Algorithms for structural variant detection

Lin, J.

Chapter 6

Conclusions and perspectives

In this chapter, we present our conclusions and provide perspectives for future research.

6.1 Conclusions

It is a fact that the research discipline of computational genomics largely emerged from sequence analysis. Indeed, deciphering the language of life from DNA, RNA or protein sequences has been greatly facilitated by the advanced sequencing technologies. From short-read DNA sequencing to single-molecule-sequencing (SMS) DNA sequencing, methods for sequence alignment, genome structural variants detection, etc., are actively developed by numerous researchers in the past decade. This thesis focuses on developing novel algorithms for several pivotal parts in applying sequencing technology to SV detection in clinical settings, including detection, characterization and validation. Moreover, this thesis provides a systematic evaluation of factors affecting clinical applications.

With the rapid development of high-throughput sequencing (HTS) technology, genomic rearrangements or structural variants (SVs) have been recognized to affect more than SNPs or Indels in genome evolution and disease progression. Recently, an increasing number of simple SVs are found to be complex events, which not only misleads downstream analysis but also introduces another layer of difficulty for SV detection. So far, most of the methods detect SVs by following a model and match approach, i.e., a sequencing data specific alignment model is first created for different SV types and further matched with the observations from sequence alignment for discovery. Though mode-based approach is well-performed for detecting simple SVs

(i.e., deletions, inversions, duplications, insertions and translocations), it is neither effective nor efficient to resolve complex events due to the complicated internal structure for modeling. On the other hand, complex events are largely unexplored, which also limits the detection through a model-based approach. Therefore, novel algorithms that can detect complex events or curate existing discoveries are in great demand, especially for sequencing oriented clinical diagnosis.

In this thesis, we first design two novel algorithms, graph based and deep learning based, to detect complex structural variants (CSVs) without predefined models. Secondly, we systematically assess the reproducibility of current SV detection methods among different datasets, helping users select proper methods and datasets for their applications. In this way, we address our main research questions, dealing with detection and assessment of structural variants.

Since short-read sequencing has been widely used in large cohort studies, a graph based approach was first developed, aiming to profile complex events at a large scale. Specifically, the graph was used to represent alternative connections derived from an individual genome, from which CSVs were detected as frequent local maximal subgraphs. However, due to the limited read length, the graph-based approach based on short-read data was not able to resolve the accurate internal structure.

As the price of long-read sequencing decreases, its usage for both research and clinical settings is expected to increase dramatically in the next few years. Therefore, we further developed SVision, a deep-learning based multi-object recognition framework, to automatically detect and characterize both simple and complex SVs from sequence image. In addition, since vast amounts of sequence data and SV callsets have become available, a high-throughput orthogonal validation approach is also in demand. We thus developed a novel algorithm, SpotSV, to assess the quality of predicted SVs, including their breakpoints and type. SpotSV uses the denoised segment to examine the breakpoints of predicted SVs, improving the assessment of complex events and SVs at repetitive regions. Our results suggest that the novel detection algorithms and the validation algorithm outperformed the state-of-the-art methods.

Furthermore, it is expected that HTS based SV detection will become a routine clinical diagnosis approach, especially for complex diseases, such as cancer. We then systematically evaluated the robustness of detection algorithms by using different sequence alignment algorithms and sequencing platforms, of which the sensitivity, specificity and breakpoint accuracy were examined.

6.2 Perspectives

This section contains directions for future research.

Flexible connection graph data structure for SV detection
In Chapter 2, a graph, representing alternative connections, has been successfully used to detect simple and complex events from an individual genome. Recently, long-read sequencing has revolutionized the detection and study of SVs, and it would add extra connections to the graph built on short-read. Moreover, long-reads are able to accurately resolve the CSV internal structure as we show in Chapter 3. Thus, a flexible data structure that could integrate both the advantage of short-read and long-read sequencing is expected to improve the detection, such as finding the accurate breakpoints induced by short-reads and internal structure characterized by long-reads.

Frequent subgraph mining among population genome graph
Since a large amount of sequencing data is available for both healthy and disease genomes, one of the biggest issues is how to detect SVs at a large scale, which is critical to understand evolution and disease progression. In principle, each individual genome mapping to the reference genome could be converted to a connection graph, thereby leading to a population-scale genome connection graph. Afterwards, frequent subgraphs representing certain types of SV or CSV could be detected based on the subgraph topology. Most importantly, a population-scale genome connection graph would enable rare SV detection in personal genome because the rare SV of an individual genome might be frequent among population. This feature makes it a valuable data structure to compare SVs within population or between populations, and it could also facilitate the analysis of undiagnosed disease.

Compare and merge SVs at population-scale
In Chapter 4, we develop a novel algorithm to assess the quality of predicted SVs, especially for complex ones and SVs at repetitive regions. Similar to SV quality evaluation of an individual genome, comparison and merging SVs at population-scale is another challenging computational problem, affecting downstream analysis, such as SV formation and Mendelian disease. In general, SV comparison is difficult because they vary across individuals and are discovered through different data and methods. Therefore, an approach that could detect and merge SVs simultaneously at population-scale is able to avoid the issue of detecting from different data and methods. We would also adopt the idea of a population-scale connection graph, integrating both short-read and long-read data, for SV comparison and merging at population-scale. Specifically, if one SV is common in population, it is expected to detect

a local subgraph of dense alternative connections derived from different individuals, and these connections are approximately equal based on specific edge attributes. Finally, a merged SV call set of multiple samples could be derived from detected subgraphs in the connection graph.

SV graph for somatic SV detection

In Chapter 3, a graph is used to represent the CSV internal structure, which also enables the graph based validation via graph alignment. So far, a number of studies have shown the strength of using long-reads to analyze tumor genomes compared with short-read data, whereas algorithms for somatic SV detection based on long-reads are underdeveloped. The graph implemented in Chapter 3 provides an important hint to isolate somatic SVs from the genetic background of a patient. Briefly, the germline SVs (i.e., genetic background) represented as mini graph are first embedded into the linear reference genome, resulting in a germline graph. Secondly, the sequencing data from matched tumor tissue could be aligned to the germline graph, from which the newly formed subgraph or path is identified to be a somatic SV and the augmentation graph could be built. Since tumor tissue might contain cells originated from different clones, detecting SVs from different clones is important to help understand the tumorigenesis. The above step could be done recursively to detect subclonal SVs, and this recursive process is terminated when new path could not be identified from the graph alignments. This is a complicated computational approach, which requires optimized graph augmentation and alignment algorithm for effective SV detection.

A structural variants analysis system for clinical settings

In Chapter 5, we have evaluated the factors that might affect SV detection from the clinical perspectives. On the other hand, the downstream analysis, such as SV quality assessment (Chapter 4) and SV merging, is also critical for clinical diagnosis. Therefore, a SV analysis system from detection to result interpretation would fill the gap between research output and clinical application, which becomes even important as the number of undiagnosed cases increases and the price of sequencing decrease. Moreover, a user-friendly interface for doctors or non-computing experts is preferred and valuable to expand the usage of sequencing technology assisted diagnosis. Therefore, as an algorithm designer and implementer, future research will continue to develop algorithms for challenging biological or clinical problems. In addition, we aim to carefully implement the algorithms and provide user-friendly graphic interfaces, enabling the application of HTS technology in clinical settings.