Algorithms for structural variant detection

Lin, J.

English summary

Structural variants (SVs) are the hidden architecture of the human genome, and are critical for us to understand diseases, evolution, and so on. The development of both sequencing technologies and computational tools greatly facilitates the detection of SVs, while misinterpreting or even missing complex ones. Detecting and characterizing complex events is a typical field requiring multiple disciplines, i.e., domain knowledge and computer science algorithms.

In this thesis, we introduce novel algorithms to detect and validate complex events, and assess the reproducibility of current SV detection pipelines for clinical and research settings.

Chapter 1 begins with the introduction of DNA, various types of SVs and sequencing technologies. Then fundamental techniques and algorithms from computer science related to the thesis are briefly described. Pattern mining, graphs and deep learning are applied to detect and characterize different types of complex SVs (CSVs). CSVs usually contain multiple breakpoints and are often missed or misinterpreted by traditional detection strategies developed for simple SV detection. Most importantly, CSVs are largely underexplored, making them even challenging to detect based on existing knowledge. Considering the sequencing cost and detection accuracy for different application scenarios, we first develop algorithms for both short-read and long-read sequencing technologies without pattern matching against a database of know structures of SVs. Currently, short-read sequencing is significantly reduced in cost and has been widely applied to clinical diagnostics and cohort studies. To detect CSVs from short-read sequencing, we consider that SVs change the connections of adjacent segments with alternative connection derived from abnormally aligned paired-end reads.

Accordingly, in Chapter 2, we propose a frequent maximal subgraph mining approach (Mako) to detect both SVs and CSVs from a graph built from abnormal alignments. This graph is called signal graph, where nodes represent positions of connected genomic segments and edges indicate alternative and reference connections between genomic segments. We then apply a linearized database with prefix index schema to efficiently detect frequent maximal subgraphs from the signal graph, from which SVs and CSVs are derived from detected subgraphs. Compared to other approaches, a graph is able to depict complex genomic segment connections originating from CSVs. Moreover, detected CSV subgraphs are interpretable, making it possible to understand and compare different types of CSVs. However, limited by the read length of short-read sequencing, two simple SVs from different haplotypes might be detected as a single CSV event. On the other hand, short read length would

also be problematic for read mapping at regions with potential CSVs, where breakpoints belonging to a CSV could be potentially missed by callers. With the advances of long-read sequencing, a single read is more likely to span an entire CSV event compared to short-read sequencing. This greatly simplifies the confirmation of CSVs by investigating the difference and similarity of reads and its counterpart sequence from the reference genome. As a result, an increasing number of CSV have been revealed through intensive breakpoint analysis and visual confirmation. However, this is only applicable to small amounts of samples, which would not satisfy the ever-increasing demand of studying CSVs at population scale.

In Chapter 3, we leverage the human intelligence of identifying CSVs from visualization, and develop a multi-object recognition framework (SVision) to detect both SVs and CSVs without previous knowledge of SV structures. We first propose a sequence-to-image coding schema, which not only describes the differences and similarities of two sequences but also removes the background sequence context. This coding strategy enables us to efficiently and effectively detect CSVs even at complex genomic regions. In addition, CSV representation or interpretation is another challenging problem that hinders the definition and cross study of CSVs. Inspired by the graph structure used in Chapter 2, we also use a graph to represent and compare CSVs detected from long-read data, from which we are able to classify different types of CSVs by measuring graph isomorphisms. But different from nodes in the signal graph proposed in Chapter 2, a node of the CSV graph in Chapter 3 represents a matched sequence between two sequences. This feature makes it possible to genotype CSVs based on graph alignment. Moreover, this provides a novel idea of detecting SVs from a SV graph instead of detecting from a biased linear reference. We expected this graph-based SV detection approach will help to detect somatic SVs and SVs from tumor subclones.

Having developed two SV detection algorithms for trending sequencing technologies, we next aim to further explore the possibilities of applying long-read sequencing in various applications. In general, we observe that the high-confident SVs detected from reproducible analysis pipelines are critical for long-read applications in either clinical or research settings. Therefore, we first develop a high-throughput SV validation approach (SpotSV) to identify high-confident SVs in Chapter 4. Different from SV detection, SV validation focuses on exclude false negatives and corrects inaccurate SV characterizations, such as type and breakpoints. The idea of this validation approach is also inspired by the way in which human experts visually characterize SVs. We first apply a light-weighted local realignment method to locate different segments between two sequences. Then, we adopt a simple two-dimensional geometry

calculation to measure the confidence of a detected SV.

Additionally, in Chapter 5, we assess the reproducibility of existing pipelines on detecting germline and somatic SVs. This chapter systematically investigates the difference of assembly-based and alignment-based SV detection, highlighting major factors for discordant discoveries. We expect that this evaluation will help non-experts to understand the difference between methods and thus will help them to select proper analysis pipelines in their own applications.

Finally, in Chapter 6, we mention future research directions regarding the accurate detection of SVs for both research and clinical settings. Notably, we are confident that the combination of BioTech and InfoTech, often referred to as BT-IT, will revolutionize future health care.