, USA

Author for Correspondence: Nurit Haspel, Department of Computer Science, University of Massachusetts Boston, Boston, MA, USA. Email: Nurit.haspel@umb.edu

Doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch3

Abstract: Large-scale structural chromosomal rearrangements or structural variants, such as insertions, deletions, translocations, and inversions may result in the exchange of coding or regulatory DNA/RNA sequences between genes, which can lead to gene fusion or the loss/gain of genetic material. Gene fusion events are common in multiple types of cancer. High-throughput DNA and RNA sequencing methods produce large amounts of genomic data. Due to the massive amounts of data and the fact that structural variants account for just a small fraction of the data, efficient and accurate search methods are required for the detection of chromosomal breakpoints and structural variations. Robust identification of structural variants remains paramount for accurate inference of long-range interactions from high-throughput chromosome conformation capture (Hi-C) data. This chapter is a survey of computational methods based on paired end reading, efficient search techniques and parallel computing to detect structural variants in both whole genome and transcriptome sequences, as well as Hi-C data.

Keywords: gene fusion; Hi-C; next generation sequencing; RNA-Seq; structural variants

In: *Bioinformatics*. Nakaya HI (Editor). Exon Publications, Brisbane, Australia.
ISBN: 978-0-6450017-1-6; Doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021

INTRODUCTION

Structural variation (SV) is a disparity in the chromosomal structure of an organism (1). SVs can be small or large, and include deletions, insertions, inversions, translocations, single nucleotide variations, and copy number variations (Figure 1). The cause of such variations is normally a break in the DNA at two different locations. The broken ends are rejoined, resulting in a novel chromosomal rearrangement. The resulting form is different from the original gene order of the chromosome. While some SVs are responsible for the diversity of phenotypes and disease, others do not have obvious effects. When one nucleotide in a DNA sequence is changed, the resulting SV is called a single nucleotide variant (SNV). This is the most common type of variation (2). When an SNV occurs in a coding region, the impact is determined by whether it is a missense or synonymous mutation. In the case of non-coding regions, the effect it may have depends on its impact on gene regulation. When a fragment of a chromosome detaches and rejoins to a different, nonhomologous chromosome, the phenomenon is called translocation. Double strand breaks of DNA at two loci followed by mismatched end joining is the main method by which translocations occur. The impacts of translocations range from mild to

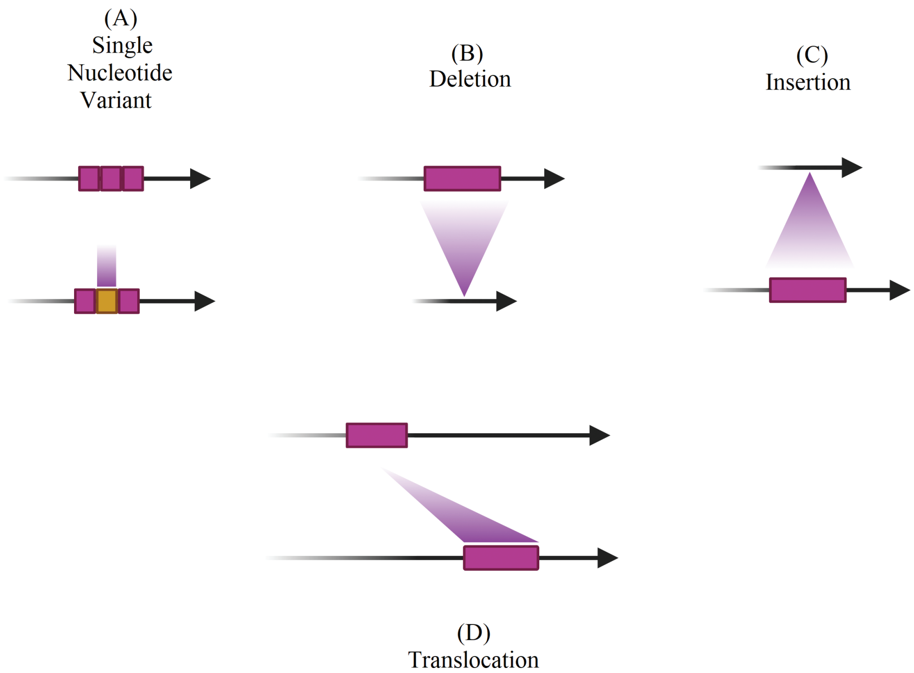

Figure 1. Different types of structural variations. A. Single nucleotide variant. **B.** Deletion. **C.** Insertion. **D.** Translocation.

catastrophic depending on the location of the affected genes in comparison to regulatory sequences. When two different nonhomologous chromosomes trade fragments in a manner that does not result in the addition or deletion of genetic information, it is called a balanced translocation. A translocation can potentially impact the phenotype due to closeness of a proto-oncogene to cis regulatory elements. Balanced rearrangements occur in approximately 1/500 to 1/625 individuals in a population. Insertions and deletions are referred to as copy number variations (CNVs). When segments share >90% of the repeated nucleotide sequences ("copies") with each other and are >1 kB in length, then it is called a segmental duplication (SD). CNVs often overlap with SDs. If a CNV is present in >1% of the population, then it is called a copy number polymorphism (CNP).

To diagnose and treat genomic disease, effective methods are required to understand the genome (2). One problem is the sporadic nature of genomic disorders. The rearrangements are novel in most scenarios. The mutation rate for a specific locus is more frequent in cases of genomic rearrangements than it is for the point mutations. Generally, genomic disorders occur with similar frequency worldwide. However, there are significant differences in incidences among populations. In some cases, population-specific SVs of the genomic region of the patients' parents have been found, showing that variation of genomic architecture can be a significant factor in disease susceptibility. Gene fusions result from SVs that occur in regulatory or coding regions, which can result in cancer (3). Gene fusions can result in irregular function or abnormal transcription of cancer driver genes. An example could be the development of chimeric transcripts, combining exons of two different genes. Cancer subtypes can be defined by these gene fusions, thus making these structural aberrations a vital class of targets for therapy. Although they do not qualify as structural variations but hold a middle ground in consequential hierarchy, we explore the topic due to its proximity to the topic at hand and its high significance.

The methods used to approach the problem of SV are broadly classified into two categories: array-based methods and sequencing-based computational methods. Microarray methods are frequently used for the detection of deletions or insertions. One problem with microarray methods is that they are incapable of detecting balanced variations. They also fail to provide the exact location of the variation. G-band karyotyping (4) is also popular. The disadvantages of this method include low throughput and low resolution. Furthermore, it is incapable of characterizing expansively rearranged genomes. Polymerase chain reaction (5) and fluorescence in-situ hybridization are also commonly used. All these techniques require pre-existing knowledge of the variation and thus novel rearrangements are not detectable. Recently, high-throughput sequencing based methods such as RNA-sequencing (RNA-Seq) and whole genome sequencing (WGS) have emerged as an effective method for SV identification. They can identify gene fusions and genomic rearrangements with high resolution; however, they have shortcomings. These short reads-based approaches cannot effectively detect SVs in repetitive regions of the genome and are limited in their ability to resolve haplotype-resolved complex SVs. There are methods that combine various techniques in an attempt to achieve efficiency goals (6).

There is also a lack of computational methods that systematically detect structural chromosomal aberrations by virtue of the genomic location of copy number alteration (CNA)-associated chromosomal breaks and identify genes that appear to be affected in a non-random way by chromosomal breakpoints across large series of tumor samples. Next generation sequencing (NGS) technology is considered one of the most recent advanced technologies in biomedical research, and it has opened more opportunities for scientific discovery of genetic information. It is particularly useful in the analysis of CNAs in the DNA.

When it comes to detecting balanced rearrangements, a drawback of using NGS based methods is the substantial sequencing depth that is required to identify false positives due to sequencing errors, and the cost associated with it. Current methods perform up to a depth of 40x (7). Despite that, the detection is hampered at repeated regions due to reduced mapability, which means certain regions (for example, heterochromatic or highly homologous regions) are often hard to map accurately. This disadvantage occurs due to the fact that many repeated rearrangements are arbitrated by fusion between homologous sequences or duplications between segments, and therefore will have one or several breakpoint mappings within repeated segments (8). RNA-Seq is primarily used to reveal the presence and quantity of RNA in a sample. The advantage of RNA-Seq over DNA sequencing is a reduction in the huge amount of data produced by NGS, as well as better coverage and higher resolution of the dynamic nature of transcriptome compared to previously used array-based methods. Using transcriptional information specifically, RNA-Seq makes it possible to focus on gene fusion, post-transcriptional modifications, single nucleotide polymorphisms (SNPs), modified gene expression over time, and mutations.

Several methods for the detection of rearrangements in chromosomes have been developed. These methods are based on the variations in the 3-D organization of the nucleus (9). 3C libraries, sequenced with low coverage, are created for this purpose. These libraries aid in the detection of variations in spatial contacts that are associated with the prevalent rearrangement. To detect translocations, high-throughput chromosome conformation capture (Hi-C)-based methods are considered the most sensitive, as intrachromosomal contact frequency is higher than inter-chromosomal contacts, resulting in a massive increase in the frequency of spatial contacts amongst translocated regions in the rearranged genome. The following sections discuss the different methods, and their performance and capabilities are analyzed in depth in subsequent sections. These methods are compared on various bases and final thoughts are presented in the conclusion.

NGS-BASED METHODS

NGS is a high throughput DNA sequencing technology that uses parallel sequencing of multiple small DNA fragments. A vast number of short reads (typically 50-150 bp) are sequenced in a single stroke. The sequence information is compared to a human genome reference sequence to identify any structural variations or mutations in the targeted sequences (10,11). Many NGS methods use paired-end reading, where two paired reads are generated at an approximately known distance in the tested genomes (usually around 500 bp). The reads are aligned to the reference genome. Pairs whose mapping distance is substantially different from the expected length, or that map in two different chromosomes, or that present with an anomalous orientation - denoted *discordant reads* - suggest potential large-scale structural variants (12). Figure 2 shows an illustration of discordant reads where the reads map into two different chromosomes. Most methods in this category use search techniques in an attempt to detect discordant reads in order to find possible chromosomal breakpoint locations. This is a big data challenge, since the number of reads is enormous, and the number of possible discordant reads is very small. Additionally, often the breakpoint itself is not known *a priori*, so the entire genome must be searched. Also, not all discordant reads necessarily indicate a breakpoint, since the reads may be prone to errors or individual differences between different genomes. A brief introduction to popular and well-known methods in this category are discussed below. Some methods are used for detecting SVs and some use a combination of methods for this purpose and then perform a comparison of the results.

DELLY (13) is an SV detection method that integrates short insert paired ends, long-range mate-pairs and split-read alignments. The method aims to accurately detect rearrangements at single-nucleotide resolution. DELLY is capable of SV detection from the 1,000 Genomes project as well as cancer genomes on real data. On simulated data, it is comparable to other prediction methods discussed in this section.

SVs are quite diverse in nature and generate various types of signals of alignment. Where most algorithms function to utilize one signal for detection and another for confidence, LUMPY (14) provides a more sensitive detection of SVs. This is particularly useful in lower coverage datasets or disparate cancer samples. The specificity increases by use of multiple signals, which include signals from read alignments or preceding evidence. The method can also incorporate additional sources of information that may be available with future technological advances.

Predicting the location of breakpoints remains a challenging problem. Utilizing paired-end reads from NGS data, TIDE (15) is a multiphase method that performs preprocessing, search, collection of the results and refinement in order to find a

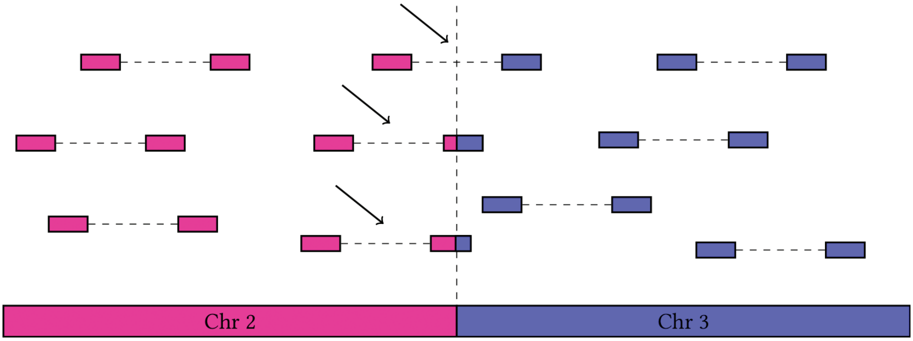

Figure 2. Discordant reads around a chromosomal translocation. The pink reads align to chromosome 2, and the blue align to chromosome 3. Paired reads where the mates align to different chromosomes (discordant reads; arrows) may indicate a chromosomal breakpoint.

set of reads that denote possible breakpoints that indicate inter-chromosomal translocations or insertions among a vast number of candidates. To reduce the number of false positives, the refinement stage uses BLAST. The number of false negatives can be controlled by the input parameters in the SketchSort algorithm. This method identifies reads containing breakpoints which allows for verification analysis of region surrounding mutation.

NGS poses a computational problem due to huge search space, as it splits the genome into small, overlapping fragments. A major challenge is to efficiently identify small numbers of breakpoints. TDJD (16) identifies the location of intra-chromosomal breakpoints corresponding to large scale structural variations. Candidate reads are split into windows that are represented as sequences of indexed arrays called *binary fingerprints*. The goal of the binary fingerprints is to reduce the dimensionality of the data. The third phase constitutes finding potential locations of chromosomal breakpoints. They use Jaccard distance to solve the exact nearest neighbor problem. What sets this apart from peer methods is the usage of multi-threaded algorithm SSE instructions to reach high performance.

To identify change points in sequential data, specifically copy number changes from various types of genomic data, a very popular algorithm is circular binary segmentation (5). ‘GeneBreak’ (17) uses a genome-wide approach to systematically identify genes recurrently affected by the genomic location of chromosomal CNA-associated breaks, which can be applied to DNA copy number data obtained by array-comparative genomic hybridization or by (low-pass) WGS. First, the method gathers the genomic locations of chromosomal CNA-associated breaks previously identified using a segmentation algorithm to obtain CNA profiles. Next, the algorithm implements an annotation approach for breakpoint-to-gene mapping. Finally, dedicated cohort-based statistics is used to correct for covariates that may influence the probability of finding a breakpoint. Additional testing correction is integrated to detect recurrent breakpoints. The software package is implemented in the R package ‘GeneBreak’, which extracts additional information from CNA data. The package provides an algorithm which handles the identification of possible breakpoints and their mapping to genes as well as providing a comprehensive statistical analysis to detect recurrent breakpoint genes from series of tumor samples. For these reasons, the method can be applied to detect CNA-associated chromosomal breaks in individual tumor samples, and it facilitates the detection of recurrent breakpoint genes across multiple tumor samples.

Chen *et al*. (18) have developed an algorithm for detecting somatic CNA using WGS data. The algorithm, CONSERTING, finds copy number segmentation by regression tree in NGS. The method performs iterative analysis of the segmentation based on changes in read depth and the detection of localized structural variations. The authors revealed novel oncogenic CNAs, complex rearrangements and subclonal CNAs by analyzing 43 cancer genomes from both pediatric and adult patients. These rearrangements were missed by alternative approaches.

High density SNP microarrays are useful to measure DNA copy number variations across the genome. Liu *et al*. (19) used SNP array data of cancer cell lines and patient samples to evaluate the CNV and copy number breakpoints for several known fusion genes implicated in tumorigenesis. Their results demonstrate the usefulness of SNP array data for predicting genetic aberrations via translocations, based on identifying copy number breakpoints within the target genes. The authors performed genome-wide analysis to identify genes that have copy number breakpoints across 820 cancer cell lines. They identified candidate oncogenes that were linked to potential translocations in specific cancer cell lines.

It is possible to analyze the NGS data for the detection of boundaries of CNV regions on a chromosome or a genome by phrasing the problem as a statistical change point detection problem presented in the read count data. Ji *et al*. (20) developed a statistical change point model to help detect CNVs using NGS read count data. The authors used a Bayesian approach to incorporate possible parameter changes in the underlying NGS read count data distribution, and derived posterior probabilities for the change point inferences. They were able to detect CNV regions in a publicly available lung cancer cell line NGS dataset. RAPTR-SV (21) is a method for SV detection that is comparable with NGS based methods such as DELLY and LUMPY. RAPTR-SV is proven to be superior for tandem duplications by recognizing twice as many duplications as DELLY. It combines paired-end and split-end algorithms. This method is available publicly with instructions for use along with test results.

Korbel *et al*. (22) have developed Paired-End Mapper (PEMer). This method is an analysis pipeline, compatible with several NGS platforms. They incorporate simulation-based error models, yielding confidence values for each SV, and a back-end database. Their simulations demonstrated their high efficiency in reconstructing structural variants for the method's coverage-adjusted multi-cutoff scoring strategy. They also showed that the method is relatively insensitive to base-calling errors.

Hayes *et al*. (23) argue that methods that use paired-ends reads are not accurate in predicting breakpoints with precision. Their method, Bellerophon, aims to resolve the issue of identification of translocations by a hybrid method that also uses "soft-clipped" reads. It also classifies translocations as balanced/unbalanced or an insertion. Compared to peer methods, Bellerophon has better accuracy in prediction and superior specificity on real cancer data, while it is similar to others in case of simulated data.

FastGT (24) is a novel method that computes SNVs by counting the frequencies of unique k-mers. Their k-mer database allows the simultaneous genotyping of 30 million SNVs, including >23,000 SNVs from the Y chromosome. It is based on counting known unique k-mers from NGS data and directly performs genotyping. Thus, it is especially suited for fast, preliminary analysis of a subset of markers before a full-scale analysis is performed.

MICADo (25) is a graph-based method that is able to distinguish patient-specific mutations from other variations. It functions by performing analysis of NGS reads for all the samples within the data context of the whole cohort. It captures the difference between a specific sample and the cohort. This technique is suitable for highly heterogeneous samples.

Another method, Churchill (26), is capable of fully automating the analytical process that takes raw sequence data through the intensive computational process of alignment and post-alignment genotyping and processing, and produces a list of SVs that can be utilized for clinical interpretation and analysis.

Synthetic long reads generated by linked-read sequencing are useful for SV detection and their analysis. Long Ranger, which wraps standard short-read variant callers to generate SNP and small indel calls, generates the essential output files necessary for downstream analyses. However, to perform downstream analysis, oftentimes users need to customize their own tools. Gemtools (27) encompasses a set of tools that provides the user with the necessary flexibility to perform basic functions on their linked-read sequencing output.

Another method combines data generated by front line methods. The method by Mimori *et al*. (28) is capable of detecting full range of SVs. Integrated structural variant calling pipeline (iSVP) incorporates existing methods for SV detection by using newly designed filtering and merging processes. In their experiment (29), large numbers of deletions were detected that included prominent peaks around 300 bp and 6,000 bp paralleled with long inspected nuclear elements. The SVs detected were consistent with those validated in other studies. FusorSV (29) uses a data mining approach in order to evaluate the performance of and merge callsets from a group of SV-calling algorithms. A simple union of SV detection methods can result in a huge number of false-positives, and therefore this novel research includes Structural Variation Engine (SVE) that includes eight popular SV detection methods and a novel algorithm, FusorSV, to merge the calls from the included methods. The best results are returned based on a performance threshold.

In summary, sophisticated algorithms are crucial to accurately detect copy number variations and breakpoints from NGS data. Although NGS technology is still emerging and typically applied to cancer studies, there is already a significant number of somatic SV and CNV detection methods for NGS data (5). Some methods are more sensitive in nature and perform well with low coverage data. As understanding in this field improves, more methods are being developed in both CNV and SV category.

RNA-SEQ BASED METHODS

RNA-Seq is a sequencing method that uses NGS to identify the presence and quantity of RNA in a biological sample at a given moment. RNA-seq allows for focus on specific events such as alternative splicing, post-transcriptional modifications, gene fusion, mutations and differential gene expression. RNA-seq results in short reads, similar to DNA sequencing described above, but only transcriptome data is given. Various algorithms use RNA-Seq data for the discovery of fusion genes. Kumar *et al.* (30) compared and evaluated the performance of 12 methods on the bases of false discovery rate, time of computation and memory used. Their results indicate that the performance of such tools is heavily dependent on the RNA-Seq data's number of reads, read length and quality. Some popular algorithms include Bellerophontes (31), BreakFusion (32), Chimerascan (33), nFuse (34), FusionCatcher (35), SOAPfuse (36) and TopHat-fusion (37). The surveyed methods were categorized into three broad groups of paired-end and fragmentation, whole paired-end and direct fragmentation. Finally, the TOPSIS method (Technique for Order of Preference by Similarity to Ideal Solution) (38) was performed on the mixed dataset results in order to rank the fusion detection methods. The paper (30) has a thorough comparison criteria and collection of reviewed methods. However, new methods have since been introduced.

Another review paper (39) benchmarks 23 fusion detection methods. The review deems STAR-Fusion (40), STAR-SEQR (41) and Arriba (42) as the best in terms of accuracy and speed. STAR-Fusion is built upon a previous technique by Stransky *et al.* (43). This publicly available tool predicts fusions with the speed of their previous technique. It performs high speed mapping of fusions to the respective reference transcript structure annotations and then screens probable artefacts to report precise fusion prediction, making use of real and simulated data. It is an efficient tool and is estimated to be faster than its counterparts. STAR-SEQR uses chimeric transcripts produced by STAR aligner (44) to detect fusions. Arriba utilizes chimeric alignments detected also by STAR aligner to get gene fusions. It applies numerous filters to the data to extract the artefacts that are present in the RNA-seq data.

The abovementioned Spliced Transcripts Alignment to a Reference or STAR (44) is quoted as the ultrafast universal RNA-seq alignment method based on the sequential maximum mappable seed search in suffix arrays that are not compressed. It later performs seed clustering and stitching procedure. This is the method of choice as it outperforms others by a factor of >50 in the speed of mapping and aligns the human genome 550 million 276 bp paired-end reads per hour on a 12-core server. It also improves alignment precision and sensitivity. STAR can discover chimeric fusion transcripts. Parallel threads are run on multicore systems with nearly linear scaling to the number of cores. Another important feature of STAR is the alignment in a continuous streaming mode, making it compatible with new sequencing technologies.

A new method that has not been covered by the previous reviews is CICERO (CI-CERO Is Clipping Extended for RNA Optimization) (45). It was primarily designed for detection of driver fusions beyond the level of recognized exon-to-exon chimeric transcripts. This is a local assembly-based algorithm that functions by integrating RNA-Seq constructed read support and thorough explanation of candidate ranking. CICERO is claimed to achieve 95% detection rate for separately verified driver fusions of different types.

Many algorithms function by aligning the sequencing reads to the reference transcriptome. Due to the perturbed nature of the cancer genome, various fusions can remain undetected. A recent and noteworthy development is ChimeRScope (46) which aims to resolve this problem by using k-mer based algorithm to accurately predict fusion transcripts from the data analysis pipeline of RNA-seq. When this novel technique is compared to SOAPFuse (36), FusionCatcher (35) and MapSplice (47) alongside other popular fusion detection methods, ChimeRScope performs better irrespective of read length, sequencing depth and expression levels of the fusion transcripts.

All things considered, as revealed by previous analysis, the methods for detecting fusions in RNA-Seq data all perform well, with each having its own advantages. Where the goal of some methods to achieve higher accuracy, methods such as STAR-Fusion, Arriba and STAR-SEQR aim for higher efficiency. Interestingly, the top ranked methods all leverage the data produced by STAR aligner. This area is budding, and new methods are being developed owing to fusion implications in cancer.

Another topic we review is the testing of the algorithms that detect fusion. One of the main difficulties in testing fusion detection algorithms is the lack of sufficiently validated fusion genes that can serve as positive controls to accurately assess performance. One way of testing the accuracy of fusion detection algorithm is simulated reads. However, these tools are limited in simulating the complexity of real samples and cannot account for all biases and errors found in real world datasets. While those tools can be of use, they often cannot correctly assess true performance on real world samples. Artifuse (48) aims to overcome this problem by simulating fusion genes by sequence modification to the genomic reference. For this reason, it can be applied to any RNA-seq dataset without the need for simulated reads. The technique is demonstrated on eight RNA-seq datasets for three fusion gene prediction tools. Overall, the performance assessed from Artifuse is lower compared to previously reported estimates on simulated reads.

HI-C BASED TECHNIQUES

Introduced by van Berkum *et al.* (49), Hi-C (Figure 3) is a technique based on 3C and uses high-throughput sequencing to discover genome-wide interactions in an unbiased manner. It starts by fixing the cells in formaldehyde, which results in the formation of covalent DNA-protein cross-links amongst the loci. The next steps are fragmentation of DNA and ligation, leading to ligation events amid DNA fragments that are cross-linked. These ligation products are marked with biotin. Later, high-throughput sequencing is performed to produce a catalog of interacting fragments. This data in the form of a contact matrix is useful for correlation and ensemble analysis.

Despite not being as sensitive as WGS data, research has proven that Hi-C data is effective for detection of translocations and CNV in spite of not providing as even a coverage (9). It can be used to complement WGS data for detecting translocations. Usage of Hi-C to characterize the organization of the genome has

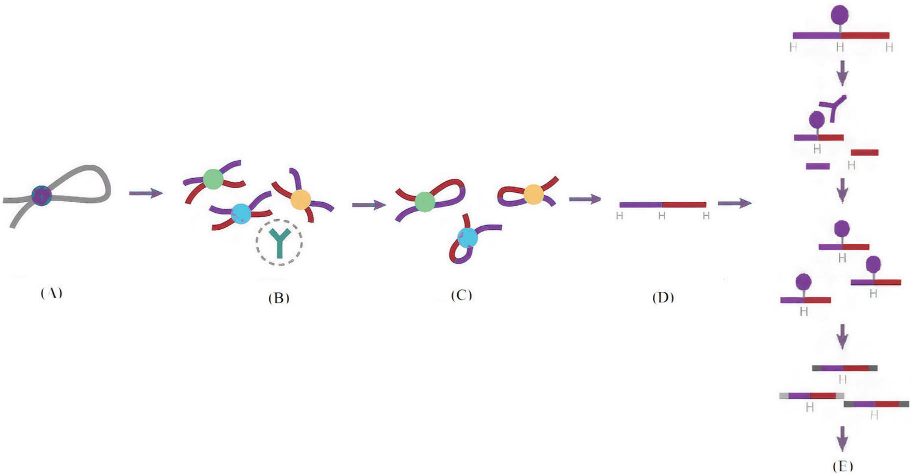

Figure 3. The Hi-C process. A. Crosslinking. **B.** Ligation. **C.** Digest crosslink chromatin. **D.** Reverse crosslink. **E.** Sequencing.

increased the knowledge of biological processes such as tumor development and progression, as well as the dynamics of cell-cycle despite it not being as sensitive as WGS data. Hi-C assays provide knowledge regarding genome-wide identification of chromatin interactions thus enabling the study of the 3D architecture of the genome and the role of gene regulation. It is also used for the characterization of topologically associating domains (TADs) which can be described as genomic regions where the frequency of DNA sequence interaction is more than outside the TAD.

Mozheiko *et al.* (9) argue that enriched 3C libraries provide more knowledge regarding detection of variants in exons than whole genome libraries. They used enriched 3C libraries sequencing for detecting intrachromosomal translocations and point mutations. These findings are invaluable for the detection of promoter-enhancer interaction. Using a relatively small depth of sequencing, the equivalent of 18 million paired readings, makes it possible to attain an average minimum coverage required to look for point mutations. Obviously, as sequencing depth increases, more coverage is achieved. This method is more expensive compared to whole genome sequencing when looking for exon variants, but cheaper with increasing depth.

HiNT (50) (Hi Number variation and Translocation detection) is another method designed for detection of copy number variations and discovery of inter-chromosomal translocations. The resolution of breakpoint detection within Hi-C data is a single base-pair. To perform a comparison with WGS methods, Wang *et al*. have applied the method to both simulated and real data (50). Not only was their false discovery rate lower using HiNT, but the single base-pair resolution is better than the pre-existing Hi-C based methodologies. Multiple input formats are supported and competent storage formats for interaction matrices are utilized. HiNT has three main components: HiNT-PRE preprocesses the Hi-C data and computes the contact matrix, which includes the contact frequencies between any two genomic loci. HiNT-CNV and HiNT-TL receive the Hi-C contact matrix as input and output a prediction of the copy number segments and inter-chromosomal translocations, respectively.

Despite the fact that Hi-C has been used for validating known rearrangements, there has been a shortage of computational methods that can distinguish true rearrangement signals from the inherent biases of Hi-C data and from the actual 3D conformation of chromatin and can precisely detect rearrangement locations *de novo*. Chakraborty *et al*. (51) developed a new set of algorithms to detect novel rearrangements from Hi-C data. These methods have their bases in how intra- and inter-chromosomal Hi-C contacts are distributed for rearranged chromosomes compared to normal chromosomes. The algorithm HiCNV is aimed at sections of genome corresponding to regions that contain CNVs including deletions and amplifications, and their algorithm, called HiCTrans, is aimed at recognizing translocations from Hi-C data.

Transposable elements (TEs), which comprise 44% percent of the genome, are a type of repeated DNA sequence scattered throughout the genome. HiTea (52), which is a short form for Hi-C based transposable element analyzer, is a method designed for the detection of TEs. Clipped Hi-C reads are capitalized in this method for the detection of insertions. A mistake to be avoided here is calling of canonical hi-C interactions as TEs. HiTea avoids this by filtering the candidates to insertion whose breakpoints, that are predicted, are on either the reference genome or TE-consensus are within 3-bases (which are user defined) of the motif of ligation. It furthers this filtration process of tentative candidates by omitting a prospect when the putative breakpoint has multiple breakpoints predicted around it as this is an indicator of complex variant. Another test performed by HiTea is that a true breakpoint should demonstrate a joint cluster of the sequences that are clipped when the mapping is performed with the TE-consensus. The insertions that do not make it through this filtration process are deemed invalid and thus removed. Simple repeat expansions are identified as insertions where clipped reads map to the PolyA sequences only and are thus removed. HiTea proceeds to identify strand information, an estimate of the magnitude of the insertion and target side duplication. HiTea is successfully applied to human genome samples and is comparable to WGS methods. Hi-C based computational methods are fairly new. The techniques explored here use DNA-ligation and high-throughput sequencing in combination with one another in order to assess spatial vicinity of pairs of loci in the genome. There is room for research and improvement in this area (53).

CONCLUSION

Large scale structural variations (SVs) result in dysregulated gene expression or fusion proteins, which may have serious effects. Gene fusions are found in all types of human cancers. Advances in computational methods to recognize SVs led to the identification of various therapeutic targets. These methods take advantage of the vast amount of molecular data available through next generation sequence DNA data, RNA-Seq or Hi-C, coupled with efficient search techniques to detect chromosomal breakpoints and gene fusion events. Despite tremendous advances over the past few years, the problem of detecting large scale SVs remains challenging due to the vast amount of data produced by sequencing techniques, the heterogeneity of cancer mutations, and the limitations of sequencing technologies. Nevertheless, we expect new methods to be developed, based on improved sequencing and search techniques to tackle these challenges and provide better detection of SV and gene fusion.

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship and/or publication of this chapter.

Copyright and permission statement: The authors confirm that the materials included in this chapter do not violate copyright laws. Where relevant, appropriate permissions have been obtained from the original copyright holder(s), and all original sources have been appropriately acknowledged or referenced.

REFERENCES

1. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7(2):85–97. https://doi.org/10.1038/nrg1767
2. Stankiewicz Pawełand Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437–55. https://doi.org/10.1146/annurev-med-100708-204735

3. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nat Rev Cancer. 2008;8(7):497–511. https://doi.org/10.1038/nrc2402
4. O'Connor C. Karyotyping for Chromosomal Abnormalities. Nat Educ. 2008;1(1):27. Available from: http://www.nature.com/scitable/topicpage/karyotyping-for-chromosomal-abnormalities-298 [Accessed on 15 Jan 2021]
5. Hultén MA, Dhanjal S, Pertl B, others. Rapid and simple prenatal diagnosis of common chromosome disorders: advantages and disadvantages of the molecular methods FISH and QF-PCR. REPRODUCTION-CAMBRIDGE-. 2003;126(3):279–97. https://doi.org/10.1530/rep.0.1260279
6. Dixon JR, Xu J, Dileep V, Zhan Y, Song F, Le VT, et al. Integrative detection and analysis of structural variation in cancer genomes. Nat Genet. 2018;50(10):1388–98. https://doi.org/10.1038/s41588-018-0195-8
7. Giannoukos G, Ciulla D, Maeder M, Gloskowski S, Skor M, Dhanapal V, et al. UDiTaSTM: A streamlined genome editing detection method for on-and off-target edits, large deletions, and translocations. In: Molecular Therapy. 2017. p. 296–7.
8. Harewood L, Kishore K, Eldridge MD, Wingett S, Pearson D, Schoenfelder S, et al. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. Genome Biol. 2017;18(1):1–11. https://doi.org/10.1186/s13059-017-1253-8
9. Mozheiko EA, Fishman VS. Detection of Point Mutations and Chromosomal Translocations Based on Massive Parallel Sequencing of Enriched 3C Libraries. Russ J Genet. 2019;55(10):1273–81. https://doi.org/10.1134/S1022795419100089
10. Mu W, Li B, Wu S, Chen J, Sain D, Xu D, et al. Detection of structural variation using target captured next-generation sequencing data for genetic diagnostic testing. Genet Med. 2019;21(7):1603–10. https://doi.org/10.1038/s41436-018-0397-6
11. Qin D. Next-generation sequencing and its clinical application. Cancer Biol Med. 2019;16(1):4–10. Available from: https://pubmed.ncbi.nlm.nih.gov/31119042
12. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 2009;6(11):S13–20.https://doi.org/10.1038/nmeth.1374
13. 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–i339. https://doi.org/10.1093/bioinformatics/bts378
14. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6):R84. https://doi.org/10.1186/gb-2014-15-6-r84
15. Vetro R, Farhoodi R, Kotla R, Haspel N, Weisman D, Rosen J, et al. TIDE: Inter-chromosomal translocation and insertion detection using embeddings. In: 2014 IEEE International Conference on Big Data (Big Data). 2014. p. 64–70. https://doi.org/10.1109/BigData.2014.7004395
16. Mohebbi H, Vajdi A, Haspel N, Simovici D. Detecting chromosomal structural variation using jaccard distance and parallel architecture. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2017. p. 1959–64. https://doi.org/10.1109/BIBM.2017.8217962
17. van den Broek E, van Lieshout S, Rausch C, Ylstra B, van de Wiel MA, Meijer GA, et al. GeneBreak: detection of recurrent DNA copy number aberration-associated chromosomal breakpoints within genes. F1000Research. 2016;5. https://doi.org/10.12688/f1000research.9259.1
18. Chen X, Gupta P, Wang J, Nakitandwe J, Roberts K, Dalton JD, et al. CONSERTING: integrating copy-number analysis with structural-variation detection. Nat Methods. 2015;12(6):527–30. https://doi.org/10.1038/nmeth.3394
19. Liu H, Zilberstein A, Pannier P, Fleche F, Arendt C, Lengauer C, et al. Evaluating translocation gene fusions by SNP array data. Cancer Inform. 2012;11:CIN–S8026. https://doi.org/10.4137/CIN.S8026
20. Ji T, Chen J. Modeling the next generation sequencing read count data for DNA copy number variant study. Stat Appl Genet Mol Biol. 2015;14(4):361–74. https://doi.org/10.1515/sagmb-2014-0054
21. Bickhart DM, Hutchison JL, Xu L, Schnabel RD, Taylor JF, Reecy JM, et al. RAPTR-SV: a hybrid method for the detection of structural variants. Bioinformatics. 2015;31(13):2084–90. https://doi.org/10.1093/bioinformatics/btv086
22. 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 Biol. 2009;10(2):R23. https://doi.org/10.1186/gb-2009-10-2-r23

23. Hayes M, Li J. Bellerophon: a hybrid method for detecting interchromosomal rearrangements at base pair resolution using next-generation sequencing data. In: BMC Bioinformatics. 2013. p. S6. https://doi.org/10.1186/1471-2105-14-S5-S6
24. Pajuste F-D, Kaplinski L, Möls M, Puurand T, Lepamets M, Remm M. FastGT: an alignment-free method for calling common SNVs directly from raw sequencing reads. Sci Rep. 2017;7(1):1–10. https://doi.org/10.1038/s41598-017-02487-5
25. Rudewicz J, Soueidan H, Uricaru R, Bonnefoi H, Iggo R, Bergh J, et al. MICADo--looking for mutations in targeted PacBio cancer data: an alignment-free method. Front Genet. 2016;7:214. https://doi.org/10.3389/fgene.2016.00214
26. Kelly BJ, Fitch JR, Hu Y, Corsmeier DJ, Zhong H, Wetzel AN, et al. Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. Genome Biol. 2015;16(1):6. https://doi.org/10.1186/s13059-014-0577-x
27. Greer SU, Ji HP. Structural variant analysis for linked-read sequencing data with gemtools. Bioinformatics. 2019;35(21):4397–9. https://doi.org/10.1093/bioinformatics/btz239
28. Mimori T, Nariai N, Kojima K, Takahashi M, Ono A, Sato Y, et al. iSVP: an integrated structural variant calling pipeline from high-throughput sequencing data. BMC Syst Biol. 2013;7(6):1–8. https://doi.org/10.1186/1752-0509-7-S6-S8
29. Becker T, Lee W-P, Leone J, Zhu Q, Zhang C, Liu S, et al. FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. Genome Biol. 2018;19(1):1–14. https://doi.org/10.1186/s13059-018-1404-6
30. Kumar S, Vo AD, Qin F, Li H. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. Sci Rep. 2016;6:21597. https://doi.org/10.1038/srep21597
31. Abate F, Acquaviva A, Paciello G, Foti C, Ficarra E, Ferrarini A, et al. Bellerophontes: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model. Bioinformatics 2012;28(16):2114–21. https://doi.org/10.1093/bioinformatics/bts334
32. Chen K, Wallis JW, Kandoth C, Kalicki–Veizer JM, Mungall KL, Mungall AJ, et al. BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. Bioinformatics.2012;28(14):1923–4. https://doi.org/10.1093/bioinformatics/bts272
33. Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics [Internet]. 2011/08/11. 2011 Oct 15;27(20):2903–4. https://doi.org/10.1093/bioinformatics/btr467
34. McPherson A, Wu C, Wyatt AW, Shah S, Collins C, Sahinalp SC. nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. Genome Res. 2012/06/28. 2012 Nov;22(11):2250–61. https://doi.org/10.1101/gr.136572.111
35. Nicorici D, Satalan M, Edgren H, Kangaspeska S, Murumagi A, Kallioniemi O, et al. FusionCatcher-a tool for finding somatic fusion genes in paired-end RNA-sequencing data. BioRxiv. 2014;11650. https://doi.org/10.1101/011650
36. Jia W, Qiu K, He M, Song P, Zhou Q, Zhou F, et al. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biol. 2013;14(2):R12. https://doi.org/10.1186/gb-2013-14-2-r12
37. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. 2011;12(8):R72. https://doi.org/10.1186/gb-2011-12-8-r72
38. Hwang C-L, Lai Y-J, Liu T-Y. A new approach for multiple objective decision making. Comput Oper Res. 1993;20(8):889–99. https://doi.org/10.1016/0305-0548(93)90109-V
39. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. 2019;20(1):213. https://doi.org/10.1186/s13059-019-1842-9
40. Haas B, Dobin A, Stransky N, Li B, Yang X, Tickle T, et al. STAR-Fusion: fast and accurate fusion transcript detection from RNA-Seq. BioRxiv. 2017;120295. https://doi.org/10.1101/120295
41. Jasper J, Powers JG, Weigman VJ. STAR-SEQR: Accurate fusion detection and support for fusion neoantigen applications. AACR; 2018. https://doi.org/10.1158/1538-7445.AM2018-2296
42. Uhrig S, Fröhlich M, Hutter B, Brors B. PO-400 Arriba--fast and accurate gene fusion detection from RNA-seq data. ESMO Open. 2018;3(2). https://doi.org/10.1136/esmoopen-2018-EACR25.426

43. Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C. The landscape of kinase fusions in cancer. Nat Commun. 2014;5:4846. https://doi.org/10.1038/ncomms5846
44. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635
45. Tian L, Li Y, Edmonson MN, Zhou X, Newman S, McLeod C, et al. CICERO: a versatile method for detecting complex and diverse driver fusions using cancer RNA sequencing data. Genome Biol. 2020;21(1):1–18. https://doi.org/10.1186/s13059-020-02043-x
46. Vellichirammal NN, Albahrani A, Li Y, Guda C. Identification of Fusion Transcripts from Unaligned RNA-Seq Reads Using ChimeRScope. In: Chimeric RNA. Springer; 2020. p. 13–25. https://doi.org/10.1007/978-1-4939-9904-0_2
47. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010/08/27. 2010 Oct;38(18):e178–e178. https://doi.org/10.1093/nar/gkq622
48. Sorn P, Holtsträter C, Löwer M, Sahin U, Weber D. ArtiFuse-computational validation of fusion gene detection tools without relying on simulated reads. Bioinformatics. 2020;36(2):373–9. https://doi.org/10.1093/bioinformatics/btz613
49. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, et al. Hi-C: a method to study the three-dimensional architecture of genomes. J Vis Exp. 2010 May;(39). https://doi.org/10.3791/1869
50. Wang S, Lee S, Chu C, Jain D, Kerpedjiev P, Nelson GM, et al. HiNT: a computational method for detecting copy number variations and translocations from Hi-C data. Genome Biol. 2020;21(1):1–15. https://doi.org/10.1186/s13059-020-01986-5
51. Chakraborty A, Ay F. Identification of copy number variations and translocations in cancer cells from Hi-C data. Bioinformatics. 2018;34(2):338–45. https://doi.org/10.1093/bioinformatics/btx664
52. Jain D, Chu C, Alver BH, Lee S, Lee EA, Park PJ. HiTea: a computational pipeline to identify non-reference transposable element insertions in Hi-C data. bioRxiv. 2020; https://doi.org/10.1101/2020.04.27.060145
53. Hombach D, Schuelke M, Knierim E, Ehmke N, Schwarz JM, Fischer-Zirnsak B, et al. MutationDistiller: user-driven identification of pathogenic DNA variants. Nucleic Acids Res. 2019;47(W1):W114–W120. https://doi.org/10.1093/nar/gkz330