DOCTORAL THESIS

Detection copy number variants profile by multiple constrained optimization

Zhang, Yue

Date of Award:
2017

Abstract

Copy number variation, causing by the genome rearrangement, generally refers to the copy numbers increased or decreased of large genome segments whose lengths are more than 1kb. Such copy number variations mainly appeared as the submicroscopic level of deletion and duplication. Copy number variation is an important component of genome structural variation, and is one of pathogenic factors of human diseases. Next generation sequencing technology is a popular CNV detection method and it has been widely used in various fields of life science research. It possesses the advantages of high throughput and low cost. By tailoring NGS technology, it is plausible to sequence individual cells. Such single cell sequencing can reveal the gene expression status and genomic variation profile of a single-cell. Single cell sequencing is promising in the study of tumor, developmental biology, neuroscience and other fields.

However, there are two challenging problems encountered in CNV detection for NGS data. The first one is that since single-cell sequencing requires a special genome amplification step to accumulate enough samples, a large number of bias is introduced, making the calling of copy number variants rather challenging. The performances of many popular copy number calling methods, designed for bulk sequencings, are not consistent and can not be applied on single-cell sequenced data directly. The second one is to simultaneously analyze genome data for multiple samples, thus achieving assembling and subgrouping similar cells accurately and efficiently. The high level of noises in single-cell-sequencing data negatively affects the reliability of sequence reads and leads to inaccurate patterns of variations.

To handle the problem of reliably finding CNVs in NGS data, in this thesis, we firstly establish a workflow for analyzing NGS and single-cell sequencing data. The CNVs identification is formulated as a quadratic optimization problem

with both constraints of sparsity and smoothness. Tailored from alternating direction minimization (ADM) framework, an efficient numerical solution is designed accordingly. The proposed model was tested extensively to demonstrate its superior performances. It is shown that the proposed approach can successfully reconstruct CNVs especially somatic copy number alteration patterns from raw data. By comparing with existing counterparts, it achieved superior or comparable performances in detection of the CNVs.

To tackle this issue of recovering the hidden blocks within multiple single-cell DNA-sequencing samples, we present an permutation based model to rearrange the samples such that similar ones are positioned adjacently. The permutation is guided by the total variational (TV) norm of the recovered copy number profiles, and is continued until the TV-norm is minimized when similar samples are stacked together to reveal block patterns. Accordingly, an efficient numerical scheme for finding this permutation is designed, tailored from the alternating direction method of multipliers. Application of this method to both simulated and real data demonstrates its ability to recover the hidden structures of single-cell DNA sequences.

Keywords: Copy number variants, read depth, sparsity, total variation, single-cell sequencing, next generation sequencing, permutation

Table of Contents