Universiteit
Leiden
The Netherlands

Algorithms for structural variant detection

Lin, J.

Citation
Lin, J. (2022, June 24). *Algorithms for structural variant detection*. Retrieved from https://hdl.handle.net/1887/3391016

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3391016

Note: To cite this publication please use the final published version (if applicable).

Chapter 1

Introduction and background

This thesis is about developing algorithms for structural variant detection, validation and analysis. We focus on long-read sequencing technologies. In this chapter, we explain the biological background, the sequencing technologies and computational approaches for the analysis of human genomes. We also mention our contributions and research questions.

1.1 Computational genomics

Computational genomics is an interdisciplinary field, combining biology, computer science, information engineering, mathematics and statistics, that develops and applies computational methods to analyze deoxyribonucleic acid (DNA) sequences for predictions or novel discoveries.

DNA is a molecule composed of two polynucleotide chains that form a double helix structure carrying genetic instructions for development, functioning, growth, reproduction, etc. The two DNA strands consist of monomeric units called nucleotides, where each nucleotide is composed of one of four nitrogen-containing nucleobases, cytosine (C), guanine (G), adenine (A) or thymine (T). From the computational perspective, a genomic sequence is a special type of string, consisting of four characters (i.e., A, T, C and G), and contains many repeated substrings.

One of the common applications of computational genomics is to assess the similarity between strings or in a set of strings, such that the candidate genes, genome evolution, genetic variants, etc. can be inferred or identified. Given that genetic variants are the major sources to form population differences and to drive diseases (i.e., cancer, autism disorder, Alzheimer, etc.), the detection of genetic variants has become a major focus in the

field of computational genomics since the development of high-throughput-sequencing (HTS) technologies [1]. Briefly, genetic variants are identified by comparing an individual genome (alternative sequence, ALT) with a reference genome (reference sequence, REF). To detect genetic variants in the sequencing era, computer science and statistical approaches have been applied, the Burrows-Wheeler Transform [2] and FM-index [3] were used to perform efficient sequence alignment, the convolutional neural network [4] was used to identify single-nucleotide-polymorphism (SNP), etc. Genome rearrangement or structural variants (SV) is another form of genetic variants, and usually affects a substring containing more than 50 characters, whereas a SNP only replaces one single character. In the past decade, great efforts have been made to generate longer DNA sequences and to optimize algorithms for the discovery and genotyping of genome rearrangements.

1.2 Emerging DNA sequencing technologies

The hybridization-based microarray approaches (i.e., for comparative genomic hybridization (CGH) and SNP microarrays) are first used to infer copy number gains or losses compared to a reference sample or population, whereas these approaches cannot identify balanced SVs (i.e., inversion), as well as their structures [1]. Another approach is the single-molecule analysis, such as fluorescent in situ hybridization (FISH) and spectral karyotyping, providing the first glimpses of common and rare SVs, such as the translocation mediated BCR-ABL fusion in Leukemia [1]. However, their low throughput and low resolution limit their application to a few individuals and to particularly large SVs (≈500kb to 5Mb).

The advent of next-generation-sequencing (NGS) technology or the so-called short-read sequencing promises to revolutionize the SV studies, and replaces the microarrays for high-throughput personal genomes variant detection. Most importantly, the NGS technology opens the field of detecting and genotyping SVs with HTS technologies, and DNA sequences produced by HTS technologies are termed as read [1]. So far, the most widely-used NGS technology is the read-pair technology, which has been applied to several population-scale genome studies, such as the 1000 Genomes Project [5], International Cancer Genome Consortium (ICGC) [6], Genome Aggregation Database (gnomAD) [7], etc. Starting from 2015, a considerable increase of novel HTS technologies that leverage single-molecule-sequencing (SMS) strategies, has led to platforms that produce reads several orders of magnitude longer than short-read data, enabling the direct detection of many previously

undetected SVs. The most representative SMS platforms are single molecule real-time sequencing (SMRT) invented by Pacific Bioscience (PacBio) and single stranded DNA nanopore sequencing invented by Oxford Nanopore Technology (ONT). The average DNA sequence length, i.e., read length, generated by PacBio and ONT is around 15kbp. To get the entire human genome of 3Gbp, an individual genome is usually sequenced multiple times, called sequencing coverage. For example, if a genome is sequenced at 30X coverage, the fragmented DNA sequences could span the entire genome 30 times. In this thesis, NGS or short-read data is referred to as paired-end sequencing, and long-read data or long-read sequencing is referred to as DNA sequences produced by PacBio and ONT sequencers.

1.3 Genome structural variations are important

In the past decade, widespread application of whole-genome HTS technology for the genetic variant detection has shown that difference between individuals is presented as single-nucleotide-variants (SNVs), small insertions and deletions (indels, <50bp) and SVs. Compared with SNVs and indels, SVs are extremely diverse in size and type, ranging from 50bp to megabases of the genome. SVs (Figure 1.1A) consist of copy number variations (CNVs), which include deletions (DEL), insertions (INS) and duplications (DUP), as well as balanced rearrangements, such as inversions (INV) and inter- or intra-chromosomal translocations (TRA, Figure 1.1B) [8]. These four types were discovered and defined in the early stages of the Human Genome Project (HGP) based on short-read sequencing, and we define them as simple SVs or canonical SVs.

Recently, based on the most advanced single-molecule-sequencing (SMS) technology, producing long-read data, a series of studies conducted by the Human Genome Structural Variation Consortium (HGSVC) has estimated that each human genome contains approximately 20,000–25,000 SVs, which doubles the number of SVs estimated by next-generation-sequencing technology (NGS) [9]. Remarkably, SMS facilitates the high-quality haplotype-aware human genome assembly and phased SV detection. The Phased Assembly Variant (PAV) allows researchers to establish their population frequency, identify ancestral haplotypes and discover new associations with respect to gene expression, splicing, and candidate disease loci [10].

Additionally, another special type of SVs, consisting of multiple combinations of the simple SV types, is called complex SV [11] (CSV, Figure 1.1C). In 2015, the 1KGP first profiled the CSVs of a healthy genome, of which the

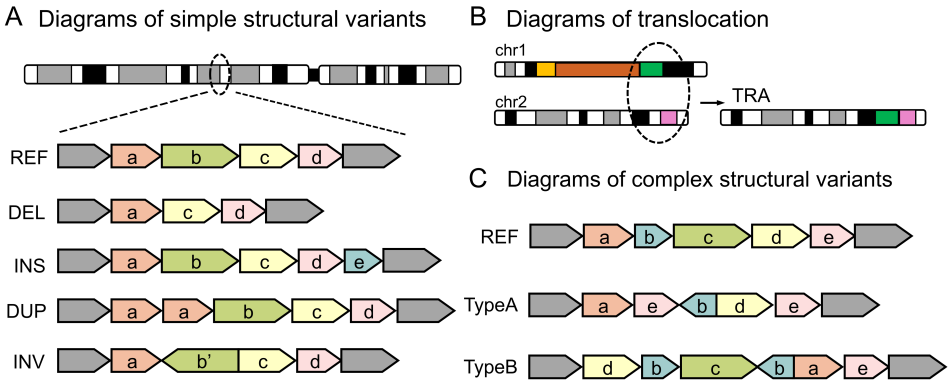

Figure 1.1: Diagrams of simple and complex structural variants. (A) Diagrams of four simple structural variants, including deletion (DEL), insertion (INS), duplication (DUP) and inversion (INV). (B) The diagram of a translocation (TRA), combining sequences from two different chromosomes. (C) The diagrams of two complex structural variant types (i.e., TypeA and TypeB).

CSVs were detected with intensive breakpoint analysis and manual curations based on SMS. This study first applied long-read data to resolve the structure of CSVs and suggested that 8% and 68% of the simple deletions and inversions are complex events [5]. In 2017, a group of researchers systematically analyzed the CSVs in a cohort of 689 patients with autism spectrum disorder and other developmental abnormalities, which was the biggest CSV study based on linked-read sequencing [12]. They identified 11,735 distinct large SV sites, and estimated each genome harbors 14 large CSVs on average. Notably, this study also found a high percentage of inversion associated CSVs, which took 84.4% of the detected CSVs.

Cancer is another complex disease, where the genome of cancer patients was changed dramatically during tumorigenesis, resulting in a great number of simple and complex SVs. The study conducted by ICGC profiled the SVs in 2,685 samples of 38 tumor types based on NGS, and first identified a group of unclassified or complex SV types in tumor genomes [6].

In Chapter 2 and Chapter 3, novel algorithms are developed to detect both simple and complex SVs from short- and long-read data, respectively. In addition, though SVs could be identified, an orthogonal approach to validate the correctness (i.e., breakpoint accuracy and type) is also important for future downstream analysis and clinical applications. Therefore, we developed

a novel algorithm to assess the quality of SVs detected by different algorithms in Chapter 4. Moreover, accumulating studies have revealed the unique strength of using long-read data to detect SVs from disease genomes, such as cancer and Mendelian disease. For example, a study of undiagnosed rare disease patients successfully identified three pathogenic CSVs that cannot be resolved by short-read data, suggesting the strength of using long-reads to characterize the exact breakpoints and structure of CSVs. In Chapter 5, we systematically evaluate the performance of the state-of-the-art long-read algorithms for both germline and somatic SV detection.

Besides the influence in downstream molecular and cellular processes, such as transcription and regulation [13], SVs are also important sources to understand the DNA damage repair mechanisms in the pathophysiological process of complex diseases such as cancer [14]. For example, the homologous recombination deficiency (HRD) has been used as an important biomarker to select drugs for a certain group of cancer patients [15].

In general, SVs are usually classified as recurrent and non-recurrent rearrangement to investigate their formation separately, where the recurrent SVs share the same size and genomic content in unrelated individuals, while the nonrecurrent ones have unique size and genomic content at a given locus in unrelated individuals [14]. CSVs often have more than one breakpoint junction and genomic interval of copy number change that can be observed at loci with susceptibility to nonrecurrent rearrangements, and replication-based mechanisms have been proposed to underlie the formation of CSVs as a result of interactive DNA template switches during replicative repair of single-ended, double-stranded DNA breaks [14]. In Chapter 2 and Chapter 3, the microhomology was identified to be the major mechanism for CSV formation, and we identified that different microhomology configurations at the breakpoint junction led to different forms of CSV. It should be noted that correct characterization of CSV formation requires accurate configuration of the breakpoint and structure, which is usually difficult to achieve using short-read data.

1.4 Detecting structural variation

Indeed, SVs of an individual genome manipulate the sequence of the reference genome, resulting in the so-called alternative sequence, and different types of SVs alter the reference sequence in different ways. In principle, all reads would be properly aligned if the sample's genome is identical to the reference, whereas the abnormally aligned reads footprint the signatures of SVs. For

instance, a deletion event indicates the sample genome missed one fragment of DNA sequence that was found in the reference genome (Figure 1.2A). The start and end position on the reference genome of the altered sequence are called *breakpoints* or *breakpoint junctions*, which are the junctions between alternative and reference sequence of the sample.

It should be noted that the SV breakpoint is defined according to the reference coordinate system, thus insertion only has one breakpoint junction on the reference compared with deletion, inversion and duplication (Figure 1.2B). The number of breakpoint junctions is often used to distinguish the simple and complex SVs, where CSVs usually have more than two breakpoint junctions. Afterwards, according to the altered sequence originating from different SV types, detection algorithms first build the SV signature model from the abnormally aligned reads for each type, where the model essentially depicts the pattern indicating how reads are aligned across the breakpoint junctions (Figure 1.3A). Therefore, to detect SVs, it is important to know how a specific SV type alters the reference sequence and its corresponding pattern inferred from the alignment. Once the SV signature models are built for each type, the detection algorithm would fit the observed read alignments with the expected model to make the detection. This approach is considered as a model-based approach, containing two major steps: i) SV signature modeling and ii) model fitting.

The model-based approach is initially designed for short-read data due to lack of SV spanning sequences, while assembly of short-reads provides longer DNA sequences and improves the detection performance. Briefly, the assembly based approach first collects all abnormally aligned reads to produce longer sequences based on the De Bruijn graph [21, 22] or string graph [23, 24]. Then, the assemblies are realigned to the focal regions, which is used to fit the prebuilt SV signature models for discoveries. Moreover, because the assemblies might span multiple breakpoint junctions, the assembly approach is widely used to detect CSVs from short-read data. Though the assembly approach is able to detect multiple breakpoints of CSVs, it often requires further efforts to filter redundant breakpoints and identify breakpoints belonging to the same events [5, 11].

The long-read technology produces even longer sequences than the assemblies from short-read data. It greatly simplifies simple SV detection from both signature modeling and model fitting because of variant spanning reads (Figure 1.3B). For CSV detection, though the long-read data avoid the assembly issues, it also follows the model-based approach, such as Sniffles [18], which is the only algorithm that detects two specific types of complex events with extra models (Figure 1.3C). However, CSVs are largely unexplored and

contain complex breakpoint configurations [25], making them challenging to model in a brute-force way. Moreover, current studies interpret and define CSVs in various ways, hindering the generalization of CSV study between researchers. Thus, one of the major objectives of this thesis is to develop novel algorithms for CSV detection without models.

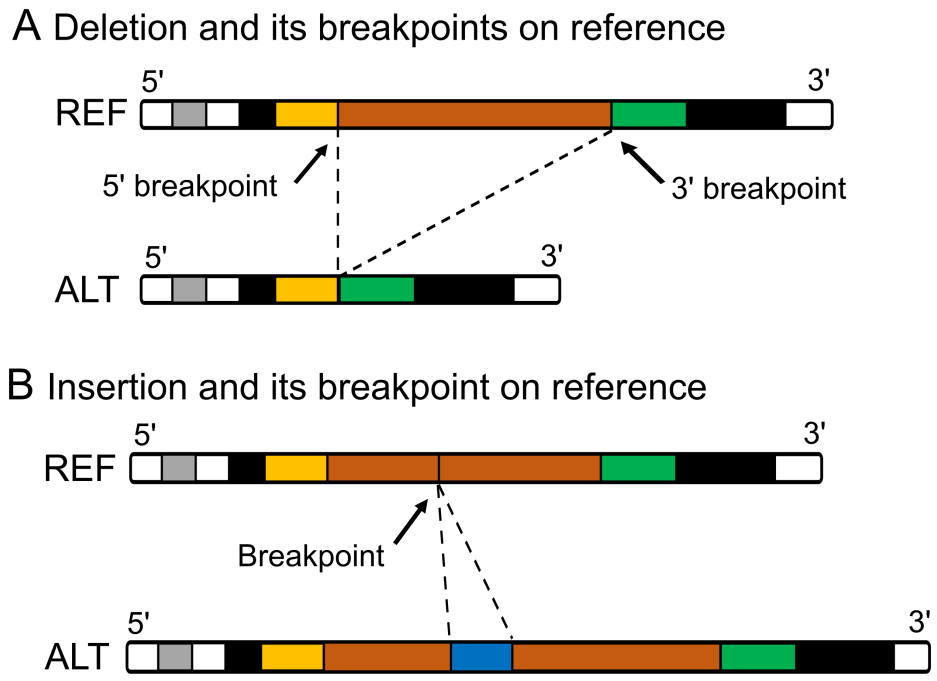

Figure 1.2: Breakpoints of two simple structural variants. (A) The breakpoints defined for a deletion, including the 5' breakpoint and 3' breakpoint on the reference. (B) The breakpoint defined for an insertion, which only has one breakpoint.

1.5 Pairwise sequence alignment for nucleotide sequences

Pairwise nucleotide sequence alignment has been used to investigate the differences between multiple genomes, to create the evolutionary tree for species, etc., which is one of the classical computational biology problems.

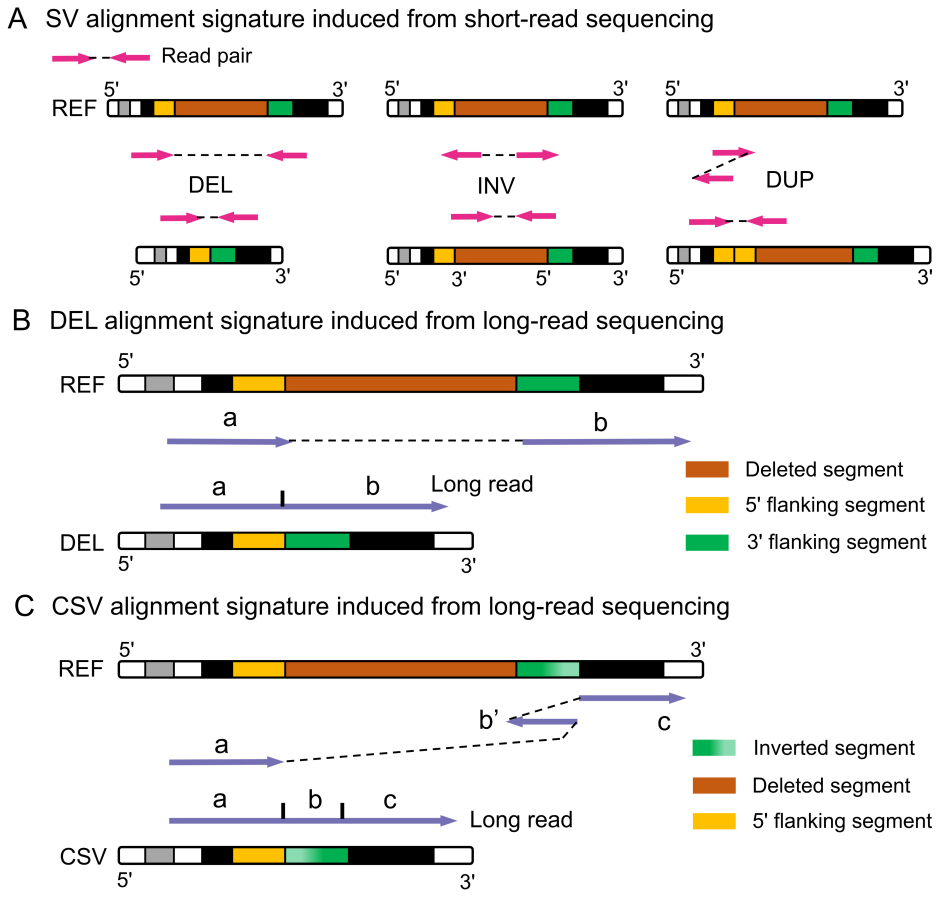

Figure 1.3: Structural variant alignment signature derived from short-read and long-read sequencing. (A) The short-read alignment signatures of three simple structural variants (SV), i.e., deletion (DEL), inversion (INV) and duplication (DUP). (B) The DEL alignment signature derived from long-read alignment. (C) The alignment signature of a complex structural variant (CSV) that is derived from long-read data.

MUMMer [16], using suffix-trees, is the most widely used algorithm for large-scale genome alignment, and it has been used to investigate the genome rearrangements between genomes.

Detecting SVs is similar to genome rearrangement detection between genomes, whereas the most advanced SMS technologies only produce fragmented DNA sequences, making it difficult for a MUMMer like approach to detect SVs genome-wide. Therefore, in order to detect SVs from an individual genome, the fragmented DNA sequences are first aligned to the human reference genome. The most common whole genome alignment algorithms, i.e., minimap2 [17] and ngmlr [18], adopt a typical seed-chain-align procedure to map the sequenced reads to the reference genome.

Briefly, for each query sequence (DNA sequence), minimap2 takes query minimizer as seeds, i.e., a longest exact match between query sequence and reference, and identifies sets of colinear matches as chains. Afterwards, dynamic programming is used to extend from the ends of the chains and to close regions between adjacent matches in chains. Fortunately, the long-read data spanning the SV site enables pairwise read and reference sequence comparison, promoting correct characterization of CSV structure. In Chapter 3, a light-weighted focal sequence realignment is proposed to refine the potential breakpoints of CSVs. This realignment approach is also based on seed-and-extension, whereas the gaps between nonlinear matches are not extended and considered to contain breakpoints.

The reference genome was first published in 2001 by HGP and has been significantly improved due to long-read technologies [19]. Although studies based on long-read data suggest that the reference could not easily serve as a standard genome, the routine genomic analysis, such as SV detection, still uses the reference genome as a universal genome. Thus, it should be noted that SVs of an individual genome are the different sequences compared with the reference genome, and the same reference is used to explore SVs in populations. Currently, the human genome is at version 38 (GRCh38), which now has fewer than 1,000 reported gaps, driven by the efforts of the Genome Research Consortium (GRC) [19].

The standard format of the alignment output is the Sequence Alignment Map (SAM) [20], and the Binary Alignment Map (BAM) is the binary version of a SAM file. The BAM file is usually used as input for SV detection, while recently another form of compressed BAM file (CRAM) is introduced for processing the large data volumes for population scale genome studies.

1.6 Usage of graphs for structural variants detection and analysis

A graph, consisting of nodes and edges, is an important data structure to model many types of relations and processes in physical, biological, social and information systems, and has a wide range of useful applications. There are three major graph types, i.e., undirected graphs, directed graphs and weighted graphs, and they have been broadly used for computational genomics. One of the most important applications of graphs is genome assembly, especially since the development of HTS technologies. The ultimate purpose of genome assembly is to build each chromosome from the fragmented DNA sequences. The method can be classified into reference guided assembly and de novo assembly, where de novo assembly achieves a real personal genome [26].

The development of long-read sequencing greatly promotes the de novo genome assembly, where two major graph data structures (i.e., De Bruijn graph and string graph) are used to produce long contiguous pieces of sequence (contigs). For the De Bruijn graph, each k-mer (a length k substring of a DNA sequence) is an edge directed from node A to node B if the $(k-1)$-mer in node A is a prefix, and that in B is a suffix of the k-mer [21]. Different from the De Bruijn graph, the nodes in string graphs are reads and edges connect two overlapping reads [24]. In the past decade, several optimized graph data structures based on either De Bruijn graph [22] or string graph [23] have been proposed to achieve the longest continuous sequence. Currently, with the PacBio hifi-fidelity (HiFi) reads, assemblers such as hifiasm [23] perform graph trio binning on the string graph to generate the final haplotype-resolved assembly of human genomes.

As we mentioned in the above section, realignment of short-read de novo assembly is a popular approach for both simple and complex SV detection. Another approach uses graphs but avoids assembly, aiming to identify fragmented DNA sequences that originated from the same longer piece of sequence, from which SVs could be accurately detected with short-read. For example, CLEVER [27] organized all abnormally aligned short-reads into a read alignment graph, where max-cliques were detected and statistically evaluated for their potential to reflect insertion or deletion based on the pre-built signature models. Inspired by CLEVER and the nature of SV, either SVs or CSVs would alter one or more genomic segments at a focal region and lead to disordered segment connections compared to the reference. Specifically, SVs or CSVs change the connection relation of DNA segments at the breakpoint junctions, and the reads across the junction will

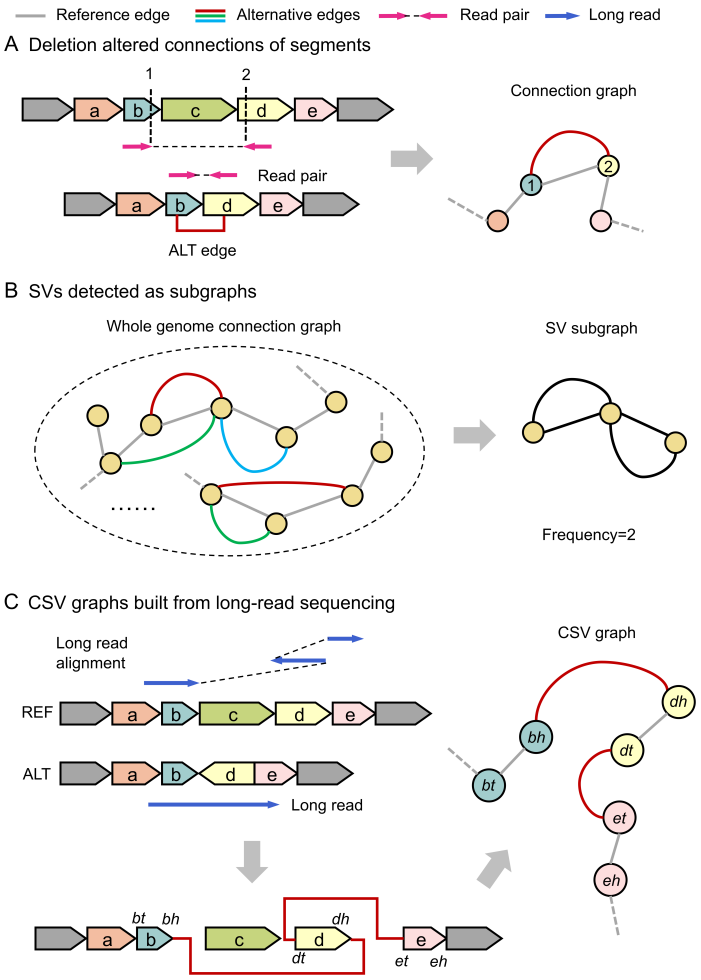

Figure 1.4: Overview and examples of detecting structural variants with a graph. (A) The connection graph created from short-read alignment on a deletion event, where node 1 and 2 indicate the mapping position of the read-pair on the reference genome (i.e., anchor position of read-pair). The connection provided by read-pair alignment is considered as alternative (ALT) edge in the graph. (B) Given the whole genome graph built based on (A), a structural variant is detected as a subgraph. (C) The graph is used to represent a complex structural variant based on long-read alignment, each node in the graph indicating a segment with tail and head. E.g., segment `b` produces two nodes in the graph: *bt* and *bh*. Similar to an edge in a short-read connection graph, the ALT edge is obtained from long-read mapping.

connect segments that are distant or discontinuous on the reference genome. For example, a deletion indeed connects two distant segments that are not adjacent on the reference genome (Figure 1.4A). Thus, we are able to create a segment connection graph (Figure 1.4A) from short-read (i.e., paired-end reads) alignments according to the signature model (Figure 1.3A), where the anchored positions of a mapped read are used as nodes and two nodes could be obtained from one paired-end read mapping (i.e., one end corresponds to one node). In terms of the edge set, one part of the edges are derived from the reference connection, indicating the identical connections between two adjacent segments on the reference, while the alternative edges are given by the abnormal aligned paired-end reads. Then, a SV or CSV is modeled as a subgraph that is involved in the genome wide segment connection graph (Figure 1.4B).

In Chapter 2, this segment connection graph is called signal graph, and we add extra attributes to the nodes and edges for CSV detection. However, using the current linear reference genome for variant detection, some major allele events have been observed and shown to be population specific variants, which might mislead further variant analysis. Therefore, another important application for graphs is to encode the population genetic diversities, producing the so-called graph genome [28, 29] or variation graph [30, 31, 32], which is then used as reference for variant detection. Given a graph reference, algorithms for efficient building, augmenting, storing, querying and variant calling are under active development [33, 34]. In 2018, the first genome-wide full pipeline of using graphs is developed, named vg [32], which improves read mapping sensitivity and increases the variant calling recall, and effectively removes reference bias.

Inspired by the above applications, we consider the graph to be a powerful data structure to represent complex events, such as the junction-balanced genome graph [35] which has been proposed to infer and classify complex rearrangements observed in tumor genomes. In Chapter 2, the CSV detected as a subgraph from the signal graph could be interpreted from the graph connections, which also enables the comparison of different types of CSVs. Inspired by the SV subgraph introduced in Chapter 2, the graph is used to represent and interpret complex events detected from long-read data in Chapter 3. This CSV mini-graph induced from long-read data contains nodes from matched segments between reference and alternative sequence, where the edges originate from the alternative sequence (Figure 1.4C). Note that in the long-read induced graphs, each node not only indicates the position on reference but also the matched segment sequences, which is different from the one derived from short-read data. Moreover, we introduce head and tail

annotation of each segment to indicate potential inversions. For example (Figure 1.4C), a segment is inverted, and the head is connected with an alternative edge observed in long read. Given the CSV mini-graph, we are able to compare different CSV events and identify those of the same type based on isomorphic graphs. Moreover, this mini-graph provides a so-called SV graph reference, such that the same event at identical loci could be identified or genotyped among populations via graph-based sequence alignment.

1.7 Frequent subgraph mining

Frequent subgraph mining raised great interest in the data mining community since 2000, and had a broad application in many fields, such as social media, chemical compound analysis, etc. The idea behind frequent subgraph mining is to "grow" candidate subgraphs, in either a breadth first or depth first manner, and then determine if the identified candidate subgraph occurs frequently enough in the graph data set for them to be considered interesting. Effective candidate subgraph generation is required to avoid the generation of duplicate or superfluous candidates, and the occurrences counting needs examination of graph isomorphism.

According to different applications, the optimization of frequent subgraph mining algorithms usually focuses on i) candidate generation strategy; ii) reduction of search space and iii) graph structure comparison. In Chapter 2, the abnormal short-read alignments, footprinting potential SVs, were coded in a signal graph (Figure 1.4B). Since SVs alter the focal genome by adding alternative edges or removing reference edges in the signal graph, the SV or CSV could be detected as a local maximal subgraph. In addition, a real SV or CSV graph structure usually occurs more frequently than the subgraph induced from alignment artifacts, thus the algorithm of detecting frequent local maximal subgraphs was developed to identify both simple and complex events (Figure 1.5). According to the specific application, this algorithm optimized the approach of saving the signal graph and comparing graph structure.

1.8 Objective and outline of the thesis

The main research questions for this thesis include i) detecting complex structural variants without prior knowledge and ii) reproducing structural variants detection among different datasets. The three main perspectives of this thesis are i) to develop novel algorithms for the detection of genomic

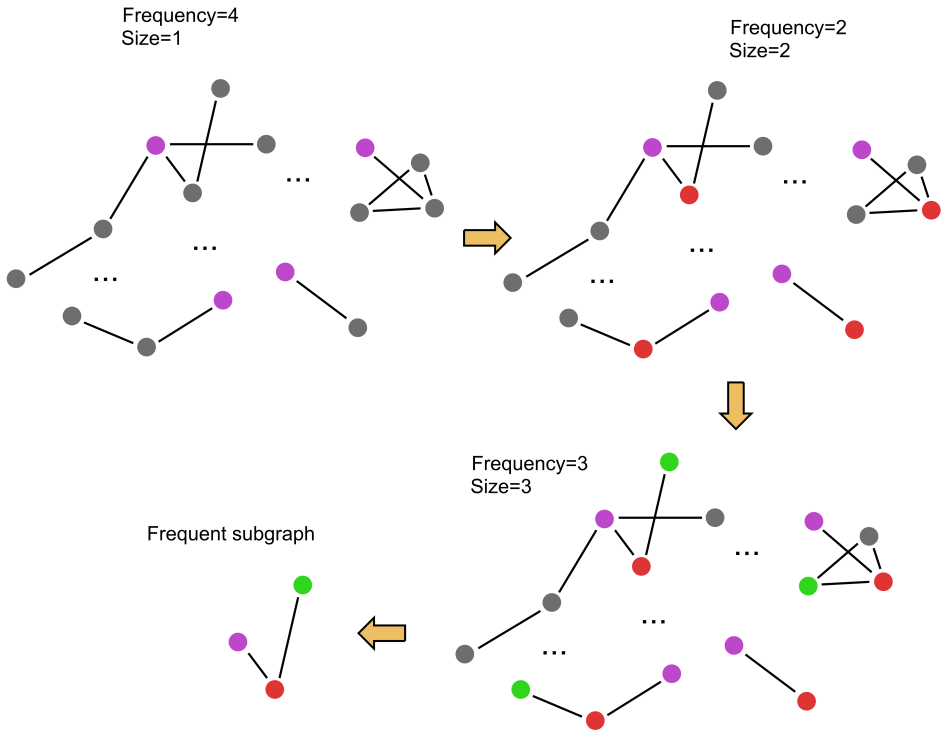

Figure 1.5: A toy example for detecting frequent maximal subgraphs from the signal graph. The subgraphs starts growth with purple node, from which subgraphs of size 2 and frequency 2 are obtained by adding red node. Furthermore, green node that satisfies the growth constraint are added to existing graph of size 2, resulting in subgraphs of size 3 and frequency 3. Finally, a frequent subgraph consisting of purple, red and green node is detected.

structural variation, especially for complex structural variations, from the short-read data (i.e., for the Illumina sequencer) and long-read data (i.e., for the PacBio and ONT platforms), ii) to develop structural variation validation algorithms and iii) to evaluate a pipeline or pipelines for detecting germline and clinically relevant structural variations.

First, for the paired-end sequencing, we use a graph data structure to encode the footprints of SVs from a given alignment and detect both simple and complex SVs by mining the frequent maximal subgraphs (Chapter 2). In the graph, a node consists of abnormal alignments, such as split-reads and discordant alignments. For the edges, two adjacent nodes are connected via a reference edge, and paired-end reads could connect two distinct nodes, referring to alternative edges. Afterwards, a frequent maximal subgraph is considered as a SV, and we set a frequency threshold for users to filter potential subgraphs made by background noise. However, since the paired-end reads are only approximately one hundred base pairs in length, it becomes challenging or even impossible to detect and interpret the whole structure of SVs, especially for complex ones. Given that the SMS provides long DNA sequencing that could cover the entire SV structure, in Chapter 3, we automate the detection and interpretation of both simple and complex SVs by recognizing the sequence differences coded in the image. Briefly, we visualize each alignment as a sequence similarity image, for which the multi-object recognition framework is applied to detect SVs without any predefined signature models. Then, we introduce a graph to represent and classify different classes of complex SVs.

Given that SMS based population scale SV study becomes common, the orthogonal approach to validate detected SVs is in great demand for the community, especially for the potential clinical application. In Chapter 4, inspired by the sequence similarity image, we provide a novel approach called SpotSV, to validate and characterize simple and complex SVs occurring in genomic regions of different complexity. In general, SpotSV validates a given SV in two major steps: i) simulating an alternative sequence with SV profile and ii) pairwise comparison of the reads and the simulated alternative sequence.

As the sequencing price drops, long-read technologies have been applied to study population genetic diversities, evolution, etc. Notably, in the past two years, several studies have shown the power of using long-read data to investigate disease genomes (i.e., for tumor and Mendelian diseases). Therefore, in Chapter 5, we aim to evaluate the existing long-read detection algorithms for both germline and somatic SV discovery. Specifically, we use five samples sequenced by PacBio HiFi and ONT, two alignment algorithms

and six widely-used detection algorithms, to examine and compare the performance of each detection algorithm.

In Chapter 6, conclusions are drawn and further perspectives are discussed.

In this thesis, Chapter 2 and Chapter 3 are based on the following publications:

- Jiadong Lin, Xiaofei Yang, Walter Kosters, Tun Xu, Yanyan Jia, Songbo Wang, Qihui Zhu, et al. "Mako: a graph-based pattern growth approach to detect complex structural variants." Genomics, proteomics & bioinformatics, 2021.

- Jiadong Lin, Songbo Wang, Peter Audano, Deyu Meng, Jacob Flores, Walter Kosters, Xiaofei Yang, Peng Jia, Tobias Marschall, Christine Beck and Kai Ye. "SVision: A deep learning approach to resolve complex structural variants." Nature Methods (Under revision, submission ID: NMETH-BC48137), 2022.

Moreover, we contributed to the following publications:

- Peter Ebert, Peter A. Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jiadong Lin, et al. "Haplotype-resolved diverse human genomes and integrated analysis of structural variation." Science, 2021.

- Peng Jia, Xiaofei Yang, Li Guo, Bowen Liu, Jiadong Lin, Hao Liang, Jianyong Sun, Chengsheng Zhang, and Kai Ye. "MSIsensor-pro: fast, accurate, and matched-normal-sample-free detection of microsatellite instability." Genomics, proteomics & bioinformatics, 2020.

1.9 List of abbreviations

Aligner Algorithm that maps long-read data to human reference genome

BAM Binary alignment map

Caller A certain algorithm for structural variant detection

Callset A set of structural variants discovered by detection algorithms

CCS Circular consensus sequencing

CGH Comparative genomic hybridization

CLR Continuous long read

CNV Copy number variation

CRAM Compressed BAM file

CSV Complex structural variant

GIAB Genome in a bottle

HGSVC Human genome structural variation consortium

HiFi Long-read data generated by PacBio CCS sequencing

HGP Human genome project

HRD Homologous recombination deficiency

HTS High throughput sequencing

Long-read Long-read data generated by SMS technology, such as a PacBio sequencer

NGS Next-generation sequencing

ONT Oxford nanopore technology

PacBio Pacific Bioscience

PAV Phased assembly variant

NGS Next generation sequencing

SAM Sequence alignment map

Short-read Paired-end data generated by NGS technology, such as an Illumina sequencer

SMS Single molecule sequencing

SNP Single nucleotide polymorphism

SNV Single nucleotide variant

SV Structural variant

1KGP 1000 Genomes project