

DOCTORAL THESIS

Detection copy number variants profile by multiple constrained optimization

Zhang, Yue

Date of Award:
2017

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

Abstract

Copy number variation, caused by 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 sub-microscopic 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

Declaration	i
Abstract	ii
Acknowledgements	iv
Table of Contents	vi
List of Tables	ix
List of Figures	x
Chapter 1 Introduction	1
1.1 Organization of the Dissertation	4
1.1.1 Introduction and Organization	4
1.1.2 Next-generation Sequencing and Preprocessing Workflows . . .	5
1.1.3 Detecting Copy Number Variants from NGS by Both Sparse and Smooth Constraint Model	7
1.1.4 Revealing Common Copy Number Patterns by a Total-variation Constrained Permutation Model	9
1.1.5 Conclusions and Future Work	11
Chapter 2 Next-generation Sequencing and Preprocessing Workflow	12
2.1 Development of Sequencing Technologies	12
2.2 Variation Calling for Next-generation Sequencing Data	15
2.2.1 Single-nucleotide Variation	16
2.2.2 Copy Number Variable Regions	17
2.2.3 Single-cell Sequencing	17

2.3	Copy Number Variation	20
2.3.1	Definition of Copy Number Variation	20
2.3.2	Current Studies on Copy Number Variation	21
2.3.3	Methods of Detecting Copy Number Variation	24
2.4	Preprocessing Workflow for Sequencing Data	31
2.4.1	Preprocessing of Bulk Sequenced Data	31
2.4.2	Preprocessing Framework for Single-cell Sequencing Data	33
2.5	Chapter Summary	36

Chapter 3 Detecting Copy Number Variants from NGS by Both Sparse and Smooth Constraints Model 37

3.1	Introduction	37
3.2	Overview of Existing Models for Detecting CNV	39
3.3	Problem Modeling for CNV Detection Model with Multiple Norm Constraints	41
3.4	Numerical Solution by Alternating Direction Minimization (ADM) Model	42
3.4.1	Composite Penalty Minimization and ADM Model	42
3.4.2	Fast Numerical Algorithm	44
3.4.3	Convergence and Complexity Analysis	46
3.5	Experimental Results	47
3.5.1	Running Environments and Parameters for Competing Methods	48
3.5.2	Synthetic Studies by Comparing with CNV-TV	49
3.5.3	Comparative Studies on Simulated Data	49
3.5.4	Comparative Studies on Real Sequencing Data	52
3.5.5	Hyper-parameters Pruning	55
3.5.6	Computational Time of the CNV Detection Methods	57
3.6	Chapter Summary	57

Chapter 4 Revealing Common Copy Number Patterns by a Total-

variation Constrained Permutation Model	67
4.1 Introduction	67
4.2 Overview of Current Methodologies for Multiple Genome Samples . .	68
4.3 Problem Modelling for Revealing Common Copy Number Patterns in SCS Data	71
4.4 Numerical Solution for the TCP Model	72
4.4.1 Fast Numerical Solution by ADM	72
4.4.2 Convergence and Complexity Analysis	75
4.5 Experimental Results	76
4.5.1 Robustness and Accuracy Test over Noise Contaminations on Synthetic Data	77
4.5.2 Recovery Uniform Pattern Test over Synthetic Irrelevant Read Depth Signal	77
4.5.3 Comparison of TCP with Popular Methods on Synthetic Data	79
4.5.4 Experiment on Empirical Single-cell-sequencing Tumor Data .	83
4.6 Chapter Summary	84
Chapter 5 Conclusions and Future Work	86
5.1 Conclusions	86
5.2 Future Work	88
Bibliography	90
Curriculum Vitae	104