9-4-2017

Detection copy number variants profile by multiple constrained optimization

Yue Zhang

Follow this and additional works at: https://repository.hkbu.edu.hk/etd_oa

Recommended Citation
Zhang, Yue, "Detection copy number variants profile by multiple constrained optimization" (2017). Open Access Theses and Dissertations. 439.
https://repository.hkbu.edu.hk/etd_oa/439

This Thesis is brought to you for free and open access by the Electronic Theses and Dissertations at HKBU Institutional Repository. It has been accepted for inclusion in Open Access Theses and Dissertations by an authorized administrator of HKBU Institutional Repository. For more information, please contact repository@hkbu.edu.hk.
DATE: September 4, 2017

STUDENT'S NAME: ZHANG Yue

THESIS TITLE: Detecting Copy Number Variants Profile by Multiple Norms Constrained Optimization

This is to certify that the above student's thesis has been examined by the following panel members and has received full approval for acceptance in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairman: Prof. Leung Ho Cheung Clement
Professor, Computer Science and Technology
Division of Science and Technology, UIC

Internal Members: Dr. Zhang Hui
Associate Professor, Computer Science & Technology
Division of Science and Technology, UIC

Dr. Jeong Seon-Phil
Assistant Professor, Computer Science & Technology
Division of Science & Technology, UIC

External Members: Prof. Tan Kay Chen
Professor
Department of Computer Science
City University of Hong Kong

Prof. Li Guiqing
Professor
School of Computer Science & Engineering
South China University of Technology

In-attendance: Prof. Cheung Yiu Ming
Professor, Department of Computer Science, HKBU

Issued by Graduate School, HKBU
Detection Copy Number Variants Profile by Multiple Constrained Optimization

ZHANG Yue

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Principal Supervisor:
Dr. CHEUNG Yiu-ming (Hong Kong Baptist University)

September 2017
DECLARATION

I hereby declare that this thesis represents my own work which has been done after registration for the degree of PhD as appropriate at Hong Kong Baptist University, and has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree, diploma or other qualifications.

I have read the University’s current research ethics guidelines, and accept responsibility for the conduct of the procedures in accordance with the University’s Committee on the Use of Human & Animal Subjects in Teaching and Research (HASC). I have attempted to identify all the risks related to this research that may arise in conducting this research, obtained the relevant ethical and/or safety approval (where applicable), and acknowledged my obligations and the rights of the participants.

Signature: ______________
Date: September 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 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
Acknowledgements

Part of the this thesis has been published. In particularly, part of material in Chapter 3 has been published in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

I would like to take the opportunity to thank the people who have supported me through my Ph.D. study.

My deepest gratitude goes first and foremost to Prof. Cheung Yiu-ming, my principal thesis supervisor at Hong Kong Baptist University for leading me to the topics of this thesis at the beginning. His enthusiasm, inspiration and great efforts to explain things clearly have helped overpass many stones, for which I had thought been impossible for me to get over. With his strong support, warm encourage and keen consideration, I have gained valuable experiences. Throughout my thesis-writing, he has provided encouragement, sound advice and great patience to my work.

I would like to express my infinite gratitude to Prof. Su Weifeng, my co-supervisor for granted me a precious opportunity to purse Ph.D study in HKBU. His humor and strong determination deeply influence me. He also provided valuable advices for this thesis.

I also have learned a lot from the discussions with Dr. Gu Fangqing and Dr. Jia Hong. Additionally, I want to thank all my friends in Prof. Cheung’s Lab. They are young for the life, energetic for today, enthusiasm tomorrow and joyful for day and night.

Lastly, and most importantly, I wish to thank my parents, for their unwavering support during the many long days that went into this endeavor. To my beloved father, a true man of integrity, he supported our whole family with his strong shoulder and selfless. He always tolerated my ill temper and complaints. He passed away
before I am ready to take over the responsibility of my family. I missed him very much with tears and my heart. To him I dedicate this thesis.
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
### Chapter 3  Detecting Copy Number Variants from NGS by Both Sparse and Smooth Constraints Model

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-
## List of Tables

3.1 Number of true positive/false positive detections for the 30 trails in simulation experiment. ............................................. 52  
3.2 A summary of a Venn-style analysis five-way tabulated Venn diagram obtained from sample of NA19240. The integer 1 represents common CNV blocks that can be detected by the corresponding method, while 0 means it is not detected. The proposed ADM-CNV could detect as many as 24,564 blocks, all being marked in DGV. It ranked the second position among the seven methods. ................................. 61  
3.3 F-scores of top 10 CNVs detected by each method. For visual comparison, the best and the second best value were highlighted in red and blue, respectively. .................................................. 64  
3.4 The computational time of the proposed method and the other six competing CNV methods. ........................................... 65  
3.5 Average distribution (in percentage) of F-scores of detected CNVs. . 66
List of Figures

2.1 The preprocessing workflow for single-cell sequenced data include five steps, including genome segmentation, sequence alignment, duplicates removing, read depth estimation and bias correction. . . . . . . . . . 33

3.1 Experiment on simulated read ratios following a Poisson distribution with various mean values, followed by Gaussian noise degraded by various variances; The reads with Gaussian noise with (a) small value of variance (0.01); (b) intermediate value of variance (0.1); and (c) high variance (1) are shown in the first column. In comparison, the corresponding detected CNV curve are shown in the second column with the raw reads overlapped. The results by the proposed ADM-CNV is highlighted in black, while the one after CNV-TV is represented in red color. The proposed ADM-CNV can accurately detect all of the CNVs in different scenarios. . . . . . . . . . . . . . . . . . . . . 59

3.2 A sample CNV detection result with $d = 6e3$, $cp = 6$ by repeating the detection scheme 30 times. The horizontal and vertical axis represent the experiment index and the genomic coordinate, respectively. The fraction between the true positive ratio and false positive ratio is formatted as (TPR/FPR) and is shown in the parenthesis followed each method. The detected CNVs are highlighted in blue lines, while the ground truth, given by the starting and ending position with known CNVs at the reference genome is in red dots. . . . . . . . . . . . . . . . . . . . . . 60

3.3 Chromosome 21 from NA19240. The blue curve is the read depth signal. The CNV regions are shown in red dots, and the green lines indicate the bounds of the zoom region. A zoom between the region within the two vertical green lines is displayed in Fig. 3.4. . . . . . . . . . . . . . . . . . . . . . . . . . . 62
3.4 The zoomed CNVs detected by ADM-CNV. The gray lines are the raw depth of the coverage data while the blue solid line is the reconstructed CN. The two horizontal green lines represent the global cutoff values for the copy number variation. Therefore, the ratio beyond the cutoff values is recognized as the CNV and is highlighted in red.

3.5 False positive rates to detect copy-number alterations is considered as a function of the cut-off value. The red circle shows that when the cut-off value is 0.2, the genome-wide false positive rate is minimized. Thus, the cut-off value is set to be 0.2.

4.1 TCP analysis of a synthetic data set. A synthetic data set consisting of 100 samples is created, and Gaussian noise with zero mean value and (a) low, (b) intermediate, or (c) high variance is added to the data. The SNRs are (a) 3.57, (b) 2.82, and (c) 1.55. The raw observation is in the first row. The results of 100 and 300 iterations of TCP are shown in the second and third row, respectively, while the final result of TCP is shown in the last row.

4.2 Recovery ability of TCP analysis. A synthetic data set consisting of 100 samples is created, and (a) 100, (b) 1,000, and (c) 10,000 artificial-read-depth signals are randomly inserted into the data. The raw data are shown in the top row. In each case, four levels of degraded noise are introduced as indicated by the SNRs, and the data are analyzed by TCP to recover uniform block patterns. In most cases, the four patterns are largely revealed by TCP in the presence of low or intermediate noise levels.
4.3 Performance comparison of TCP with the other methods. TCP is compared with the Sparse NMF, Quick Sort and hierarchical clustering methods on synthetic data under contamination with (a) low noise, (b) intermediate noise, and (c) high noise, as demonstrated by the SNRs of the raw data. TCP shows a high level of robustness to noise degradation, recovering the four distinct patterns in the true signal under all three noise levels.

4.4 TCP analysis of empirical data. (a) The RD signals of 100 single-cell sequencing data; (b) result by quick sort scheme; (c) result after using the TCP to assemble similar samples at contingent positions; (d) result by hierarchical clustering. The data contains three sub-populations of (D) diploids or pseudodiploids; (A) aneuploids; and (H) hypodiploids. The result after TCP produces three blocks and is consistent with the previous characterization of the cells by FACS analysis [59]. From the top to the bottom, the ploidy of the clustered subpopulations in (d) is A, H, D, respectively. Five of the diploids or pseudodiploids are mis-classified to the hypodiploids by HC, as highlighted in red arrow.
Chapter 1

Introduction

The Human Genome Project (HGP) was first proposed in 1985 and launched in 1990. Scientists from Britain and America jointly participated in the HGP with a budget of approximately US$3 billion. Since this epoch-making scientific project was completed, it has greatly promoted the development of genomics. The HGP was a landmark in revealing the mysteries of the human genome. This project not only provided a complete reference genome sequence, but also boosted the development of high-throughput sequencing technologies. Researchers are now able to perform detailed and comprehensive analyses of the genome of different species at a cost of less than US$1000 and with a sequencing time of weeks. Human genome sequencing has thus now entered the era of high throughput and low cost.

Substantial in-depth sequencing data have been provided through the constant improvements in sequencing technology developed within a short period of time. In this context, it has become increasingly urgent to achieve efficient and accurate analyses of such data as well as to apply the associated findings in the field of human health. For example, scientists are attempting to identify molecular biomarkers, which are vital for connecting genome sequencing to improvements in human health. In 2008, the 1000 Genomes Project was also launched by a large international research team led by Shenzhen Huada Gene Research Institute, National Human Genome Research Institute, and Sanger Institute. This project involved sequencing of the genomes of at least 1000 people from all over the world to explore the relationship between genes and human diseases. The large-scale and globally shared dataset is intended to help drive medical advancements to develop in the direction of accuracy and personalization. In this project, about 3.8 million single-nucleotide variations (SNVs) and at least 14,000 structural variations were identified. The detected average length of structural variations was 8 Kb, approximately four times
the length detected by microarray technology [56]. SNV refers to the mutation of a single base, which is very common in the human genome, with the frequency of variation being identified to exceed 1%. However, despite this large number of single-nucleotide variations, it has not been possible to fully reveal the causes of human genetic lesions [66], and such data can only provide some guidance for disease research on a small scale.

Structural variation is also present in the genome, which includes copy number variation (CNV, greater than 1 KB), short insertions/deletions (indels), inversions, and ectopia. Each unit of a reference gene fragment can produce 1.2% CNVs and 0.1% SNVs, which suggests that CNVs account for a larger proportion of genomic variation. Nevertheless, in terms of the entire number of variations across the genome, that of CNVs is far less than the number (about 12 million) of single-nucleotide variations. However, CNVs cover a far wider range (total coverage length throughout the chromosome of about 150 MB) than any other genetic marker. Moreover, the distribution of CNVs in the entire human genome is clearly nonrandom, and has close connections with mitosis, exons, and movable components (such as Alu repetitive sequences).

Research in many fields such as genetic disease has indicated that the structural variation of the genome leads to complex diseases such as cancer. As an important type of structural variation of the human genome [99], it has been increasingly understood that CNVs have an intimate relationship with complex human diseases, and the research achievements in this field have also emerged in an endless stream. In 2008, a research team led by Steven carried out intensive research on Crohn’s disease and the IRGM gene [52] and found that about 20,000 base pairs upstream of the gene contained some defect resulting in the decline of patients’ immunity against invasive bacteria, finally leading to occurrence of the disease. In 2009, the relationship between CNVs and body mass index (BMI) was also explored [95]. Common variations in FTO and MC4R2 gene fragments were found to have a significant correlation with BMI, and the CNVs affect the expression of genes associated with
hypothalamic nerves that control hunger and metabolism, eventually leading to obesity. Such findings demonstrate the close relationship between CNV and human diseases. CNVs also influence human phenotypes by affecting gene expression, and then exert great influence on the genetic variation and evolution of humans.

Research on CNVs in the genome has also provided a wealth of information about human evolution. Thus, rapid and precise detection of CNVs in the human genome from high-throughput sequencing data has been of great significance for research on both human health and human evolution. Against this background, this thesis is firstly focused on the calling of CNVs in data obtained by next-generation sequencing. We model CNV calling as a quadratic minimization problem, constrained by multiple norms. We test the performance of the proposed model through extensive experiments, and demonstrate its superiority over several popular models. Besides, the success of this model will be of high theoretical significance. Quadratic minimization models with multiple norm constraints are widely applied in signal processing, image processing, computer vision, and other fields. The successful implementation of this model will not only promote the development of such fields, but also further provide biological theory and algorithm tools for analyzing the structural variations of the genome.

Once CNV profiles for multiple samples have been obtained, it is natural to simultaneously analyze the genomic data to assemble and subgroup similar cells together accurately and efficiently. This subgrouping facilitates the exploration of heterogeneity among single cells, providing important information in relation to cellular evolution. With the help of recent single-cell sequencing technology, it is now possible explore gene-expression heterogeneity among single cells, and thus provide important information on cell evolution. However, single-cell DNA sequencing data usually have low amount of starting genome, which requires an extra step of amplification to accumulate sufficient samples, introducing noise and making regular pattern-finding challenging. To directly achieve subgrouping by finding common patterns within alignments of sequencing data from multiple individual cells that
might otherwise be obscured by noise, this thesis further investigate the problem of finding common CNV patterns hidden in multiple sequenced data. To achieve this goal, this thesis propose to tackle this issue of recovering the hidden blocks within single-cell DNA-sequencing data through continuous sample permutations such that similar samples are positioned adjacently. The proposed model involves continuous permutation of the data to achieve a pattern in which similar samples are organized into clusters and named as total-variation constrained permutation (TCP) model. The permutation is guided by the total variational norm of the recovered copy number profiles, and is continued until the total variational norm is minimized when similar samples are stacked together to reveal block patterns. 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. The results reported in this thesis not only contribute to the development of sequencing technologies and related analytical tools, but also show the promise of the model for widespread use in innovative drug development, cancer prediction, and precision medicine.

1.1 Organization of the Dissertation

This thesis is divided into five chapters as described below:

1.1.1 Introduction and Organization

Chapter 1 briefly introduces the field of genomics, summarizes current research activities, and underlines the significance of various structural variations. The organization of this thesis is described.
1.1.2 Next-generation Sequencing and Preprocessing Workflows

Chapter 2 introduces several key technologies for this thesis. It first introduces next-generation sequencing, including its principles, developments, and applications to investigating the structural variation of the human genome. It then introduces copy number variation, including its basic concept, research status, and some existing methods of detecting it. Then, this chapter introduces the schemes for processing sequencing data. The popular CNV calling schemes with respect to various sequencing methods are also summarized and categorized. Finally, we establish a pipeline for the preprocessing of both bulk and single-cell sequencing data.

Within the traditional bulk sequencing area, single-nucleotide and copy number variation are two hot topics in studies of genetic variation. Studies of genetic loci associated with diseases have continuously deepened our understanding of disease, but the association analysis between single-nucleotide genome-wide and common variations still hinders the further development of disease research. Most genetic diseases are closely associated with rare variations, indicating that rare variations play a prominent role in many phenotypes. Recent studies have also suggested that CNVs are more closely related to many genetic diseases than single-nucleotide variations are. The average length of CNVs is 456 KB [73] and they are often located in gene duplication fragments, such as in centromeres and heterogeneous points. CNVs have also been shown to be closely associated with complex diseases, such as autism [68]. By studying them, it is hoped that we can increase our understanding of human disease-related mechanisms, guide the genetic diagnosis of diseases, and finally improve bio-medicine.

The advent of single-cell sequencing, involving sequencing the DNA of a single cell by amplifying the trace amount of DNA contained in it, allows scientists to approach complicated issues such as growth, cancer, and neuroscience from a new perspective. However, single-cell genomic and transcriptomic sequencing requires
more DNA than that contained in a single cell, leading to significant challenges associated with nucleotide amplification. Owing to the low DNA content of cells for use as templates, damage, contamination, or degradation of any sample in the sequencing process will exert a major influence on the results. Therefore, efficient and accurate experimental amplification techniques are needed to sequence a single cell, which will be elaborated on below. To compensate for the bias introduced by experimental amplification, computational methods for calling single-nucleotide and copy number variation are urgently needed for single-cell sequencing data.

Traditional CNV detection methods are mainly designed for bulk sequencing data. Therefore, most of the popular CNV detection methods are for massive high-throughput data. The current CNV detection methods are mainly divided into two types, pair-end mapping (PEM) and depth of coverage (DOC) [32]. DOC is based on next-generation sequencing technology and is more commonly used. It can reconstruct copy number signals, restore the initial copy numbers, and provide relevant information of CNVs. In comparison, PEM-based detection methods can detect the indels, inversions, and other abnormal conditions of gene fragments. Popular DOC methods include CNV-TV [21], CNV-seq [96], readDepth [57], CNVnator [2], SegSeq [16], FREEC [8], and the time-based intelligent detection method EWT [101]. The basic principles of DOC-based algorithms involve first aligning short sequence samples to the human reference genome and then estimating the copy number variations using different model assumptions.

In comparing, the PEM-based sequencers include a pair of gene sequences. Information on the copy number can be obtained using the alignment results of this pair of sequences against the reference genome. The calling of CNVs is achieved by analyzing the discordant PEM, in which the alignment of the pair of sequences differs. The popular calling methods include Hydra [72], GASV [86], BreakDancer [12], and genome STRiP [28]. However, owing to problems such as alignment bias, amplification bias, and sequencing bias in the sequencing process, the detection results of this algorithm still have a false-positive rate.
To facilitate sequencing data analysis, we built two pipelines to deal with both next-generation and single-cell sequencing data.

1.1.3 Detecting Copy Number Variants from NGS by Both Sparse and Smooth Constraint Model

Copy number variations are a common type of structural variation. Recent studies aimed at uncovering the intricacies within the human genome have revealed CNVs to be potential candidates for explaining various phenotypic differences and genetic diseases [25], as well as other complex diseases including autism, schizophrenia, osteoporosis, and certain tumor types [43, 88, 81]. Next-generation sequencing has been adopted as a popular strategy for genotyping, which has enabled the comprehensive characterization of CNVs by generating hundreds of millions of short reads in a single run [56].

The depth of coverage [DOC, also known as read-depth (RD) method] [101, 19] method is one of the major. There are multiple schemes for obtaining copy number profiles from sequenced data. These involve counting the number of reads that fall in each prespecified window of a certain size [2, 96, 101]. CNV calling is achieved by detecting deviations between the number of reads mapped to a chromosome and its expected value.

Most existing tools for CNV calling based on read depth can be broadly divided into two categories: parametric versus nonparametric schemes. The performance of parametric approaches heavily depends on the underlying statistical distribution. For example, Magi [51] used the stochastic process of a shifting level model to simulate multisequential samples. ExomeCNV [79] and CNV-seq [96] assume a Gaussian distribution of the read ratios. They also assume that the proportion of reads matching a specific sample usually follows a binomial distribution whose success rate is determined by the genome-wide read count ratio between the test sample and the reference set, as well as the potential presence of CNVs. However, it has been shown that the binomial assumption is actually violated in practice.
due to technical variability, noise induced by library preparation, and capturing or sequencing error. ExomeDepth [69] uses a beta-binomial model to approximate the distribution of the read count ratio, and then employs statistical inference to detect the CNVs. CLImAT [102] takes tumor impurity and ploidy into consideration for identifying genomic aberrations. DeAnnCNV [107] is an online integrated tool to precisely detect and systematically annotate copy number variations from whole-exome sequencing (WES) data. Hidden Markov models (HMMs) and many variations thereof [93] have dominated the parametric approach. The unique characteristics of HMMs lie in their flexibility for handling several common complications, including variable single-nucleotide polymorphism (SNP) frequencies [48], variable distances between adjacent SNPs, linkage disequilibrium, and relationships between study subjects.

Nonparametric methods attempt to reconstruct the CNV-associated copy number from the observed read depth signal in concert with the application of subsequent statistical inference on the estimated signal to detect the CNVs. For simplicity, let us consider a plateau/basin in the reconstructed signal to be a duplication/deletion event. Therefore, the CN is assumed to be a piecewise constant function with two fundamental characteristics: sparsity because of few CNVs and smoothness because of contingency positions possessing similar CNs. In such a configuration, CNV detection is considered to be a piecewise linear signal recovery problem. A classical approach for addressing this is by the fused lasso scheme [90] with its refinements [75]. It was also tailored for use in CNV detection on array CGH [91]. Recently, the conditions in which the fused lasso could consistently recover a piecewise constant pattern have been investigated [71]. A numerical scheme for solving fused lasso-type problems includes an approximation method of path coordinate descendent optimization [4], and a linearization method within an ADM framework [7, 19].

However, the empirical sequencing process itself inevitably introduces certain biases, such as nonuniform genome content and low mappability ratios. Therefore, the observed copy numbers of the genome appear to have spurious signals due to over-
or undersampling. While a number of successful approaches have been presented for copy number calling, there is still a paucity of methods for analyzing it through a systematic framework to achieve high accuracy and good robustness.

To address this problem, Chapter 3 reports our proposed CNV detection model, called ADM-CNV. The proposed ADM-CNV formulates the RD signal reconstruction problem into a quadratic minimization problem involving two constraints. The numerical solution is obtained within the popular alternative direction minimization framework. Our proposed alternating direction minimization-total variational (ADM-CNV) method offers two main contributions as opposed to the previous methods: 1) The RD signal reconstruction is characterized by a well-designed model to reveal intrinsic structure, thus facilitating subsequent statistical testing for CNV detection; and 2) an efficient numerical method using a classical alternating direction minimization framework is tailored to solve the ADM-CNV model. The key characteristic of the ADM method that makes it different from standard methods in solving TV-related minimization problems is that it does not require matrix inversion, thus greatly reducing computational cost. This is particularly useful in NGS data analysis, in which hundreds of thousands of signals need to be recovered. Extensive experiments were conducted to demonstrate the performance of the proposed model on both synthetic and empirical sequencing data. The superior performance of the proposed method suggests its potential value in the calling of CNV from NGS data.

1.1.4 Revealing Common Copy Number Patterns by a Total-variation Constrained Permutation Model

The rapid development of single-cell sequencing technology has enabled exploration of gene expression heterogeneity among single cells, providing important information on cell evolution [59, 60]. Single-cell sequencing combines flow sorting of single cells, whole-genome amplification, and next-generation sequencing to characterize the genome-wide copy number in single cells. Existing whole-genome amplification (W-
GA) techniques, such as degenerate-oligonucleotide-primed polymerase chain reaction [94], multiple-displacement amplification [42], and multiple-annealing looping-based amplification cycling [115] inevitably introduce varying degrees of amplification bias when the whole genome of a single cell is amplified to microgram levels for next-generation sequencing [29, 84]. Besides, copy number profile detection requires only sparse sequence coverage [6] and it contributes to the intrinsic noise associated with single-cell sequencing data. Technical noise that results from amplification bias is over-dispersed compared with Gaussian noise, and differs from the noise that occurs in bulk sequencing, which does not involve amplification.

As single-cell-sequencing data accumulate, recent efforts have focused on simultaneously analyzing genome data from multiple samples to assemble and subgroup similar cells accurately and efficiently. This subgrouping will be helpful for the exploration of heterogeneity among single cells, providing important information on cellular evolution. The simplest way to subgroup multiple samples into a block pattern is to sort their reads in ascending or descending order. However, the high levels of noise in single-cell sequencing data negatively affect the reliability of sequence reads and lead to inaccurate patterns of variation. Clustering is of the first choice by categorizing the expression data into sub-groups [70, 24]. Matrix factorization is another most widely used techniques for finding a suitable representation of data in real applications by decomposing the expression measurements with linear basis matrix and its loading coefficients. Ideally, this representation reveals the latent structure in the data, and reduces the dimensionality of the data so that further computational methods can be applied [33]. A variety of methods have been developed to reveal hidden structures in observed data [74, 100], such as social network clustering [97, 39] and mining of protein or gene interactions [104, 85]. However, matrix factorization suffers from two drawbacks. First, finding an accurate linear basis and its spanning dimensions remains a challenging problem. Second, matrix factorization has high computational cost, which limits its application to high-dimensional data analysis.
An alternative way to subgroup multiple samples into a block pattern is to obtain their copy number profiles and then recover hidden block patterns [105]. The copy number profiles could be obtained by the popular CNV detection method, or by the ADM-CNVC proposed in Chapter 3. Thus, the copy number calling scheme could be viewed as a preprocessing step. The derived copy number profiles are then used to find the common block patterns within alignments of sequencing data from multiple single cells that might otherwise be obscured by noise. However, directly using the copy number profile is prone to being affected by the bias introduced by each individual cell. It is therefore difficult to find common patterns for multiple cells.

In Chapter 4, we propose tackling this issue of recovering the hidden blocks within single-cell DNA sequencing data through continuous sample permutations, such that similar samples are positioned adjacently. The permutation is guided by the total variational norm of the recovered copy number profiles, and is continued until the total variational norm is minimized when similar samples are stacked together to reveal block patterns. This model is named the total-variation constrained permutation (TCP) model. In TCP, the permutation is not stopped until the total variational norm of the recovered matrix is minimized, when similar samples are stacked together to reveal block structures. An efficient numerical scheme for finding this permutation is designed and 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.

1.1.5 Conclusions and Future Work

Chapter 5 concludes this thesis by summarizing the major contributions of this study and acknowledging several drawbacks within current work. Possible future research directions associated with the work is then presented in this chapter.
Chapter 2

Next-generation Sequencing and Preprocessing Workflow

2.1 Development of Sequencing Technologies

With the ongoing development of sequencing technologies, the cost, precision, and speed of sequencing have significantly improved. The development of this technology enables researchers all over the world to carry out large-scale gene sequencing on different individuals, promoting our knowledge of single-nucleotide variation and CNVs, and improving whole-genome sequencing, single-cell sequencing, and other fields, as well as laying a solid foundation for the vigorous development of bioinformatics analysis.

In the 1970s, Sanger proposed a DNA sequencing method called the chain termination method, which was performed completely manually. At that time, automated sequencers were not available, so the sequencing technology proposed by Sanger greatly accelerated the development of DNA sequencing; for his efforts, Sanger won the Nobel Prize in Chemistry in 1980 [78]. This method involved the use of a radioisotope to image the base sequence, and featured four lanes to detect the products of the four chain termination reactions. However, the reaction time was rather long, a lot of reagent was used up, and the sequencing throughput was also seriously limited. Therefore, only by developing nonradioactive high-throughput sequencing technologies could the growing demand for sequencing be met. A few years later, another research team made some modifications to the conventional Sanger sequencing method. They adopted fluorescent labeling to replace the isotope labeling, so that the four labeled DNA terminators could undergo chain termination reactions in the same reaction vessel [87], greatly shortening the reaction time. In the second year, by further improving this method, this research team achieved semiautoma-
ic sequencing technology and developed the first automated sequencer, ABI 370, which freed the sequencing from cumbersome manual work and marked the beginning of a new era of automated sequencing. The capillary electrophoresis method of the sequencer ABI 370 replaced the conventional plate electrophoresis separation method, greatly reducing reagent consumption. With sequencers based on capillary electrophoresis technology, parallel operation could easily be achieved, greatly increasing the sequencing throughput. The high concurrent sequencing technology allowing four terminators to react at the same time greatly improved the sequencing speed and also enabled the sequencing of several samples simultaneously. The invention of this automated sequencer not only substantially enhanced the speed and quality of sequencing, but also laid a solid foundation for subsequent completion of the Human Genome Project.

Despite the rapid development of first-generation sequencing technology, both the cost and the precision of sequencing soon reached certain limits. Although the average cost of unit base sequencing was only $0.0005 and the sequencing bias was merely 0.001%, owing to its overdependence on electrophoresis separation technology, the first-generation sequencing technology failed to make further progress in terms of high parallelization and low cost. For example, the Human Genome Project cost about US$300 million and took years to complete the sequencing of approximately 3 billion base pairs by using conventional sequencing methods. The associated urgent demand to explore the human genome quickly and cheaply boosted the rapid development of sequencing technologies, which gave birth to next-generation sequencing.

Next-generation sequencing works largely based on circular array synthesis technology, and its core idea is to combine both synthesis and sequencing. It determines the labeling of a newly synthesized end to determine the corresponding base arrangement. Its main operation steps include: the preparation of a DNA template library, the fixation of DNA fragments, the unimolecular amplification of DNA fragments, the parallel sequencing reaction, the collection and processing of optical images,
and sequence splicing and assembly. Although there are different next-generation sequencing technologies with different operating steps of library preparation and DNA molecular amplification, most of their principles are similar. This method configures ordered or unordered DNA sample arrays by a large-scale microarray analytical technique, improving the degree of parallelization of the reaction and the intensity of information, and thus increasing the sequencing throughput. In addition, the introduced miniaturized equipment avoids the strong reliance on electrophoresis and reduces reagent consumption. Moreover, this technology can process several hundred gigabases of DNA data within a few weeks, so as to obtain large amounts of gene sequencing data, effectively control the sequencing costs, and promote the rapid development of genomics and bioinformatics. The main next-generation sequencers used in the USA include the 454 Genomic Sequencer of Roche, SOLiD Sequencer of ABI, and Solexa Genetic Analyzer of Illumina. Different sequencers have different merits in terms of their sequencing principles, but adopt similar strategies developed based on cyclic chip sequencing method. This method refers to repeated performance of the polymerization reaction, such as deformation, annealing, and extension, on a DNA chip of a fixed size. Next-generation sequencing technology greatly reduces time and cost, making large-scale parallel sequencing possible. Compared with that in conventional sequencing methods, the throughput is markedly improved in next-generation sequencing technology, producing substantial sequencing data. However, this approach also has some disadvantages, such as producing smaller gene fragments. For instance, Illumina platforms generally produce read fragments with a length of 36 bp, raising difficulties for the subsequent genetic analysis, such as gene annotation, sequence splicing, and gene assembly. Second, although the next-generation sequencer based on PCR amplification can take advantage of trace amounts of DNA, the bias in the amplification process causes excessive amplification of some base areas, resulting in bias.

The aforementioned next-generation sequencing technologies determined the base types based on the signals released by fluorescence or other luminescent materials
during DNA polymerase reactions. However, owing to the expensive optical image acquisition and processing system needed in the detection procedure and the overreliance on biochemical reactions, the cost could not be further reduced. The development of next-generation sequencing technology had thus reached a bottleneck. In 2007, [47] reported a sequencing method based on nanopore technology. By randomly segmenting the target genome into read fragments of 100 KB, it is reported to achieve DNA sequencing at the single-molecule level. The sequencing accuracy of this method is even comparable to that of the first-generation sequencing technology. Through this development, the sequencing speed of a single base was reduced to 0.1 s, which is 20,000 times faster than that of the chemical sequencing method. Its sequence length could be above 10,000, greatly increasing the sequencing speed and the sequence length in a single detection. However, many assumptions about this method remain to be verified, and this technology is still at an immature stage. Thus, next-generation sequencing technology is still dominant.

In summary, the continuing innovation and development in science and technology has brought about a new era of genome sequencing. Human sequencing technologies have developed from the original purely manual operation to the current fully automated approaches. Sequencing costs have been further reduced and sequencing throughput has grown exponentially, making it possible to sequence a whole genome for just US$1000. The rapid development of next-generation sequencing technology has laid a solid foundation for the study of human diseases and genetic variation, and promoted the revolutions in medical sequencing and personalized medicine.

2.2 Variation Calling for Next-generation Sequencing Data

The Human Genome Project (HGP) not only deepened our understanding of genome sequencing, but also promoted the development of sequencing technologies. These pioneering methods have fundamentally changed the strategies adopted by researcher-
s to study problems. The next-generation high-throughput sequencing technologies have triggered the equivalent of a new industrial revolution and successfully replaced the microarray-based sequencing methods. The higher reaction speed in the detection process has provided researchers with a more in-depth understanding of genome regulation. Through massive high-throughput sequencing data, we can also now accurately detect the drug resistance of viruses and bacteria, reveal the genes of unknown species, understand nucleic acid treatment, complete related biological information at the individual level, and improve the existing medical services from the gene level. The development of next-generation sequencing technology also revealed a closed connection between genome variations and diseases, as described below:

### 2.2.1 Single-nucleotide Variation

Single-nucleotide polymorphism refers to an insertion, deletion, or change of a single base in the genome; such polymorphisms are widely distributed throughout the genome. Initially, researchers used gene polymorphism fragments with a finite length as molecular biomarkers. Success in searching for factors that play a major role in the development of human genetic diseases depends largely on having good molecular biomarkers [37]. Initially, researchers tended to use single-nucleotide variation as a new molecular biomarker. High resolution and gene mapping describing common single-nucleotide variations can effectively promote the research of high-throughput sequencing fragments. Studies of genetic loci associated with diseases have deepened our understanding of such diseases, but the association analysis between single-nucleotide genome-wide and common variations still hinders the further development of disease research. Only a small number of common genetic diseases can be explained by their association with common variations. Most genetic diseases are instead closely associated with rare variations, indicating that rare variations play a prominent role in many phenotypes. The continuous development of high-throughput sequencing technologies should greatly promote research on the association between rare genetic variations and phenotypes.
2.2.2 Copy Number Variable Regions

Copy number variation refers to large-scale gene rearrangements, such as inversions, insertions, and deletions, which can even involve whole chromosomes (aneuploidy) or gene fragments of megabases in size. In the early stage of genetic research, genome structural variations were considered to be related to human diseases. It is worth noting that, in previous research, it was deemed that only insertions and deletions of chromosomes would lead to cancer and developmental defects. Until recently, with the development of high-resolution whole-genome sequencing technology, some undiscovered submicroscopic variations have been detected in the genomes of both patients and healthy individuals. Subsequently, CNVs also came to be regarded as specific molecular biomarkers. The advent of sequencing technology manifested that CNVs are also present in healthy individuals. Initially, CNVs were considered to have benign or unknown clinical significance. However, in-depth studies gradually suggested that CNVs are even more closely related to many genetic diseases than single-nucleotide variations are. The average length of CNVs is 456 KB [73, 106], and they are often located in gene duplication fragments, such as in centromeres and heterogeneous points. CNVs are closely associated not only with genetic diseases and rare diseases, but also with complex diseases, such as autism [68]. Therefore, the studies of CNVs promote research on human disease-related mechanisms, guide the genetic diagnosis of diseases, and improve medical treatments.

2.2.3 Single-cell Sequencing

Single-cell sequencing involves the sequencing of cells by amplifying the trace amount of DNA contained in a single cell. Its main steps are as follows: First, appropriate cells are selected by screening cells using a flow cytometer, as described in [59]. Second, the original RNA molecules are captured by reverse transcription of the selected cells. Finally, the products are amplified, and combine random primers with specific genomes by the specific polymerase to form a duplicated branch structure
and further obtain Klenow fragment DNA products. In recent years, an increasing number of articles related to single-cell sequencing have been published. Since its invention in 2010, this method was highlighted as the most important methodology of the year in *Nature Methods* in 2014. Single-cell genomic and transcriptomic sequencing requires more DNA than contained in a single cell itself, bringing significant challenges regarding the need for nucleotide amplification. Owing to the low level of DNA for use as a template, damage, contamination, or degradation of any sample in the sequencing process will exert a major influence on the results. Therefore, efficient and accurate amplification techniques are needed to sequence a single cell, as elaborated below. Single-cell sequencing technology can be used to study individual organisms and complex diseases at the single-nucleotide level, and thus gains further development and extensive application.

Single-cell sequencing technology leads to a boom in bioinformatics, allowing scientists to approach challenging issues such as growth, cancer, and neuroscience from a new perspective. The first application of SNS technology to study breast cell heterogeneity was reported [59]. To mitigate the bias that arises during the amplification stage of SNS, an original amplification scheme was reported [115]. Based on the properties of cells and the characteristics of surface markers, Navin et al. put cells into a cell separometer by a streaming classification method, and obtained single or multiple cells. With its simple operation, this method enables single cells to be obtained with high precision, but is associated with problems such as sample pollution and a decline of cell activity. The DNA is extracted and separated from a single cell that has been isolated from the cell line. Since a single cell contains a small amount of DNA, the above-mentioned methods are used to amplify the small amount of DNA and obtain sufficient DNA for sequencing. Two types of single cell were used in the experiment. The first type was polygenic tumor cells composed of tumor cells with different genotypes, for which the evolutionary process and gene copy number were unknown. The other type was monogenic tumor cells, which evolved from a single tumor cell and had the same genotype and copy
number, and the tumor had spread to the liver. These two types of tumor cell were separated by the described steps to finally obtain 100 single cells. Then, a single-cell sequencing analysis algorithm was used to detect the CNVs in the corresponding cells, and massive genome deletions, duplications, and mutations [59] were found (detection resolution can reach up to 50 kb). The results show that single cells of a polygenic tumor can be divided into three different subtypes, and these cells have different gene expression. In contrast, the monogenic tumor cells can only be obtained from the primary tumor cells by monoclonal expression. This experiment reveals the process by which a minority of diseased cells evolve into a complex mixture, which deepens our understanding of tumor cell evolution. In 2012, [115] developed MALBAC to preamplify single cells for five rounds; here, the amplified products are formed into closed ring molecules. Since the ring molecules cannot be further amplified, this proceeds into linear amplification. Then, normal PCR amplification is conducted. The adopted template is more homogeneous, so it is unlikely that large bias arises in the amplification process. The amplification products of the human genome obtained in this process have a coverage degree of up to 93%, and a large number of CNVs are also detected. Research work has achieved sequencing of the whole genome of sperm, successfully built a personal genetic map with a high degree of coverage, and greatly enriched and promoted research on embryological genetics.

In summary, next-generation sequencing technology can be used to sequence the whole genome, thus bringing about a new era in the field of genomics, and greatly promoting the development and clinical application of molecular biology. With the constant advancement of science and technology, sequencing technology should develop towards alternatives that are cheaper, faster, more efficient, and have a higher degree of coverage; guide human healthcare towards treatments based on genetic information; and further promote healthcare to develop in the direction of highly precise and individualized treatments.
2.3 Copy Number Variation

Genetic diversity is generally derived from CNVs, especially the duplication of CNVs, which exerts an even more significant influence on the genome than their loss. Although the frequency of CNVs occurring in the human genome is low, they directly influence the differences between individuals. In the early days, scientists studied and analyzed the genome mainly by karyotype analysis, chromosome imprinting, and other relatively blunt instruments. For instance, conventional Sanger sequencing technology was used to detect the chromosome variation in patients with lymphoma and leukemia, and also to detect the serious variation in such patients. With the development of sequencing technologies, scientists have increasingly recognized that CNVs have a close association with human diseases and genetic variation. Therefore, a highly efficient and accurate model needs to be established for comprehensive and in-depth research of massive high-throughput sequencing data.

2.3.1 Definition of Copy Number Variation

Research has shown that CNVs are common in the human genome, and are an important source of human genetic diversity. As a kind of genome structural variation, the term “CNV” generally refers to the structural variation involving the duplication or deletion of DNA fragments of more than 1 kb in size, and is mainly caused by genomic rearrangement. For example, an A-B-C-D chromosome may be mutated into an A-B-C-C-D (duplication of C) chromosome or an A-B-D (deletion of C) chromosome. The frequency of CNVs in humans is relatively high (over 10%), and the average size of CNVs in an individual genome is about 3-4 MB [34]. CNVs have been linked to complex diseases such as cancer [11], schizophrenia [88], and Alzheimer’s disease [77].
2.3.2 Current Studies on Copy Number Variation

CNV was discovered by researchers studying fruit flies in the 1970s. It was widely believed that changes in genetic structure affect gene expression, leading to the different expression of genes among individuals. Different CNVs will change the structure of a gene and then affect the organic functions.

With the continuous development of sequencing technologies in the 21st century, we have developed a deeper understanding of CNVs. In 2004, two independent laboratories announced at almost the same time that DNA fragments (CNVs) from 1 kb to several Mb existed in the human genome. In addition, in 2006, a number of laboratories overcome the obstacles associated with human genome mapping and establish the first map of the human genome [73]. The map of CNVs in the human genome revealed the universality and richness of genome variation, and was published in the journal Nature, leading to a boom of CNV research. It was asserted that CNVs would become the gold standard of research on all phenotypic variations (including disease susceptibility). Initially, it was generally accepted that CNVs were only present in patients. However, with the intensification of CNV research, numerous CNVs were also subsequently detected in healthy individuals. CNVs in healthy individuals are easily confused with those in patients, hindering the detection of CNVs that cause serious and rare diseases. In 2009, in the DECIPHER [26] project, CNV information about a few severe and rare diseases was found, greatly promoting the development of research on copy number variation. A map containing a large number of candidate recessive alleles was also published, enabling researchers to choose the best CNVs on which to perform in-depth research based on the published datasets. With the improvement of CNV detection and the intensification of research on the mechanisms behind such variation, research related to CNVs has also gradually penetrated into various fields. Relevant applications are introduced below in a subject-specific manner.
**CNV and Cancer**

Compared with SNPs, CNVs can influence more gene fragments, and thus exert a greater influence on human diseases. Researchers generally believed that there is a close relationship between CNVs and cancer, but this had not been scientifically validated. It was not until 2004 that Nature published a paper [17] proving the close link between CNVs and neuroblastoma in children. Sharp et al. performed a comparative analysis of the CNVs from 846 neuroblastoma patients and 803 healthy individuals and found that a common CNV located at q21.1 of chromosome 1 was closely associated with neuroblastoma. This CNV was found to be produced during the polymerase reaction, which was confirmed by tumor hybridization and fluorescence pair analyses. Part of the base sequence in the CNV was shown to share significant similarities with the genes of neuroblastoma, and to play a significant role in fetal sympathetic nerve and brain development. Moreover, fetal phenotype has a direct correlation with the degree of copy number variation. Actually, there has been no success in specifically clarifying the relationship between CNVs and cancer, such as how the degree and frequency of base variation affect the transcription level of genes, which in turn affect gene expression and fetal phenotypes. Nevertheless, research has provided strong evidence of the correlation between CNVs and cancers, leading to a boom in the study of cancers at the molecular level.

**CNV and Schizophrenia**

Researchers have also found that CNV is closely linked to schizophrenia. Schizophrenia is a common and serious mental illness, for which the global incidence may be as high as 1%. Its main symptom is that one’s personality, emotion, and mentality cannot coordinate with the surrounding environment. It features a series of symptoms grouped together into the same mental illness, instead of being a single independent disease, and is mainly divided into two types. The main clinical manifestations of the first type of schizophrenia include anxiety disorders, acousma, confused thinking, delusions, and obsessive-compulsive disorder. The main clinical manifestations
of the other type include hypobulia, depression, pessimism, and frustration. Clinical studies have shown that only 20% of patients will fully recover after the initial onset of the disease. Instead, for most patients, schizophrenia is a lifelong condition, and about 10% of the patients with depression even commit suicide to put an end to their suffering. Inevitably, this also confers a heavy burden on their families, hospitals, and society as a whole. The causes of schizophrenia are both environmental and genetic. Sullivan et al. proved that the effect of hereditary on schizophrenia accounted for 81% (95% confidence interval of 73% – 90%), as determined by carrying out experiments on several groups of twins [89]. This research enabled exploration of the pathogenesis and evolutionary process of such diseases at the genetic level, which heralded a new dawn for the clinical treatment of schizophrenia. In 2008, Sharp et al. [83] also studied the correlation between rare CNVs and schizophrenia. They found that these rare genetic variants were mainly distributed on chromosome 1 and chromosome 15, and these CNVs were not present in healthy individuals but frequently occurred in the genomes of patients. In addition, in 2009, Need et al. [61] further studied the influence of SNPs and CNVs on patients with schizophrenia, and found a close relationship between rare CNVs and this disease, whereas mutual CNVs in human body had little relationship with patients. The findings of this research greatly promoted the study of mental illnesses, and laid a solid foundation for the precise treatment of patients with schizophrenia.

**CNV and Genome-wide Association Analysis**

With the intensification of CNV research, experts revealed the close relationship between complex diseases and genome-wide CNVs. With the advancement of CNV genome-wide association analysis, its range of applications has also gradually expanded. Genome-wide association studies (GWAS) involve investigation of the relationship between complex gene mutations and human diseases by searching for abnormal sequences across the human genome [44]. In 2005, Nature Reviews Genetics published a paper on GWAS [31], which described a study of the connection
between genetic susceptibility and complex characters. With the vigorous development of this analytical approach, more than 1000 related research papers had been published by the end of 2010, particularly reporting on more than 120 characters, and covering the research fields of wildlife genome-wide CNVs and association analyses of human genome-wide CNVs and some characters, obesity, and schizophrenia. GWAS was also extensively applied in the medical field for the first time. In this field, diseases can be divided into monogenic and polygenic types. Monogenic diseases are those in which the mutation of a single gene yields a change in an amino acid that comprises a protein, leading to a change in the character of the human body, in accordance with Mendel’s law of inheritance. Polygenic diseases refer to that the individual gene sequence is changed under the influence of both environmental and genetic factors. Researchers have revealed the mechanisms of genetic variation of many complex diseases through GWAS. For example, Lee [44] et al. performed in-depth analysis on the association between CNVs of psychiatric patients on a genome-wide scale in 2012, and found some gene fragments significantly associated with susceptibility to psychiatric conditions at the genome-wide scale, such as NRGN, TCF4, and ANK3 on chromosome 6. The results also indicate some commonalities in the mutated genes present among different psychiatric patients, greatly promoting the development of the field of CNV genome-wide association analysis. The increasingly sophisticated research methods in this field have also promoted the development of human genetics, conferred breakthroughs in research on diseases, and attracted substantial attention from researchers in the field of genetics.

2.3.3 Methods of Detecting Copy Number Variation

Copy number variation has a very close relationship with human diseases, so a CNV detection method specially designed for massive high-throughput data needs to be put forward at the current sequencing level. At present, CNV analysis methods based on existing sequencing technologies are relatively immature, but there are great differences among the different detection methods. As such, the different
methods can be used to supplement each other, overcoming each of their disadvantages. The results obtained by using all methods share some similarities. At the methodological level, the current CNV detection methods are mainly divided into two types: pair-end mapping (PEM) and depth of coverage (DOC) [32]. DOC is based on next-generation sequencing technology and is more commonly used. It can reconstruct copy number signals, restore the initial copy numbers, and reveal relevant information about CNVs. Moreover, PEM-based detection methods can detect the insertions, deletions, inversions, and other abnormalities of gene fragments. The significant span of paired genome sequence on the reference genome means the occurrence of insertions or deletions. The two types of CNV detection method are elaborated below:

**DOC-based Detection Methods**

The algorithms described in this thesis are designed based on DOC, and DOC-based algorithms are used in most CNV detection methods. DOC-based CNV detection methods assume that the number of CNVs in each bin is directly proportional to the read count in this bin; typical methods include CNV-TV [21], CNV-seq [96], readDepth [57], CNVnator [2], SegSeq [16], FREEC [8], and a time-based intelligent detection method EWT [101]. The detection principles of DOC-based algorithms are as follows. First, short sequence samples are selected from the genome. Then, BWA [45], Bowtie2 [40], and other alignment tools are used to align short sequence samples to a human reference genome, such as Human Genome 18. According to the above assumption, a specific algorithm is used to obtain information about copy number variations, so as to determine whether there is any CNV at this position. A read depth algorithm is an important method to detect CNVs. Read depth technologies can be classified with regard to the angles of parameters and bin sizes according to different model assumptions, as discussed below.
Parameter-based Model

According to different CNV model assumptions, the analytical models can be divided into parametric models and nonparametric models. Parametric models restore and reconstruct information on the original copy number by signal recovery after preprocessing of the initial read depth signals and making the assumption that they are the continuous functions of specific features. Parametric models are mainly represented by CNV-TV, among others. CNV-TV [21] assumes that CNVs are rare and controls signal sparsity through the lasso penalty term. This penalty term can ensure that the values of copy numbers are normal in most cases and mutated in only a few cases. Nonparametric models obtain the breakpoints of CNVs by modeling the processed initial signals according to the underlying probability statistical model. The main nonparametric models include CNV-seq [96], readDepth [57], and FREEC [8].

CNV-TV is a representative example of a parametric model [21]. It detects human CNVs by a least square model with a penalty term. In this process, first, the read depth data are collected. Then, the read depth data on the whole reference genome are collected by a sliding but nonoverlapping bin, so as to obtain the required read depth (RD) signals. With the reference genome used as a contrast, it is not necessary for this method to further normalize (including the correction of GC content and alignment quality) the read depth signals, as used in other approaches (CNV-seq, SegSeq, etc.). This thesis proposes a least square model with total variance (TV), in which variation breakpoints are obtained by the self-adaptive statistical detection of RD signals. This method assumes that the distribution of RD signals has a peak or a nadir within the scope of the genome, which means that copy numbers increase or decrease in this region from the perspective of gene variations. Subsequently, a piecewise constant function is used for the least square fitting of the signals. To balance the sensitivity and specificity of the algorithm, the model frequently needs to obtain the iterative solution of the TV coefficient. Therefore, Schwarz Information Criterion (SIC) [80] is introduced to find the best coefficient.
This method is suitable for paired samples with a control group (such as tumor and its corresponding control group), and a bias-corrected single sample (such as GC bias in the amplification process or mapping quality bias in the sequencing process). CNV-TV is characterized by its robustness. Even when the CNV length of the input sample is very short or its copy number is close to the normal level, its detection results still exhibit good sensitivity and specificity.

CNV-seq [96] is another popular nonparametric model. It features a statistical model with better robustness and allows the prediction of copy number ratio within a certain confidence interval. First, the fragments of both the fluorescently labeled genome and the reference genome are sampled and sequenced, after which they are aligned to the genome template. Then, the number of mapped fragments in each bin is mapped by sliding bin technology, and the obtained ratios constitute the initial data needed by the algorithm. The random sampling from the shotgun sequencing data with an uneven degree of coverage will affect the copy number ratio. Therefore, the statistical model assumes that the mean RD value in any region of the genome depends on the total number of samples, that is, the lengths of this genome and this region. Subsequently, the size of the sliding bin is calculated under this assumption, and thus the minimal confidence coefficient of the observed value is obtained. In summary, CNV-seq sets up a statistical test model to detect the CNVs in the shotgun sequencing method. CNV-seq combines the advantages of microarray genome-wide sequencing and high-throughput sequencing, and effectively evaluates confidence intervals using a powerful statistical model. Finally, this model is applied to both simulated data and empirical data. The results show that CNV-seq still maintains good specificity and sensitivity in cases with a low degree of sequencing coverage and can obtain the optimal detection results at a given confidence level.

**Bin-based Model**

According to the strategy for selecting model bins, the algorithms can be divided into two categories: fixed bin models and variable bin models. Fixed bin models,
such as Segseq [16], adopt a sliding bin with a constant size of 100 bp, and obtain the initial read depth data by calculating the amount of sequencing data within each bin. It should be noted that a larger bin size will result in less precision in determining the CNV breakpoints, leading to abnormal detection of small CNV fragments (about 1000 bp). Moreover, with a coverage degree of 3, the read depth distribution of 100-bp bins fits the normal distribution perfectly, while for larger bins this is not the case. A typical representative of a variable bin model is Ginkgo proposed by Cold Spring Harbor Laboratory in 2012 [5]; Ginkgo calculates the size of the variable bin of each gene fragment using Varbin’s algorithm. First, the algorithm simulates about 200 million short sequence read fragments from the human genome (HG18/NCBI36), and each fragment has a size of 48 bp. Then, the algorithm introduces single-nucleotide bias of a fixed frequency according to the features of sequencing data on the Illumina Sequencing Platform. Subsequently, the simulated data are input into the alignment software according to the given parameters, and aligned with the human genome. Next, the chromosomes are segmented according to the number of read fragments mapped in each bin, so as to ensure that there is an equal number of read fragments in each variable bin. Finally, the algorithm obtains about 50,000 variable bins without public areas. The empirical results have proven that, compared with fixed bin technology, variable bin technology can correct the number of read fragments in each bin within the scope of the whole genome, and thus is more accurate in terms of breakpoint resolution and has higher sensitivity.

DOC-based detection methods are mostly dependent on the read counts of short sequences, while next-generation sequencing can produce a large number of short sequence read counts. Experimental results have shown that the decisive factor for improving the resolution of CNV detection is the number of short sequence read counts rather than the length of sequence fragments. When the resolution remains the same, the sensitivity and specificity of the detection results will increase with the increase of sequencing number. As a result, the existence of a large number of short sequence fragments has a great advantage over a small number of long
sequence fragments when performing research on gene copy number. Based on the current sequencing technology, the further improvement of DOC-based CNV detection methods is also extremely promising in the field of genetic diagnosis.

**PEM-based CNV Detection Methods**

Most PEM-based sequencers include a pair of gene sequences. The relevant information about CNVs can be obtained according to the alignment results of the pair of sequences on a reference genome. If the alignment direction of a pair of sequences on a reference chromosome is the same as that of the chromosome, and the distance between the pair of sequences on the reference genome falls within a certain range, this means that concordant PEM prevails for the read counts of the sequences and the reference genome; otherwise, the case involves discordant PEM. This algorithm obtains the sizes, types, and corresponding positions of CNVs by analyzing the discordant PEM of the pair of gene sequences and the reference genome, and further obtains the status of variation of the base sequence to be detected.

When the first PEM-based detection method was proposed [92], the research group made a detailed plan of the entire detection process and provided solutions to problems such as judging the positions, sizes, and types of CNVs. Currently, the emerging methods in this field include Hydra [72], GASV [86], BreakDancer [12], and genome STRiP [28], along with other leading algorithms gradually developed based on this research. However, the preparation of a library during sequencing tends to be constrained by the available technology, which exerts some influence on the insertion fragments needed in the detection process. Therefore, in the subsequent analysis, clustering analysis of the CNV fragments of approximate positions and similar sizes should be performed so as to increase the precision of CNV detection. The clustering effect is one of the key factors influencing the performance of an algorithm, which has been subjected to intensive analysis by different researchers. For example, Hydra [72] used a heuristic clustering method to cluster the pair-ends, performed a comparative analysis of discordant PEM to obtain a pair set, and then
analyzed whether the PEMs were located in the same gene fragment and had the same direction. If two PEMs fulfill the above conditions, they are considered to support each other. The advantage of this strategy is that it can singly map the PEM candidate sets to the greatest extent, maximize the number of mutually supported PEM candidate sets, and increase the sensitivity of CNV detection, but it sacrifices the specificity of detection to some extent. Meanwhile, this algorithm can be used to study the origin of SVs. The author performed a clustering analysis of 3316 abnormal sequence breakpoints at the single-nucleotide point mutation level, and found that the structural variations of about 16% of the nontransposons were consistent with the complex breakpoint types of the reference template in DNA replication or DNA repair. The GASV [86] algorithm calculates variations through clustering polygonal intersection by a polygonal clustering algorithm, and can precisely detect the variation boundary of CNVs. Meanwhile, BreakDancer [12] segments the reference genome into bins with fixed lengths, counts the discordant PEM sequences, and obtains accurate results through clustering analysis. With high operational efficiency, the algorithm has been widely applied in the field of large-scale sequence analysis. However, owing to problems such as alignment bias, amplification bias, and sequencing bias in the sequencing process, the detection results of this algorithm still have a false-positive rate. STRiP [28], combines mass information with sequencing depth to lower the false-positive rate, greatly improving the precision of PEM-based sequencing technology. PEM-based variation detection models can detect various structural variations such as ectopia and inversions, and make up for the corresponding defects of DOC-based CNV algorithms, but cannot provide information on the types of CNVs.
2.4 Preprocessing Workflow for Sequencing Data

2.4.1 Preprocessing of Bulk Sequenced Data

Preprocessing of raw DOC data mainly involves sequence alignment, bias correction, and variation calling. We provide a detailed introduction to each step below.

Alignment refers to the sequencing genome being aligned with a reference genome, such as HG18, through alignment software such as BWA to obtain the positions of the short sequences on the reference genome. Then, the reference genome is segmented in a certain pattern to acquire a number of bins. The number of short sequences mapped in each bin is also counted to obtain initial read depth signals.

Bias correction mainly refers to the bias introduced in the amplification process being analyzed and processed to avoid it influencing the detection results of the algorithm. Previous research has shown that the bias introduced by whole-genome sequencing technology affects the sensitivity and accuracy of CNV detection, with GC-content bias having the most significant influence on the results. GC-content bias can directly affect the ability of polymerase and DNA primers to function, resulting in over- or under-amplification. Since DOC-based detection algorithms readily produce false-positive results under the interference of GC content, the GC content in each bin needs to be corrected. It has been shown that the number of fragments mapped in the bin is correlated with GC content [98]. This number (> 45%) in GC-enriched region is smaller than the mean for the whole genome and decreases with increasing GC content, leading to significant bias. In order to quantify the quality of the data, a measurement called relative read number (RRN) is defined as the ratio of the number of mapped fragments in each bin to the mean for the whole genome. The ideal value of RRN is 1, with deviation from 1 occurring in the case of amplification bias. Current majority methods normalize the mapped fragments in the current bin based on GC content, and the RRN in each bin is approximately pme [10].

Calling of variations is thus based on the bin signals after normalization, in
which $RRN \approx 1$. The values of CNVs are recovered according to the strength of signals. Meanwhile, the positions of CNVs in the whole genome and the sizes of copy numbers can be obtained according to the positions, quantity, and sizes of the bins. A few methods, such as CNV-seq and SegSeq, require an initial control sample, the normalization of read depth signals is thus unnecessary. Otherwise, normalization is required to achieve GC content correction. Read depth signals after correction can be processed by the following two methods: (1) They can be segmented through the breakpoint detection (or fragment segmentation) method, and then restored through a merge program. For example, readDepth [57] uses the Circular Binary Segmentation (CBS) algorithm, FREEC [8] uses the Lasso-based Segmentation Algorithm, and CNVnator [101] uses the mean shift algorithm. (2) Statistical hypothesis testing is performed for the read depth signals in each bin or several neighboring bins, such as Event-wise Testing (EWS). The algorithms based on breakpoint detection usually have good detection accuracy, and the detection algorithms based on statistics have good sensitivity.

In summary, the most popular detection algorithms based on read depth signals include SegSeq, CNV-seq, FREEC, CNVnator, readDepth, and EWT. Among these, readDepth, CNVnator, and EWT perform better when detecting the breakpoints of CNVs. Meanwhile, CNVnator, readDepth, and CNV-seq can more accurately predict the copy number values. FREEC, SegSeq, and EWT have a higher true positive rate, and readDepth, EWT, and CNVnator have a lower false positive rate. According to the author of FREEC, the Lasso-based Segmentation Algorithm has strong robustness when processing extreme data. However, owing to the bias produced by the norms used by Lasso during judgment of the copy number peak, FREEC is not as accurate as other algorithms in copy number prediction. Nevertheless, FREEC and EWT have excellent performance in terms of equilibrium time and memory loss.
2.4.2 Preprocessing Framework for Single-cell Sequencing Data

The aforementioned pipeline for bulk sequencing cannot be directly applied to analyze the single-cell sequenced data. The major reason for this is the need for an extra genome amplification step for SNS. The current popular amplification schemes invariably introduce a large amount of bias, and thus bias correction is indispensable for high-quality data. Therefore, we established a preprocessing step for single-cell sequenced data, as described in this section. Our preprocessing pipeline is similar to the one used elsewhere [5]. The workflow consists of five key steps, as shown in Fig. 2.1. The five key steps are described below.

**Figure 2.1:** The preprocessing workflow for single-cell sequenced data includes five steps, including genome segmentation, sequence alignment, duplicates removing, read depth estimation and bias correction.
Step 1: Segment by Variable Bin Size

The software Bowtie is adopted to first create indexes for the human reference genome. The reference genome used in this thesis is a new version from which duplicate fragments have been removed. The required input file is chromlist, which includes the numbers, initial positions, and end positions of all of the chromosomes in the reference genome.

The sequenced data are then aligned to the new hg19 as the input of Bowtie, and the specific command used is:

```
/path/to/bowtie-0.12.7/Bowtie-S-t-n2-e70-m1—best—strata
—solexa1.3—quals hg19/path/to/sequence.part.0.k50.txt
/path/to/sequence.part.k50.sam.
```

The positions that can be well mapped to the reference genome are termed good positions, while the others are termed bad ones. The number of good positions on each chromosome is calculated, by which the reference chromosome hg19 is segmented into 50,000 variable bins. Within each bin, the number of read fragments is adjusted to be the same. Once we segment the chromosome into variable bins, Bowtie is used again to align the sequence to the reference genome to obtain the required SAM files. This step takes about 500 h and is the longest step in the whole experiment.

Step 2: Remove PCR Duplicates Body

The PCR duplicates is removed by using Samtools, an open source software for handling NGS data. It is designed to deal with various file types, the Samtools are for SAM / BAM / CRAM format conversion, the BCF2 is for BCF2 / VCF / gVCF and calling SNPs as well as short insertion deletion mutations (INDEL).

Step 3: Align Single-cell Sequencing Data

The tumor single-cell sequencing data output by the Illumina sequencing platform are aligned to the corrected hg19 reference genome. The sample SRR054616 can be
downloaded from the short sequence library of NCBI ([http://trace.ncbi.nlm.nih.gov/Traces/sra.cgi?cmd=viewer&m=data&s=viewer&run=SRR054616](http://trace.ncbi.nlm.nih.gov/Traces/sra.cgi?cmd=viewer&m=data&s=viewer&run=SRR054616)). This sample originates from a single tumor cell studied by Navin et al. and the retrieval ID is the same as the sample name SRR054616. In the alignment process, the command line adopted by Bowtie is:

```
/path/bowtie-0.12.7/bowtie-S--n2--e70--30--50--m1
    --best--strata hg19/path/SRR054616.fastq
/path/SRR054616.sam.
```

The parameters -3 and -5 stand for the bases without the 3’ end and the 5’ end before alignment, respectively, and the default value is 0. The parameter -best indicates that the mapped short sequences must be sorted in descending order of mapping quality. The parameters -strata and -best must be used at the same time to represent the highest part of the mass fraction. The parameter -m1 indicates that the number of mis-paired bases must be within a tolerable range, and the best one should be randomly selected from the report files after alignment for description. Then, the average length of short sequences for alignment is 36 bp because the segmented bins are obtained by aligning the simulation data with a length of 50 bp. Thus, the short sequence length should be as close to 50 bp as possible to improve the accuracy of the algorithm. However, if all of the read counts of short sequences generated by the sequencing platform in a project are 36 bp, then 36 bp can be selected as the length of the simulated data to segment bins in the second step.

**Step 4: RD Calculation**

The matching position by the output in SAM file is used to extract the position of each each read sequence. The RC of each window is normalized to obtain RD.

**Step 5: Format Conversion and GC Bias Correction**

The Sam file obtained in the last step is converted into a bam file by the conversion tool Samtools, and an example of the command line is as follows:
Then, the bam file is sorted and the duplicate fragments introduced in the amplification process are removed. The sample code is as follows:

```
/path/to/samtools -0.1.16/samtools rmdupCs
/path/to/SRR054616.sorted.bam
```

After that, Locally Weighted Scatterplot Smoothing (LOWESS) is used to fit the number of short sequences within each bin according to GC content to obtain the results after correction.

### 2.5 Chapter Summary

This chapter briefly introduces the principles, development, technical breakthroughs over time, current developmental status, and applications of next-generation sequencing technology. It introduces current research for the detection of copy number variation. It also describes in detail the major steps involved in a pipeline that we established for analyzing single-cell sequencing data. The background introduced in this chapter is of great importance for subsequent modeling and experiments, and lays a solid foundation for the copy number variant calling described in Chapter Three.
Chapter 3

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

3.1 Introduction

Copy number variants (CNVs), a common type of structural variation, involve a duplication or deletion of a DNA segment larger than one kbp. Rapid development of new techniques for uncovering the intricacies within the human genome has revealed CNVs to be potential candidates for explaining various phenotype differences and genetic diseases [25]. Recent studies have revealed a strong correlation between CNVs and classic Mendelian diseases, plus other, less well characterized conditions, including autism, schizophrenia, osteoporosis, and certain tumor types [43, 88, 81]. Various laboratory techniques have been developed to measure the DNA copy number. Traditional major approaches include array-based comparative genomic hybridization (arrayCGH) and single-nucleotide polymorphism (SNP) array methods [67]. The next-generation sequencing are adopted as a popular strategy for genotyping and has included comprehensive characterization of CNVs by generating hundreds of millions of short reads in a single run [56]. The NGS could achieve higher coverage and resolution, thus allowing for more accurate estimation of copy numbers and detection of breakpoints with high throughput, than arrayCGH does. This has dramatically increased our capability of detecting CNVs. However, due to the complexity of the genome and the short read lengths associated with NGS technology, there are still many challenges associated with the analysis of NGS data for CNVs.

There are multiple approaches to having the copy number profiles from sequenced data. One of the major and popular approaches is based on counting of depth of coverage (DOC, also known as read-depth (RD) methods) [106, 101, 19]. The
DOC methods count the number of reads that fall in each pre-specified window of a certain size [2, 96, 101] provided that the sequencing process follows a Poisson distribution with its mean value being proportional to the number of copies. The CNV calling is then obtained by detecting deviations between the number of reads mapped to a chromosome window with its expectation. However, the empirical sequencing process itself will inevitably introduce certain biases, such as non-uniform GC-content and low map-ability ratios. Therefore, the observed copy numbers of the genome appear to have spurious signals due to over- or under-sampling. While a number of successful approaches have been presented for copy number calling, there is still a paucity of methodology for analyzing it through systematical framework to achieve high accuracy and nice robustness.

In this thesis, we tackle the problem by formulating the RD signal reconstruction problem into a quadratic minimization problem involving two constraints. Our proposed alternating direction minimization-total variational (ADM-CNV) method offers two main contributions as opposed to the previous methods: 1) The RD signal reconstruction is characterized by a well-designed model to reveal intrinsic structure, and thus facilitating the subsequent statistical testing for CNV detection; 2) An efficient numerical method using a classical alternating direction minimization framework is tailored to solve the ADM-CNV model. The key characteristic of the ADM method that makes it different from standard methods in solving TV related minimization problems is that it does not require matrix inversion, thus greatly saving computational cost. This is particularly useful in NGS sequencing data analysis, in which hundreds of thousands of signals need to be recovered.

The remainder of this chapter is as follows: Section 3.2 overviews the existing models for analyzing copy number variants. In particular, the methods based on read-depth counting for CNV analysis are briefly reviewed. In Section 3.3, we reformulate the detection of copy number variants into signal reconstruction problem, which is constrained by multiple conditions to satisfy the characteristics of copy number distribution. An exact numerical solution is presented and its theoretical
convergence is analyzed in Section 3.4. Section 3.5 demonstrates the performance of the proposed model through extensive experiments on both synthetic and empirical sequencing data. Finally, concluding remarks are given in Section 3.6.

3.2 Overview of Existing Models for Detecting C-NV

The current methodology for reconstructing CNV information from observed data can be broadly divided into four categories: (i) depth of coverage (DOC, also known as read-depth (RD) methods) [101, 19], (ii) paired-end mapping (PEM, also known as read-pair methods) [50], (iii) split-read (SR) [1, 53], and (iv) assembly-based (AS) methods [3]. All of these methodologies, except for the last one, require initially mapping the sequenced reads to a known reference genome. Read depth analysis is particularly effective for exome sequence data, as it does not rely on sequencing into or near CNV breakpoints.

Most of the existing tools for CNV calling based on read depth can be broadly divided into two categories: parametric versus non-parametric schemes. The performance of parametric approaches heavily depends on the underlying statistical distributions. For example, Magi [51] used the stochastic process of a shifting level model to simulate multi-sequential samples. ExomeCNV [79] and CNV-seq [96], assume a Gaussian distribution of the read ratios. Also, they assume that the proportion of reads matching to a specific sample usually follows a binomial distribution whose success rate is determined by the genome-wide read count ratio between the test sample and the reference set, as well as the potential presence of CNVs. However, it has been shown that the binomial assumption is actually violated in practice due to technical variability, induced noise in library preparation, and capturing or sequencing error. ExomeDepth [69] uses a beta-binomial model to approximate the distribution of the read count ratio, and then employs statistical inference to detect the CNVs. CLImAT [102] takes tumor impurity and ploidy into considera-
tion for identifying genomic aberrations. DeAnnCNV [107] is an online integrated tool to precisely detect and systematically annotates copy number variations from whole-exome sequencing (WES) data. Hidden Markov models (HMMs) and many variations thereof [93] have dominated the parametric approach. The unique characteristics of HMMs lie in its flexibility for handling several common complications, including variable single nucleotide polymorphism (SNP) frequencies, variable distances between adjacent SNPs, linkage disequilibriums, and relationships between study subjects.

Nonparametric methods attempt to reconstruct the CNV copy number from the observed read depth signal in concert with the application of subsequent statistical inference on the estimated signal to detect the CNVs. For simplicity, let us consider a plateau/basin in the reconstructed signal to be a duplication/deletion event. Therefore, the CN is assumed to be piece-wise constant function with two fundamental characteristics: sparsity because of few CNVs, and smoothness because of contingency positions possessing similar CNs. In such a configuration, the CNV detection is considered to be a piece-wise linear signal recovering problem. The signal is normalized around zero, thus possessing sparsity. Mathematically, let the parameter vector $\mathbf{r} = (r_1, r_2, \cdots, r_n)$ quantify DNA levels at $n$ successive SNPs. These levels are normalized such that $r_i = 0$ corresponds to the standard copy number 2, where SNP $i$ is represented once each on the maternal and paternal chromosomes. After such configuration, the problem is reformulated as a signal reconstruction problem. A classical approach to addressing it is by the fussed lasso scheme [90] with its refinements [75]. It was also tailored to use in CNV detection on array CGH [91]. Most recently, the conditions that the fussed lasso could consistently recover the piecewise constant pattern have been investigated [71]. Numerical scheme for solving the fussed lasso typed problem includes approximation method of path coordinate descendent optimization [4], and linearization method within ADM framework [7, 19].
3.3 Problem Modeling for CNV Detection Model with Multiple Norm Constraints

We tackle the problem by modeling the CNV reconstruction into a least-square minimization problem, penalized by two constraints:

\[
\min_{x_i} f(x_i) = \frac{1}{2} \sum_{i=1}^{n} (x_i - r_i)^2 + \lambda_1 \sum_{i=1}^{n} \| x_i - x_{i-1} \|_1 + \lambda_2 \sum_{i=1}^{n} \| x_i \|_1
\]  

where \( r_i \) is the observed reads depths at the \( i \)-th position, and \( x_i \) is the reconstructed copy number. The second term of Eq. (3.3.1) attempts to penalize the similarities between adjacent sites, whereas the third term aims to achieve a sparse signal representation.

We would like to emphasize here that the sparsity and smoothness may not be applicable to arm-level CNVs, which has a high frequency of occupying one chromosome arm exactly. However, the sensitivity of detecting focal amplifications and deletions will be greatly increased after filtering out the arm-level CNVs by using either amplitude or length thresholds [55]. An effective scheme is to firstly separate the two types of CNVs by using length threshold and then discriminate both of them individually.

As the second term in above equation is actually a TV term, we rewrite it as:

\[
\min_{x} f(x) = \| x - r \|_2^2 + \lambda_1 \| D x \|_1 + \lambda_2 \| x \|_1
\]  

where \( D \) is the first order difference matrix, \( x = (x_1, x_2, \ldots, x_n) \) and \( r = (r_1, r_2, \ldots, r_n) \).

Minimization problems involving a TV term are common in disparate areas, including signal processing and image recovering [62]. Due to the non-differentiability of the TV term, Zhang et al. [108, 109] proposed to use an \( l_2 \) norm to approximate the TV norm, and then to use a majority minimization framework to solve the problem. Such an approach will typically lead to explicit solutions, and is thus very efficient at a sacrifice of accuracy. To obtain a sparse solution attributing to
the TV norm, a normal practice is to use the Lasso numerical method to solve the problem [19, 112]. However, a common drawback of them is their highly demanding computational cost due to the necessity of calculating a matrix inversion or introducing too much slack variables.

Recent researches [62, 112] have successfully demonstrated the efficiency of ADM in solving image restoration problems involving TV norms. The numerical solution show surprisingly fast speed and high accuracy. The image restoration ADMs are variants of the classical Augmented Lagrangian method for optimization problems with separable structures and linear constraints, and they have been intensively studied in the optimization community. Inspired by this idea, we reformulated the problem in Eq. (3.3.2) into an optimization problem with favorably separable structures, thus allowing it to be efficiently solved using ADMs. The separable structure decomposition is particularly useful in sequencing data analysis due both to its large/huge dimensionality, and to the large amount of data involved; plus it dramatically reduces computational cost by overcoming the matrix inversion burden, required by Lasso and majority minimization schemes [62, 112].

3.4 Numerical Solution by Alternating Direction Minimization (ADM) Model

3.4.1 Composite Penalty Minimization and ADM Model

The proposed model (3.3.1) aims to minimize the average loss from the training data. The averaged loss is regularized by two regularizers. Such composite penalty minimization problem have been widely used in bioinformatics, text and other structured data mining tasks [110]. Optimizing such structure regularized loss minimization problem can be challenging, especially confronted with very large dataset [15, 14, 13].

If the composite penalties are nonsmooth or non-differentiable, one can neither directly take derivative as dealing with smooth regularizer nor have closed form
proximal update for simple nonsmooth term. In this cases, the simple methods based on gradient descent can not be directly applied when these complex regularizers are used. Apparently, other techniques need to be introduced. A popular line of research is a split method called alternating direction method of multipliers (ADM) [62, 9]. To alleviate computation burdens, the ADM model have been refined. The refined models include stochastic gradient ADMM methods [65], and incremental gradient ADMM method [76, 82]. Recently, proximal average has been introduced to deal with component penalties efficiently when each of its component allows cheap computation of proximal operator [15, 110]. The proximal average method is simple to implement and analysis, but it either converge slowly or suffer from high per-iteration cost.

Mathematically, the ADM is designed to solve a generally well-structured optimization problem,

$$\min f_1(x) + f_2(Bx)$$ (3.4.3)

where $f_1 : \mathbb{R}^d \to \mathbb{R}$ and $f_2 : \mathbb{R}^p \to \mathbb{R}$ are convex functions. $B \in \mathbb{R}^{p \times d}$ is a transformation matrix with full column rank.

Given a Lagrangian multiplier $\gamma$, the augmented Lagrangian function for the above function is,

$$L(x, y, \lambda) = f_1(x) + f_2(y) + \langle \lambda, Bx - y \rangle + \frac{\gamma}{2} \|Bx - y\|^2_2.$$ (3.4.4)

The above minimization problem could be addressed solved by alternating solving the following sub-problems sequentially::

$$\begin{align*}
    x^{k+1} &\in \arg \min L(x, y^k, \lambda^k) \\
y^{k+1} &\in \arg \min L(x^{k+1}, y, \lambda^k) \\
\lambda^{k+1} &= \lambda^k - \rho(Bx^{k+1} - y^{k+1}).
\end{align*}$$ (3.4.5)

where $\rho$ is step size.
3.4.2 Fast Numerical Algorithm

Mathematically, let $\omega = \lambda_1 D x, z = \lambda_2 x, B = \begin{pmatrix} \lambda_1 D \\ \lambda_2 I \end{pmatrix}$, and $y = \begin{pmatrix} \omega \\ z \end{pmatrix}$ with $I$ being the identity matrix. The minimization problem (3.3.2) could be equivalently rewritten as,

$$\arg\min \|y\|_1$$

subject to:

$$x \in \mathbb{R}^n$$
$$\omega = \lambda_1 D x$$
$$z = \lambda_2 x$$
$$\|x - r\|_2^2 \leq \alpha.$$ 

The augmented Lagrangian of the minimization problem is given by:

$$L(y, x) = \|y\|_1 + \chi_S(x) + <\lambda, Bx - y> + \frac{1}{2} \gamma \|Bx - y\|_2^2$$ 

(3.4.6)

where the parameters $\lambda$ and $\gamma$ are Lagrangian multiples. The function $\chi_S(x)$ is the usual indication function on the set $S = \{x \in \mathbb{R}^n, \|x - r\|_2^2 \leq \alpha\}$ for a predefined parameter $\alpha$.

Let $f_1(y) = \|y\|_1$, $f_2(x) = \chi_S(x)$, the minimization problem falls into the standard framework of Alternating Direction Method (ADM) with the following notations:

$$L(y, x) = f_1(y) + f_2(x) + <\lambda, Bx - y> + \frac{1}{2} \gamma \|Bx - y\|_2^2$$ 

(3.4.7)

This new formulation allows it to be decoupled into two separate sub-problems with variables $x$ and $y$, respectively, and, therefore, to be solved in an iterative manner.

**Step 1:** Find

$$x^{k+1} \in \arg\min <\lambda, Bx - y^k> + \frac{1}{2} \gamma \|Bx - y^k\|_2^2$$

subject to:

$$\|x - r\|_2^2 \leq \alpha.$$ 

(3.4.8)
With an algebra transformation, minimization of Eq. (3.4.2) is equivalent to the following constrained least square problem:

$$x^{k+1} \in \arg \min \frac{1}{2} \| Bx + \frac{\lambda}{\gamma} - y^k \|^2_2$$

subject to:

$$\| x - x_0 \|^2 \leq \alpha.$$ 

Let \( \hat{x} = x - x_0 \), then the above minimization can be rewritten as,

$$\hat{x}^{k+1} \in \arg \min \frac{1}{2} \| B\hat{x} + Bx_0 + \frac{\lambda}{\gamma} - y^k \|^2 + \delta \| \hat{x} \|^2,$$

where \( \delta \in [0, +\infty) \) is a Lagrange multiplier. This is equivalent to solving the least square problem of \( \min_{\hat{x}} \| \hat{B}\hat{x} - \hat{c} \|^2 \), where \( \hat{B} = \begin{pmatrix} B \\ \delta I_n \end{pmatrix} \), \( \hat{c} = \begin{pmatrix} c \\ 0 \end{pmatrix} \), and \( c = -(Bx_0 + \frac{\lambda}{\gamma} - y^k) \). The analytical solution is explicitly given by:

$$\hat{x}^* = (\hat{B}^T \hat{B})^{-1} \hat{B}^T \hat{c} = (B^T B + \delta^2 I)^{-1} B^T c = (\lambda_2^2 D^T D + \lambda_2^2 I + I)^{-1} B^T \hat{c}.$$ 

One may note that the matrix \( D \) is circular and thus could be diagonalized by Fourier transform as \( D = F^T K F \), where \( F \) is 2-D discrete Fourier transform (DFT) and \( K \) is a diagonal matrix containing the DFT coefficients of the difference operator \( D \) positive definite and thus its inverse can be computed efficiently by singular value decompositions. It follows that

$$\hat{x}^* = F^T (\lambda_1^2 K^T K + \lambda_2^2 I + I)^{-1} F \hat{B}^T \hat{c}.$$ 

(3.4.9)

**Step 2:** Find

$$y^{k+1} = \arg \min f_1(y) - \lambda^k y + \frac{1}{2} \gamma \| Bx^k - y \|^2$$

(3.4.10)

Its solution is given by Moreau proximity operator, which is simply a soft thresh-
old as follows:

\[
y^{k+1} = \arg \min f_1(y) - \lambda^k y + \frac{1}{2} \gamma \|Bx^k - y\|^2_2 \\
= \arg \min f_1(y) + \frac{\gamma}{2} \|y - (Bx^{k+1} + \frac{\lambda^k}{\gamma})\|^2 \\
= \text{shrinkage}_{1/\gamma}[(Bx^{k+1} + \frac{\lambda^k}{\gamma})].
\]

Therefore, the analytical solution \( y^* \) is given by,

\[
y^* = y_0 - \min(\frac{1}{\gamma}, |y_0|) \frac{y_0}{|y_0|}
\] (3.4.11)

where \( y_0 = (Bx^{k+1} + \frac{\lambda^k}{\gamma}) \).

The above two steps are updated iteratively until convergence to achieve the final optimal solution of \( x \).

The prominent characteristic of the aforementioned numerical solution is its separable structure that allows for finding a solution quickly. After simple algebra operations, explicit analytical solution for the two sub-problems, Eq. (3.4.2 and 3.4.10) are obtained. In both sub-problems, their solutions involve low-cost calculations and thus could be solved efficiently.

### 3.4.3 Convergence and Complexity Analysis

**Theorem 1 (Eckstein-Bertsekas [22])** Consider problem (3.4.3), \( f_1 : \mathbb{R}^d \to \mathbb{R} \) and \( f_2 : \mathbb{R}^p \to \mathbb{R} \) are closed, proper, convex functions. For arbitrary \( \beta > 0 \) and \( y_0, \lambda_0 \in \mathbb{R}^p \), if there exist two convergent sequences of \( \{\eta_k \geq 0, k = 0,1,\cdots\} \) and \( \{\rho_k \geq 0, k = 0,1,\cdots\} \), such that there are three sequences \( \{x_k \in \mathbb{R}^d, k = 0,1,\cdots\}, \{y_k \in \mathbb{R}^p, k = 0,1,\cdots\}, \{\lambda_k \in \mathbb{R}^p, k = 0,1,\cdots\} \) that satisfy

\[
\|x_{k+1} - \arg \min_x f_1(x) + \frac{\mu}{2} \|Bx - y^{k+1} - \lambda_k\|^2_2\| \leq \eta_k \\
\|y_{k+1} - \arg \min_y f_2(y) + \frac{\mu}{2} \|Bx^{k+1} + y - \lambda_k\|^2_2\| \leq \rho_k \\
\lambda_{k+1} = \lambda_k - (Bx^{k+1} - y^{k+1} - b)
\]
Then, if Eq. (3.4.3) has a solution \( x^* \), it follows that \( x_k \to x^* \).

The convergence of ADM-CNV can be established by Theorem 1.

**Corollary 1** The algorithm in section 3.4.2 for the problem of (3.4.7) converges to a minimizer.

**Proof:** The proposed ADM-CNV is an instance in Theorem 1, where \( f_1(y) = \|y\|_1 \) and \( f_2(x) = \chi_S(x) \) are closed, proper and convex. The matrix \( B = \begin{pmatrix} \lambda_1 D \\ \delta \lambda_2 I \end{pmatrix} \) has full column rank. According to Theorem 1, ADM-CNV is convergent to a minimizer of the objective function.

The optimal solution is obtained through alternative updating the sequence of \( \{x^k, y^k, \lambda^k\} \), whose solution is given in Eq. (3.4.9) and (3.4.11), respectively. In Eq. (3.4.9), the inverse of diagonal matrix \( (\lambda_1^2 K^T K + \lambda_2^2 I + I) \) costs \( O(n) \). The products by \( F, F^T \) with the inverse diagonal matrix need \( O(n \log n) \). Similarly, the product of \( \hat{B}^T \hat{c} \) costs \( O(n \log n) \) too. Thus, the total computation cost in Eq. (3.4.9) is \( O(n \log n) \).

In Eq. (3.4.11), the soft thresholding operator \( \min(\frac{1}{\gamma}, |y_0|) \) costs \( O(n) \). In total, the computational cost of ADM-CNV scales as \( O(n \log n) \).

### 3.5 Experimental Results

The proposed method for analyzing raw NGS data consists of four steps. In the first step, the short reads are aligned to a reference genome using standard tools such as MAQ [46] and Bowtie [41]. The pipe-line we built in section 2.4.1 is used to achieve the task. In the second step, the aligned reads are used to estimate the read depth signal \( r \) to measure its density of the aligned reads. Next, our proposed method is used to reconstruct the copy number from the raw ratio reads. Finally, statistical testing is borrowed from other methods to detect the suspicious CNVs.

We conducted the experiments to demonstrate the performance of the proposed model. The testing datasets were categorized into three types: synthetic, simulated and empirical data. Due to the proximity of the proposed method to CNV-TV [20],
we firstly designed an experiment with synthetic data to illustrate the superiority of our proposed model over CNV-TV. In the second experiment, human genome data were employed to serve as fundamental testing data. CNVs with different length and copy numbers were introduced artificially into the human genome data to simulate real sequencing process. The proposed ADM-CNV method and six representative CNV detection methods were compared to evaluate their performances. Finally, experiment on an empirical data of chromosome 21 of NA19240 (Yoruba female) was comparatively studied.

3.5.1 Running Environments and Parameters for Competing Methods

Five popular CNV calling methods, including CNV-seq [96], FREEC [8], readDepth [57], CNVnator [2] and CNV-TV [19], are borrowed for comparing with the proposed ADM-CNV. The running environments and configurations are listed hereinafter.

1: CNVnator with version of 0.3 was installed in linux system. The default parameters were used except the bin size set at 100bp. The input file is in the format of BAM.

2: readDepth with version of 0.9.8.4 was installed in linux system. The default parameters were used. The input file is in the format of BED.

3: ADM-CNV, the same as CNV-seq and CNV-TV, was installed in linux system. The default parameters were used. The input file is in the format of HIT. It also requires a reference genome as input.

4: FREEC with version of 2.0 was installed in linux system. The default parameters were used. The input file includes chromosomes fasta file, the file with chromosome length as well as the sample file with BAM or SAM format.

5: CLImAT, with the version of 1.2.2, was installed in linux system. The default
parameters were used. The input file includes a test BAM file, an SNP file, a genome reference file(.fasta) and a mappability file(.bw).

3.5.2 Synthetic Studies by Comparing with CNV-TV

Random read depth data was created to simulate the reads from Illumina platform. At most positions in the data, the ratio was normalized to have zero mean value. At certain sites, read ratios following Poisson distributions with various mean values were added to simulate the CNVs. To test the robustness of the methods, the Gaussian noises with zero mean value and three different variances are further added to the data. The synthetic data are shown in the first column of Fig. 3.1. From the top to the bottom, the degraded noise level tends to larger. The results after using ADM-CNV and CNV-TV are shown in the second column of Fig. 3.1. The results by CNV-TV are denoted in red. The detected CN curve after ADM-CNV, highlighted in solid black, accurately characterizes the distribution of the read depth, denoted in blue. One may observe that the detected CNV in blue line performed less satisfactorily the proposed method, especially when the noises level is high, in which only the most prominent CNV was detected as shown in the right part of Fig. 3.1(c).

3.5.3 Comparative Studies on Simulated Data

The synthetic read ratios were created by a Poisson distribution with various mean values. It was then corrupted by Gaussian noises with three different variances(0.01,0.1,1). To test the robustness of the proposed method, different CNV profiles were artificially inserted into the read ratios. For example, the read ratios between 200 and 219 were thin and shrill, those between 1500 and 1519 were wide and short. In total, 16 different CNV shapes were created.

In creating the simulated dataset, a random DNA sequence with the size of 2 Mbps was firstly generated, whose base pairs obey the uniform distribution, i.e., the probability to be A, T, C or G is 0.25. A genome of 2 Mbp is large enough because we
only introduced one CNV in each genome for efficient comparison. To simulate the
diploidy, we concatenated this random DNA sequence with its duplication, yielding
the reference genome of length 4 Mbp. Then, a CNV with predefined copy number
and copy length was introduced to generate the test genome. The SNPs and indels
were also introduced in our experiment. Their frequency is 5 SNPs/kbp and 0.5
indels/kbp, respectively. The indels have the random length of 1-3 bp. Then,
short reads were created by sampling at the test genome to simulate the short-gun
sequencing. Each read has the length of 35 bp to agree with the Illumina platform.
The short reads were aligned to the first half of the reference genome by using
Bowtie2. Since a read may align to multiple locations, here we chose the uniquely
mapped read only. The proposed ADM-CNV and the other six CNVs detection
methods were called in the simulated data. Their outputs were compared with the
ground truth to evaluate the performances.

In summary, the simulation steps are briefly summarized as follows:

Step 1: Extract a real sequence with length 2 Mbp from chromosome 21 as the con-
trol genome. It was concatenated with its duplication, yielding the reference
genome of length 4 Mbp;

Step 2: Introduce CNV with copy number $c$ and single copy length $d$ artificially to
generate the test genome. In our experiments, $d$ was set to be 6kbp and $c$
being tested at different values of 3 and 6;

Step 3: Sample the test genome to simulate the single-end short reads. The location
of each read is uniformly distributed at the given genome (genomic coordinate
from 1 to 4e6), and each read has length 35 bps to agree with the Illumina
platform. The sample coverage is set at 1. In total, we generate $\frac{1 \times 4e6}{35} \approx 114286$
short reads. The reference genome is also sampled and handled to obtain the
read depth ratio.

Step 4: Align the short reads to the first half of the reference genome with Bowtie2 [41]
(the second half is only the duplication of the first half). Since a read may
align to multiple locations, only the uniquely mapped reads are used for testing to minimize the bias brought by misalignment;

**Step 5:** Use a non-overlapping window of size 800 bps to calculate the read depth ratio. Then, we call ADM-CNV and other popular CNV detection methods to evaluate their performance. The result by each method is compared with the ground truth.

We compare the proposed ADM-CNV with five other popular CNV detection methods, including CNV-seq [96], FREEC [8], readDepth [57], CNVnator [2] and CNV-TV [19]. In the model of CNV-TV, the detection of CNV is modeled as a change-point detection scheme from the read depth signal, penalized with a total variational term. FREEC adopted a similar idea, yet estimation is obtained by using a non-overlapping sliding windows (raw CNP). CNV-seq and readDepth are based on statistical testing model. The former conducts the statistical confidence assessment of observed copy number ratios modeled as Gaussian ratio distribution. The latter one models the distribution of reads, which are uniquely mapping to the genome by negative-binomial distribution. CNVnator is based on mean-shift tracking to broaden the range of discovered CNVs.

For each pair of the two parameters (copy length $d$ and copy number $c$), the experiment was repeated for 30 times. The results of CNV detection by each method are shown in Fig. 3.2. The horizontal axis denotes the experiment index, ranging from 1 to 30, while the vertical axis is the genomic coordinate. The blue line represents the detected CNV regions while the red dots denote the ground truth. The performance of each method was quantified by the accuracy of the detected CN matching with the ground truth. The better the matching is, the better performance of the CNV detection method is. Among the six methods, readDepth achieved superior performance by matching perfectly to the ground truth. CNVnator was less satisfactory in that it failed to detect true CNV twice and produced a large false positive in the last experiment. FREEC and CNV-TV resulted in inaccurate CNVs during the 30 experiments. In comparison, the proposed ADM-CNV
and CNV-seq achieved very nice performance by almost perfectly matching to the ground truth. The ADM-CNV achieved slightly over-performance than CNV-seq by having smaller false positives. Such visual observations were further validated by quantitatively comparisons. The true positive ratio (truly detected CNVs located within the ground truth) over false positive ratio (falsely detected CNVs within the ground truth) for the 30 experiments was calculated and shown in Table 3.1. The performance of the ADM-CNV was suboptimal to the readDepth, but was comparable or better than the others. It performed robustly and steadily across the 30 repetitions.

Table 3.1: Number of true positive/false positive detections for the 30 trails in simulation experiment.

<table>
<thead>
<tr>
<th>Method</th>
<th>CNV-seq</th>
<th>FREEc</th>
<th>readDepth</th>
<th>CNVnator</th>
<th>CNV-TV</th>
<th>ADM-CNV</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPR</td>
<td>1</td>
<td>0.994</td>
<td>1</td>
<td>0.93</td>
<td>0.997</td>
<td>1</td>
</tr>
<tr>
<td>FPR</td>
<td>2.7e-4</td>
<td>6.3e-4</td>
<td>0</td>
<td>5e-3</td>
<td>6e-4</td>
<td>2.5e-4</td>
</tr>
</tbody>
</table>

3.5.4 Comparative Studies on Real Sequencing Data

We tested ADM-CNV on an empirical NGS data by comparison with the aforementioned five methods, CNV-TV [19], FREEC [8], CNV-seq [96], CNVnator [2], readDepth [57] and a recently proposed method of CLImAT [102], to demonstrate its performance on empirical NGS data. The tested mapped reads data (BAM files) was downloaded from the 1000 Genomes Project at (http://ftp.1000genomes.ebi.ac.uk/). The raw data was firstly aligned to a reference genome by alignment tools of Bowtie2 [41]. The reference samples were obtained by randomly sampled at the reference genome to obtain a .hit file. The location of each read is uniformly distributed at the given genome (genomic coordinate from 1 to the length of the chromosome 21), and each read has a length of 35 bps to agree with the Illumina platform. The sample coverage is set to be same as that of the bam file. Then the copy number signal was calculated by dividing the read count of each window’s test
The DGV database of genomic variants was downloaded from http://projects.tcag.ca/variation/ to mark all of the discovered CNVs reported in the literature, and these known CNVs were chosen to serve as ground truth. Here, we used the beta version of DGV, from which CNVs can be retrieved with respect to the sample name, the platform, and the project [34]. The option of filter query was external sample id = NA19240, chromosome = 21, assembly = NCBI36/hg18, variant type = CNV. To calculate the CNV detection power at a tolerable genome wide false discovery rate, the genome was artificially divided into sliding windows at 800bp with overlaps at 200bp.

The proposed ADM-CNV was applied on the data and the reconstructed C-N (highlighted blue color) as well as the detected CNV (highlighted in red) were shown in Fig. 3.3. For better visualization, a short segment interval of 45.7M-46.1M (highlighted in green) was zoomed-in and displayed in Fig. 3.4. The two horizontal lines represents the global thresholds for the CNV, and the detected CNVs are highlighted in red. The aforementioned six methods were also tested to compare with ADM-CNV. The results were analyzed by Venn diagram and summarized in Table 3.2. This table quantifies the number of overlapping regions between those detected by the seven methods. All of the detected CNVs were validated in DGV. For clarity, only those blocks with numbers greater than 1,000 were used in the comparison. The integer 1 denotes there exists of a logical overlap between the ones by the tested methods and the reported ones in DGV, while 0 implies of zero overlapping. One may observe that ADM-CNV can detect as many CNV as 24,544 blocks, which were all identified in DGV. This number is dramatically larger than the ones detected by other methods except CLImAT. For example, CNV-TV can only detect 9,272 blocks, and CNVnator detected 14,435 blocks. The performance of CLImAT was superior by detecting as many as 785,590 blocks. Such nice performance was attributed to its integrative analysis, including careful handling of data.
preprocessing. However, it also resulted in a high computational costs, as shown in Table S1. The last column in Table 3.2 summarized the overlapping block number detected by all the seven methods. It provided a consistency measurements of the results. CLImAT preformed superior by detecting as many blocks as 785590. The proposed ADM-CNV ranked the second by detecting 24544 blocks. One may also observed from the fifth column that both CLImAT and ADM-CNV found 24508 common blocks, which was more than 90% of the total number of 24544 by the ADM-CNV. It implies that the proposed method could detect highly consistent variations. Another interesting observation from the last column is that only 5678 blocks were uniformly detected by all the seven methods. Such inconsistences show that it prone to have a large number of false positives in CNV detection.

F-score quantitative measurements were also calculated for each method. The F-score measures the overlap ratio between two intervals and thus takes values between 0 and 1. A low score means the overlap quality is poor, while a higher score implies a better overlap. The formula for calculating the F-score is \( F = \frac{2PR}{P+R} \), where \( P \) is the precision (percent of detected CNVs that overlap with the ground truth CNVs from DGV), and \( R \) is the recall (percent of the ground truth CNVs which overlap with the detected CNVs). For visual comparison, the best and the second best value are highlighted in red and blue, respectively. Table 3.3 lists the top 10 F-score CNVs detected by each method. In most cases, the proposed method achieved superior or sub-superior performance in terms of F-scores. The method of CNVnator achieved the top ranking. The performance of the proposed method outperformed CLImAT except at the first column. Another prominent characteristic of the proposed ADM-CNV is its stability. Among the top 10 CNVs detected, the proposed method achieved the next best F-score for every CNV, and the smallest standard deviation.

To have a comprehensive understanding on the F-scores after each methods, the average F-scores distribution was further calculated. It reflects the averaged performance of the CNV detection methods. The CNVs detected by the aforementioned methods were categorized into 10 classes \((0−0.1, 0.1−0.2, \cdots, 0.9−1)\) by its F-score.
The results, shown in Table 3.5, demonstrated that the performance of ADM-CNV ranked the third, suboptimal to CNVnator and CLImAT at the category of 0.9-1.0. It achieved superior performance than CNVnator at the category of 0.0-0.1. In the other categories, the ADM-CNV always outperformed the CNV-TV and had comparable performance in comparison with the others.

We have illustrated that CNVnator achieved the best performance in Table 3.3. However, it obtained the second lowest (highlighted in red and blue) performance among all the methods when viewing its average distribution at class 0.0 – 0.1 in Table 3.5. The main reason is that CNVnator could detected almost all deletions, but it generated large false positives for duplications. Such inaccuracy resulted in a low precision value. For the class 0.9 – 1.0, one will find out that the method of CLImAT, the proposed method and CNVnator were the top three model. In summary, CLImAT ranked the top one by achieving even distribution of F-score. The proposed method performed suboptimal while CNV-TV performed less satisfactory among all the tested methods.

In summary, the proposed ADM-CNV performed stably under various noise corruptions when tested in synthetic data. In the simulation experiment, the performance of ADM-CNV ranked in top two positions. When comparing the F-score, ADM-CNV ranked uniformly in the first or second place to CLImAT. However, the computational cost of the CLImAT method is surprising higher than ADM-CNV by over ten times. Such high computation cost makes it difficult to be applied for real-time analysis. The proposed ADM-CNV achieves a nice balance between computational cost and performance.

### 3.5.5 Hyper-parameters Pruning

There remains a challenging problem with model (3.3.2) in that of tuning the hyper-parameters of \( \lambda_1, \lambda_2 \). The parameters of \( \lambda_1, \lambda_2 \) have a profound influence on the reconstructed CNV signal. Since \( \lambda_1 \) controls the signal difference penalty, a larger value will lead to a smoother signal. In comparison, the value of \( \lambda_2 \) controls the
sparsity of the CNV, a larger value will lead to a curve with most of its elements being zero. Although the two parameters could be set by a rule of thumb, an automatic strategy for choosing the parameters is desirable. Parameter pruning can be viewed as a model selection problem. There are a few commonly used model selection schemes, including the least angle regression [23] and Schwarz information criterion (SIC) [20]. Unfortunately, such model selection schemes do not return encouraging results in our framework, possibly due to the bias introduced by sequencing or our minimization process.

The parameter $\lambda_2$ controls the sparsity of the reconstructed ratios and thus serves as a thresholding operator. It should be large enough to smear out the noise, but not so large that over-smoothing the signal occurs [111]. A classical approach is to use the well-known hard threshold scheme in wavelet shrinkage [18],

$$\lambda_2 = \sigma \sqrt{\ln(N)}$$

where $N$ is the sample size and $\sigma$ is the standard variance of the noises that are assumed to follow $N(0, \sigma^2)$. Since CNVs are sparse in the data, we can estimate $\sigma$ from the data using the median-absolute-deviation estimator, $\hat{\sigma} = 1.48\{\text{median}\{r - \text{median}(r)\}\}$. 

As the relative weight $\frac{\lambda_1}{\lambda_2}$ balances the two terms in $\lambda_1\|Dx\|_1 + \lambda_2\|x\|_1$ and controls the recovered read ratios, we empirically set $\lambda_1 = 20\lambda_2$.

To further investigate the influence of the ratio of the two penalty parameters over the model performance, we conducted extensive experiments by changing the ratio from 15 to 50. Our experiments manifested that, when the ratio of the two parameters is less than 20, it has minor influences on the model performance. However, if the ratio is larger than 20, the smoothing term will impose heavy constraints such that the resulted CN loses variation. Consequently, small CNVs will be smeared out and resulted in a low precision value.

Through our experiments, these three parameters were optimized by a rule of thumb to minimize the false positive rate in detecting CNVs. A typical example is shown in Fig. 3.5. The false positive rates to detect copy number alterations was
considered as a function of the cut-off value, and was plotted in Fig. 3.5. It can be observed that, when the cut-off value is 20, the genome-wide false positive rate is the smallest. Thus, the calling criteria was set at 20.

3.5.6 Computational Time of the CNV Detection Methods

We recorded the running time of the proposed method as well as other six methods in empirical experiment at Section 3.5.4. The results were summarized in Table 3.4. The formats of the source code were also provided for reference. One may note that ReadDepth, coded by R, needed the least time of 134.8. Similarly, CNVnator, implemented by C++, also had nice performances. FREEC, implementing the fussed lasso regularized least square estimation model, costed comparative time than ADM-CNV. It was implemented by C and thus was efficient. CLImAT consumed the longest time, yet its performance in empirical experiments was the best. The proposed method ADM-CNV was comparable compared with the other methods. However, we would like to emphasis that our experiments were conducted by Matlab and thus the reported running time of the proposed method could be further accelerated if the other programming languages like C++ are used.

3.6 Chapter Summary

We have proposed an original model for identifying CNVs from raw NGS profiles by formulating the problem into a changing point detection optimization. The new formulation possesses a unique and fundamental characteristic distinct from its peers, in which the two constraints sparsity and smoothness of the reconstructed copy number are considered. An exact numerical solution for the convex formulation has been provided by using the classical ADM framework to guarantee a global optimal solution. Another superior characteristic of the numerical method lies in its efficiency in avoiding the matrix inversion operation, which is commonly used in signal recovering problems.
We have demonstrated the capability of the proposed method to separate CNVs from other variations in wide data types, including synthetic, simulated and empirical sequencing data. The RD ratios obtained by our method demonstrate sparsity and smoothness, thus accurately identifying CNVs.
Figure 3.1: Experiment on simulated read ratios following a Poisson distribution with various mean values, followed by Gaussian noise degraded by various variances; The reads with Gaussian noise with (a) small value of variance (0.01); (b) intermediate value of variance (0.1); and (c) high variance (1) are shown in the first column. In comparison, the corresponding detected CNV curve are shown in the second column with the raw reads overlapped. The results by the proposed ADM-CNV is highlighted in black, while the one after CNV-TV is represented in red color. The proposed ADM-CNV can accurately detect all of the CNVs in different scenarios.
Figure 3.2: A sample CNV detection result with $d = 6e3$, $cp = 6$ by repeating the detection scheme 30 times. The horizontal and vertical axis represent the experiment index and the genomic coordinate, respectively. The fraction between the true positive ratio and false positive ratio is formatted as (TPR/FPR) and is shown in the parenthesis followed each method. The detected CNVs are highlighted in blue lines, while the ground truth, given by the starting and ending position with known CNVs at the reference genome is in red dots.
Table 3.2: A summary of a Venn-style analysis five-way tabulated Venn diagram obtained from sample of NA19240. The integer 1 represents common CNV blocks that can be detected by the corresponding method, while 0 means it is not detected. The proposed ADM-CNV could detect as many as 24,564 blocks, all being marked in DGV. It ranked the second position among the seven methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>ADM-CNV</th>
<th>CNV-TV</th>
<th>FREEC</th>
<th>CNV-seq</th>
<th>CNVnator</th>
<th>readDepth</th>
<th>CLImAT</th>
<th>Number of Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADM-CNV</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CNV-TV</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>FREEC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CNV-seq</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CNVnator</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>readDepth</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CLImAT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of Blocks</td>
<td>24544</td>
<td>9272</td>
<td>14435</td>
<td>24508</td>
<td>785590</td>
<td>11520</td>
<td>4865</td>
<td>8884</td>
</tr>
</tbody>
</table>
Figure 3.3: Chromosome 21 from NA19240. The blue curve is the read depth signal. The CNV regions are shown in red dots, and the green lines indicate the bounds of the zoom region. A zoom between the region within the two vertical green lines is displayed in Fig. 3.4.
Figure 3.4: The zoomed CNVs detected by ADM-CNV. The gray lines are the raw depth of the coverage data while the blue solid line is the reconstructed CN. The two horizontal green lines represent the global cutoff values for the copy number variation. Therefore, the ratio beyond the cutoff values is recognized as the CNV and is highlighted in red.
### Table 3.3: F-scores of top 10 CNVs detected by each method.

<table>
<thead>
<tr>
<th></th>
<th>CNV-TV</th>
<th>0.88</th>
<th>0.83</th>
<th>0.74</th>
<th>0.70</th>
<th>0.68</th>
<th>Avg. ± std.</th>
<th>0.6467</th>
<th>0.7120</th>
<th>0.7398</th>
<th>0.7514</th>
<th>0.7748</th>
<th>0.8792</th>
<th>0.9158</th>
<th>0.9572</th>
<th>0.9893</th>
<th>0.8845</th>
<th>0.9000</th>
</tr>
</thead>
<tbody>
<tr>
<td>FREEC</td>
<td>0.81</td>
<td>0.75</td>
<td>0.77</td>
<td>0.77</td>
<td>0.78</td>
<td>0.78</td>
<td>0.88 ± 0.011</td>
<td>0.72</td>
<td>0.85</td>
<td>0.90</td>
<td>0.93</td>
<td>0.96</td>
<td>0.99</td>
<td>0.88</td>
<td>0.83</td>
<td>0.81</td>
<td>0.81</td>
<td>0.88</td>
</tr>
<tr>
<td>CNV-seg</td>
<td>0.86</td>
<td>0.83</td>
<td>0.57</td>
<td>0.59</td>
<td>0.57</td>
<td>0.59</td>
<td>0.60 ± 0.012</td>
<td>0.71</td>
<td>0.82</td>
<td>0.84</td>
<td>0.85</td>
<td>0.87</td>
<td>0.88</td>
<td>0.88</td>
<td>0.83</td>
<td>0.81</td>
<td>0.83</td>
<td>0.88</td>
</tr>
<tr>
<td>FReE</td>
<td>0.88</td>
<td>0.88</td>
<td>0.78</td>
<td>0.84</td>
<td>0.86</td>
<td>0.88</td>
<td>0.94 ± 0.011</td>
<td>0.77</td>
<td>0.80</td>
<td>0.83</td>
<td>0.84</td>
<td>0.87</td>
<td>0.91</td>
<td>0.88</td>
<td>0.83</td>
<td>0.81</td>
<td>0.83</td>
<td>0.88</td>
</tr>
<tr>
<td>CLImAT</td>
<td>0.91</td>
<td>0.76</td>
<td>0.75</td>
<td>0.69</td>
<td>0.68</td>
<td>0.68</td>
<td>0.70 ± 0.013</td>
<td>0.71</td>
<td>0.75</td>
<td>0.73</td>
<td>0.71</td>
<td>0.69</td>
<td>0.66</td>
<td>0.64</td>
<td>0.62</td>
<td>0.59</td>
<td>0.57</td>
<td>0.60</td>
</tr>
<tr>
<td>readsph</td>
<td>0.87</td>
<td>0.72</td>
<td>0.77</td>
<td>0.77</td>
<td>0.78</td>
<td>0.78</td>
<td>0.88 ± 0.011</td>
<td>0.72</td>
<td>0.85</td>
<td>0.90</td>
<td>0.93</td>
<td>0.96</td>
<td>0.99</td>
<td>0.88</td>
<td>0.83</td>
<td>0.81</td>
<td>0.81</td>
<td>0.88</td>
</tr>
<tr>
<td>ADWin-CNV</td>
<td>0.99</td>
<td>0.90</td>
<td>0.88</td>
<td>0.86</td>
<td>0.89</td>
<td>0.89</td>
<td>0.92 ± 0.013</td>
<td>0.74</td>
<td>0.87</td>
<td>0.89</td>
<td>0.91</td>
<td>0.93</td>
<td>0.96</td>
<td>0.88</td>
<td>0.83</td>
<td>0.81</td>
<td>0.83</td>
<td>0.88</td>
</tr>
<tr>
<td>CNV-TV</td>
<td>0.88</td>
<td>0.83</td>
<td>0.74</td>
<td>0.70</td>
<td>0.68</td>
<td>0.68</td>
<td>0.64 ± 0.010</td>
<td>0.68</td>
<td>0.60</td>
<td>0.58</td>
<td>0.56</td>
<td>0.55</td>
<td>0.55</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>CNV-seg</td>
<td>0.86</td>
<td>0.83</td>
<td>0.57</td>
<td>0.59</td>
<td>0.57</td>
<td>0.59</td>
<td>0.60 ± 0.012</td>
<td>0.71</td>
<td>0.82</td>
<td>0.84</td>
<td>0.85</td>
<td>0.87</td>
<td>0.88</td>
<td>0.88</td>
<td>0.83</td>
<td>0.81</td>
<td>0.83</td>
<td>0.88</td>
</tr>
<tr>
<td>FReE</td>
<td>0.88</td>
<td>0.88</td>
<td>0.78</td>
<td>0.84</td>
<td>0.86</td>
<td>0.88</td>
<td>0.94 ± 0.011</td>
<td>0.77</td>
<td>0.80</td>
<td>0.83</td>
<td>0.84</td>
<td>0.87</td>
<td>0.91</td>
<td>0.88</td>
<td>0.83</td>
<td>0.81</td>
<td>0.83</td>
<td>0.88</td>
</tr>
<tr>
<td>CLImAT</td>
<td>0.91</td>
<td>0.76</td>
<td>0.75</td>
<td>0.69</td>
<td>0.68</td>
<td>0.68</td>
<td>0.70 ± 0.013</td>
<td>0.71</td>
<td>0.75</td>
<td>0.73</td>
<td>0.71</td>
<td>0.69</td>
<td>0.66</td>
<td>0.64</td>
<td>0.62</td>
<td>0.59</td>
<td>0.57</td>
<td>0.60</td>
</tr>
<tr>
<td>readsph</td>
<td>0.87</td>
<td>0.72</td>
<td>0.77</td>
<td>0.77</td>
<td>0.78</td>
<td>0.78</td>
<td>0.88 ± 0.011</td>
<td>0.72</td>
<td>0.85</td>
<td>0.90</td>
<td>0.93</td>
<td>0.96</td>
<td>0.99</td>
<td>0.88</td>
<td>0.83</td>
<td>0.81</td>
<td>0.81</td>
<td>0.88</td>
</tr>
</tbody>
</table>

*Values were highlighted in red and blue, respectively.*

Table 3.3: F-scores of top 10 CNVs detected by each method. For visual comparison, the best and the second best.
Table 3.4: The computational time of the proposed method and the other six competing CNV methods.

<table>
<thead>
<tr>
<th>Cut-off value</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.5: False positive rates to detect copy-number alterations is considered as a function of the cut-off value. The red circle shows that when the cut-off value is 0.2, the genome-wide false positive rate is minimized. Thus, the cut-off value is set to be 0.2.
<table>
<thead>
<tr>
<th>F-score 0.0-0.1</th>
<th>0.0-0.2</th>
<th>0.2-0.3</th>
<th>0.3-0.4</th>
<th>0.4-0.5</th>
<th>0.5-0.6</th>
<th>0.6-0.7</th>
<th>0.7-0.8</th>
<th>0.8-0.9</th>
<th>0.9-1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNV-TV</td>
<td>0.7837</td>
<td>0.0898</td>
<td>0.0408</td>
<td>0.0204</td>
<td>0.0163</td>
<td>0.0204</td>
<td>0.0163</td>
<td>0.0041</td>
<td>0.0082</td>
</tr>
<tr>
<td>FREEC</td>
<td>0.5</td>
<td>0.14</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.024</td>
</tr>
<tr>
<td>CNV-seq</td>
<td>0.54</td>
<td>0.21</td>
<td>0.08</td>
<td>0.05</td>
<td>0.03</td>
<td>0.017</td>
<td>0.017</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>ADM-CNV</td>
<td>0.6571</td>
<td>0.0737</td>
<td>0.0641</td>
<td>0.0417</td>
<td>0.0481</td>
<td>0.0417</td>
<td>0.0288</td>
<td>0.0321</td>
<td>0.0096</td>
</tr>
<tr>
<td>readDepth</td>
<td>0.6389</td>
<td>0.1389</td>
<td>0.0278</td>
<td>0.0417</td>
<td>0.0417</td>
<td>0.0</td>
<td>0.0417</td>
<td>0.0471</td>
<td>0.0278</td>
</tr>
<tr>
<td>CNVnator</td>
<td>0.7143</td>
<td>0.0476</td>
<td>0.0</td>
<td>0.0159</td>
<td>0.0476</td>
<td>0.0159</td>
<td>0.0159</td>
<td>0.00317</td>
<td>0.00467</td>
</tr>
<tr>
<td>CLImAT</td>
<td>0.4483</td>
<td>0.0345</td>
<td>0.0690</td>
<td>0.0345</td>
<td>0.1034</td>
<td>0.0690</td>
<td>0.0690</td>
<td>0.1034</td>
<td>0.0345</td>
</tr>
<tr>
<td>ADN-CNV</td>
<td>0.6771</td>
<td>0.0737</td>
<td>0.0417</td>
<td>0.0417</td>
<td>0.0417</td>
<td>0.0641</td>
<td>0.0641</td>
<td>0.0377</td>
<td>0.05</td>
</tr>
<tr>
<td>CNV-seg</td>
<td>0.74</td>
<td>0.05</td>
<td>0.0</td>
<td>0.024</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>FREEC</td>
<td>0.652</td>
<td>0.0278</td>
<td>0.0417</td>
<td>0.0417</td>
<td>0.0417</td>
<td>0.0</td>
<td>0.0417</td>
<td>0.0471</td>
<td>0.0278</td>
</tr>
<tr>
<td>CNV-ty</td>
<td>0.483</td>
<td>0.0737</td>
<td>0.0417</td>
<td>0.0417</td>
<td>0.0417</td>
<td>0.0641</td>
<td>0.0641</td>
<td>0.0377</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 3.5: Average distribution (in percentage) of F-scores of detected CNVs.
Chapter 4

Revealing Common Copy Number Patterns by a Total-variation Constrained Permutation Model

4.1 Introduction

Variations in DNA copy number carry important information on genome evolution and regulation of DNA replication in cancer cells. The rapid development of single-cell sequencing technology enables exploration of gene-expression heterogeneity among single cells, providing important information on cell evolution. Evolutionary relationships in accumulated sequence data can be visualized by adjacent positioning of similar cells so that similar copy-number profiles are shown by block patterns. However, single-cell DNA sequencing data usually have low amount of starting genome and therefore it is necessary to add an extra step of amplification to accumulate sufficient samples, introducing noise and making regular pattern-finding challenging.

In this chapter, we will propose to tackle this issue of recovering the hidden blocks within single-cell DNA-sequencing data through continuous sample permutations such that similar samples are positioned adjacently. The permutation is guided by the total variational norm of the recovered copy number profiles, and is continued until the total variational norm is minimized when similar samples are stacked together to reveal block patterns. 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.

The remainder of the chapter is as follows: In Section 4.2, the current researches on analyzing multiple genome sequencing samples are overviewed and categorized. Section 4.3 states the problem of block pattern finding for multiple SNS samples and
reformulate it into an optimal permutation searching problem, which is guided by an
ergy term to form similar patterns. Numerical solution tailored from the efficient
alternating direction minimization model is presented in Section 4.4. Extensive
experiments on both synthetic and real sns data are conducted to demonstrate the
performance of the proposed TCP in Section 4.5. Finally, the concluding remarks
are given in Section 4.6.

4.2 Overview of Current Methodologies for Mul-
tiple Genome Samples

Profiling of genome-wide copy-number landscapes has been conducted by the ap-
plication of next-generation sequencing technologies [16]. This strategy is popular
for genotyping, and can include the comprehensive characterization of copy-number
profiles by the generation of hundreds of millions of short reads in a single run [56].
Since next-generation sequencing uses bulk DNA from tissue samples, it provides
an average signal from millions of cells, and is thus of limited utility for the charac-
terization of tumor heterogeneity at the single-cell level.

Single-cell sequencing is a technique that has been developed to address key issues
in cancer studies, including measurement of mutation rates, tracing of cell lineages,
resolution of intra-tumor heterogeneity and elucidation of tumor evolution [59, 60].
Single-cell sequencing combines flow sorting of single cells, whole-genome amplifica-
tion and next-generation sequencing to characterize the genome-wide copy num-
ber in single cells. The current whole-genome amplification (WGA) techniques
consist of three major types, multiple-displacement amplification [42], degenerate-
oligonucleotide-primed polymerase chain reaction [94], and multiple-annealing looping-
based amplification cycling [115]. The threes inevitably introduce varying degrees
of amplification bias when the whole genome of a single cell is amplified to micro-
gram levels for next-generation sequencing [29, 84]. Besides, copy number profile
detection requires only sparse sequence coverage [6] and it makes a contribution
to intrinsic noise of single-cell sequencing data. Technical noise that results from amplification bias is over-dispersed compared to Gaussian noises, and differs from the noise that occurs in bulk sequencing, which does not involve amplification.

There are four strategies that use next-generation-sequencing data to detect genome copy number, including read-depth, read-pair, split read and de novo assembly methods [53]. Read-depth-based methods are arguably most popular for detection of copy-number variation. The CNV detection method, circular binary segmentation (CBS) [64], a statistical approach used in a single-cell sequencing protocol [5], is a modification of binary segmentation to translate a noisy intensity read depth signal into regions of equal copy number. Copynumber [63] combines least squares principles with a suitable penalization scheme for a given number of breakpoints and detects copy number profiles. Control-FREE [8] (control-free copy number and allelic content caller) uses least absolute shrinkage and selection operator (LASSO) regression to identify the breakpoint and detects copy number profiles. CNV-Seq [96] (copy number variation using shotgun sequencing) uses a Gaussian distribution to model read depth signal. CNAsseg [35] (copy number abnormality segmentation) employs a hidden Markov model (HMM) and Pearson’s $\chi^2$ statistic to segment read depth signal and get the copy number profile. Recent studies have shown that the intrinsic noise of read depth signal from single-cell sequencing data could be accurately characterized by negative binomial distribution [27]. NbCN-V was then developed to reconstruct the copy number profile in single-cell DNA sequencing data [103].

As single-cell-sequencing data accumulate, recent efforts have focused on simultaneously analyzing genome data from multiple samples to assemble and subgroup similar cells accurately and efficiently. This subgrouping will be helpful for the exploration of heterogeneity among single cells, providing important information in relation to cellular evolution. The simplest way to subgroup multiple samples into a block pattern is to sort their reads in ascending or descending order. However, the high levels of noise in single-cell-sequencing data negatively affect the reliability
of sequence reads and lead to inaccurate patterns of variation. Another popular techniques for finding a suitable representation of data in real applications is matrix factorization. It usually decomposes the expression measurements with linear basis matrix and its loading coefficients. Heuristically, this new formation is hoping to reveal the latent structure in the data. Besides, the reformation could reduce the dimensionality of the raw data and thus allows further analysis [33]. There are a few popular methods been used to reveal hidden structures in observed data [74, 100], such as social network clustering [97, 39] and mining of protein or gene interactions [104, 85]. However, matrix factorization suffers from two drawbacks. Firstly, finding accurate linear basis and its spanning dimensions remains a challenging problem. Secondly, matrix factorization has a high computational cost, which limits its application to high-dimensional data analysis.

To directly recover block patterns within alignments of sequencing data from multiple single cells that might otherwise be obscured by noise, our aim is to develop an efficient method to re-alignment of the sequencing data [105]. Principally, the proposed method involved continuous permutation of the data to achieve a pattern, in which similar samples are organized into clusters. This technique, named the total-variation constrained permutation (TCP) model, is based on the alternating direction method of multipliers. In TCP, the permutation is not stopped until the total variational norm of the recovered matrix is minimized, when similar samples are stacked together to reveal block structures.

The TCP method has two main advantages over previous methods. Firstly, the hidden structures are estimated directly, and little parameter pruning is needed. This high robustness facilitates its usage with various types of sequencing data with few manual interventions. Secondly, as the major operation involved is sample permutation, the TCP method, using the classic Kuhn-Munkres algorithm, is very efficient, greatly saving computational cost. This efficiency is particularly useful for the analysis of modern sequencing data, in which hundreds of thousands of signals need to be recovered.
4.3 Problem Modelling for Revealing Common Copy Number Patterns in SCS Data

Throughout this chapter, we let matrices be denoted by bold uppercase letters and vectors by bold lowercase letters. For a matrix $X$, we use $X_i$ and $X^j$ to represent its $i$-th row and $j$-th column, respectively. For a matrix $X \in \mathbb{R}^{m \times n}$, its Frobenius norm is $\|X\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}^2}$. The trace $\text{Tr}(X)$ of a square matrix $X$ is defined to be the sum of the elements on the main diagonal. The $l_1$-norm for a vector $x \in \mathbb{R}^m$ is defined by $\|x\|_1 = \sum_{i=1}^{m} |x_i|$. A structural $l_1/l_q$-norm for matrix $X$ is $\|X\|_{l_1,l_q} = \sum_{j=1}^{n} \|X^j\|_q$. In particularly, the $(TV,l_1)$-norm is the total variational norm with respect to each row of $X$, i.e., $\|X\|_{TV,l_1} = \sum_{j=1}^{n} \|X^j\|_{TV}$, where $\|X^j\|_{TV}$ is defined as $\|X^j\|_{TV} = \sum_{i=2}^{m} |x_{i,j} - x_{i-1,j}|$.

Let $B \in \mathbb{R}^{m \times n}$ represent a copy-number-variation profile dataset obtained from multiple samples, where $m$ is the number of samples and $n$ is the number of genes. Each entry $(i,j)$ records the copy number at probe $j$ of sample $i$ and the value of $(i,j)$ is denoted by $B_{i,j}$. The $i$-th row $B_i$ corresponds to a copy-number profile obtained from sample $i$. We propose to use the following model to describe a given data set:

$$B = PX + \epsilon$$  \hspace{1cm} (4.3.1)

where $X \in \mathbb{R}^{m \times n}$ denotes the permuted copy-number-variation signals by a permutation matrix $P$, and $\epsilon$ is measurement noise, assuming a zero-mean normal distribution. We should note that although the technical noise resulting from amplification bias is over-dispersed than Gaussian noise does. For the simplify, we take the Gaussian noise for demonstration. A future study will be needed to assess the impact of non-Gaussian noise.

Our goal is to find an approximated matrix $X$, in which similar copy-number signals are rearranged together. By the rearrangement, the estimated signal $X$ could reveal the hidden-block characteristics accurately.

To make the decomposition in (4.3.1) feasible, we set up two assumptions:
**Assumption 1**: For each sample, the copy numbers are normalized around zero such that $B(i, j) = 0$ corresponds to the gain of two copy number at position $j$ for the $i$-th sample. A copy number gain smaller than 2 (0 and 1) indicates deletion event, and copy number gain larger than 2 indicates duplication event. Thus the signal possesses sparsity.

**Assumption 2**: For a set of related samples, the copy number signals are likely to share similar patterns or linear correlation with each other.

The first assumption is generally valid in modern sequencing data because the variants, including single-nucleotide polymorphisms and copy-number variants, account for a small portion of the whole genome. The second assumption is the basic motivation for researchers to analyze multiple samples simultaneously.

Based on the above assumptions, we propose estimation of $X$ by minimizing the following energy:

$$\arg\min_{X,P} \|PX - B\|_F^2 + \alpha \|X\|_{TV,l_1}. \quad (4.3.2)$$

Here, we regularize $X$ with a structural norm ($TV,l_1$). The total variational norm with respect to row-wise aims to encourage the sample similarities, while the $l_1$ norm aims to preserve the sparsity of copy number variants column-wise induced by the preprocessing step. Therefore, the regularization term satisfies the two predefined assumptions. $\alpha$ is a trade-off constant used to balance the data fidelity and structural similarity.

### 4.4 Numerical Solution for the TCP Model

#### 4.4.1 Fast Numerical Solution by ADM

Minimization problems involving a total variational term are common in disparate areas, including signal processing and image recovery [62]. To obtain a sparse solution attributing to the total variational norm, a normal practice is to use the Lasso
numerical method to solve the problem [19, 112]. However, a common drawback is the highly demanding computational cost resulting from the necessity to calculate a matrix inversion and the introduction of too many slack variables.

Recently, research results [62, 112] have successfully demonstrated the efficiency of alternative direction of minimization (ADM) in solving image-restoration problems involving total variational norms. The numerical solution shows surprisingly fast speed and high accuracy. The image-restoration ADMs are variants of the classic Augmented Lagrangian method for optimization problems with separable structures and linear constraints, and they have been intensively studied in the optimization community. Inspired by this idea, we reformulated the problem in Eq. (4.3.2) into an optimization problem with favorably separable structures, thus enabling it to be efficiently solved using ADMs. The separable structure decomposition is particularly useful for sequencing-data analysis because of both its large dimensionality and the large amount of data involved. What is more, it dramatically reduces computational cost by overcoming the matrix-inversion burden required by Lasso and majority-minimization schemes [62, 112].

Mathematically, let \( Y = DX \) with \( D \) being the first-order difference matrix, the minimization problem (4.3.2) could be equivalently rewritten as,

\[
\arg \min_{X, P} ||PX - B||^2_F + \alpha ||Y||_{l_1, l_1},
\]

subject to:

\( Y = DX \)

The augmented Lagrangian of the minimization problem is given by:

\[
\mathcal{L}(P, Y, X) = \alpha ||Y||_{l_1, l_1} + ||PX - B||^2_F + < \lambda, DX - Y > + \frac{1}{2} \gamma ||DX - Y||^2_F
\]

where the parameters \( \lambda \) and \( \gamma \) are Lagrangian multiples.

This new formulation enables it to be decoupled into two separate sub-problems with variables \( X \) and \( Y \), respectively. Therefore, it can be solved in an iterative manner:
Step 1: Find

\[ X^{(k+1)} = \arg\min_X \{ \|PX - B\|_F^2 + <\lambda, DX - Y^{(k)} > + \frac{1}{2}\gamma\|DX - Y^{(k)}\|_F^2 \} . \] (4.4.4)

It is equivalent to:

\[ X^{(k+1)} = \arg\min_X \{ \text{Tr}\{(X - P^TB)^T(X - P^TB)\} + \lambda^T\text{Tr}\{DX\} + \frac{\gamma}{2}\text{Tr}\{(X^TDX - 2Y^{(k)^T}DX)\} \} \]

\[ = \arg\min_X \text{Tr}\{X^TX - 2B^TPX + \lambda^TDX + \frac{\gamma}{2}X^TD^TDX - \gamma Y^{(k)^T}DX\} . \]

It has an explicit solution:

\[ \hat{X}^{(k+1)} = (2I + \gamma D^TD)^{-1}(2P^TB - D^T\lambda + \gamma D^TY^{(k)}) . \]

Matrix \(D\) is a circulant matrix and thus can be diagonalized by Fourier transform as \(D = F^TKF\), where \(F\) is 2-D discrete Fourier transform, and \(K\) is a diagonal matrix containing the discrete Fourier transform coefficients of the difference operator \(D\).

It follows that

\[ \hat{X}^{(k+1)} = F^T(2I + \gamma K^TK)^{-1}F(2P^TB - D^T\lambda + \gamma D^TY^{(k)}) \] (4.4.6)

Step 2: Find

\[ Y^{(k+1)} = \arg\min_Y \{ \alpha\|Y\|_{l_1,l_1} + <\lambda, DX^{(k)} - Y > + \frac{1}{2}\gamma\|DX^{(k)} - Y\|_F^2 \} \]

\[ = \arg\min_Y \{ \alpha\|Y\|_{l_1,l_1} + \frac{\gamma}{2}\|Y - (DX^{(k)} + \frac{\lambda\gamma}{2})\|^2_F \} \]

Its analytical solution could be estimated by point-wise soft thresholding:

\[ Y^{(k+1)} = S_{\frac{\gamma}{2}}(DX^{(k)} + \frac{\lambda}{\gamma}) \] (4.4.7)

The soft-threshoing operator \(S_\alpha(x)\) is defined by:

\[ S_\alpha(x) = \begin{cases} 
  x - \alpha, & \text{if } x > \alpha \\
  0, & \text{if } -\alpha \leq x \leq \alpha \\
  x + \alpha, & \text{if } x < -\alpha
\end{cases} \]
Step 3: Find

\[ P^{(k+1)} = \arg \min_P \|PX^{(k)} - B\|_F^2. \quad (4.4.8) \]

This is a so-called linear assignment problem. It is generally solvable by using the classical Kuhn-Munkres algorithm in \( O(n^3) \) time [114, 113]. Kuhn-Munkres provided a permutation matrix to capture the relative ordering between a set of raw observations and the re-organized version of that set.

The above three steps are updated iteratively until converge to achieve the final solution of \( X \) and the optimal permutation matrix \( P \).

The prominent characteristic of the aforementioned numerical solution is that its separable structure enables rapid identification of a solution. After simple algebraic operations, explicit analytical solutions for the two sub-problems, Eq. (4.4.4) and Eq. (4.4.7) are obtained. In both of the sub-problems, the solutions involve low-cost calculations, which can be solved efficiently.

4.4.2 Convergence and Complexity Analysis

The convergence of TCP can be similarly established by Theorem 1.

**Corollary 2** The numerical scheme designed in section 4.4.1 for the problem of (4.3.1) converges to a minimizer.

**Proof:** The objective function (4.3.1) is convex with respect to \( X \). By denoting

\[
\begin{align*}
    f_1(Y) &= \|Y\|_1 \\
    f_2(X) &= Tr\{X^TX - 2B^TPX + \lambda^TDX + \frac{\gamma}{2}X^TDX - \gamma Y^{(k)^T}DX\}
\end{align*}
\]

one can verify that the two functions are closed, proper and convex. The matrix \( D \) is the first-order difference matrix and thus has full rank. The linear assignment problem is a linear programming problem and thus it convergent to a global minimizer.

According to Theorem 1, TCP is convergent to a minimizer of the objective function.
The optimal solution is obtained by updating the sequence of \( \{ X^{(k)}, Y^{(k)}, P^{(k)} \} \) alternately, which are given in Eq. (4.4.6), (4.4.7) and Eq. (4.4.8), respectively. In Eq. (4.4.6), the inverse of diagonal matrix \( (2I + \gamma K^T K) \) costs \( O(m) \). The products of \( F, F^T \) with the inverse diagonal matrix need \( O(m \log m) \). Similarly, the product of \( (2P^T B - D^T \lambda + \gamma D^T Y^{(k)}) \) costs \( O(m \log m) \), too. Thus, the total computation cost in Eq. (4.4.6) is \( O(m \log m) \).

In Eq. (4.4.7), the soft thresholding operator \( S_{\alpha \gamma} (DX^{(k)} + \lambda) \) costs \( O(mn) \), where \( n \) is the sample size. In Eq. (4.4.8), estimation of the permutation matrix is achieved by the classical Kuhn-Munkres algorithm and it costs \( O(n^3) \). In total, the computational cost is \( O(n^3) \).

However, it should be emphasized that the computational cost of the proposed TCP is not high, because empirical sequence samples usually have small sample sizes. Besides, earlier techniques to find common patterns often suffered from the drawback of the “curse of dimension”. For example, the complexity of the popular technique of matrix factorization [74, 100] is on the order of \( O(n \times k \times m) \), where \( m \) is the feature dimension and \( k \) is the dimension of the reduced space. This value is much larger than \( O(n^3) \) if \( n \ll m \).

### 4.5 Experimental Results

Experiments were conducted to demonstrate the performance of the proposed TCP model. The test datasets included both synthetic and empirical single cell DNA sequencing data. The first experiment was designed to test the discriminatory power of the proposed method of TCP to group similar samples and form uniform patterns on synthetic data. Various levels of Gaussian noise were introduced artificially into the synthetic data to evaluate the robustness of the proposed TCP. In the second type experiment, the TCP was tested on a single cell DNA sequencing data set of 100 breast-tumor cells to find their optimal organization, thus revealing block patterns.
4.5.1 Robustness and Accuracy Test over Noise Contaminations on Synthetic Data

A synthetic data set was created to test the permutation capability of the proposed TCP. The synthetic data consisted of 100 samples, drawn independently from four groups. The four groups were from different linearly expanded independent spaces and were created to be orthogonal. This configuration helped to create distinct block structure of the data. To test the robustness of the methods, Gaussian noise with zero mean value and three different variances was added to the data. The unmodified synthetic data are shown in Fig. 4.1(a), with an intermediate level of added noise in Fig. 4.1(b) and a high level of noise in Fig. 4.1(c). In each sub-figure of Fig. 4.1, each column (x-axis) is copy number reads along the genome, while each row (y-axis) is the sample. The synthetic data was drawn in heatmap and thus a larger value of copy number was hotter (red), while a smaller value of copy number was cooler (blue). The noise degradation was quantified by the signal-to-noise ratio (SNR). Fig. 4.1 shows, from top to bottom, the results of TCP after 100 and 300 iterations, as well as the final result. When the noise level was low, TCP quickly organized the similar samples to group together, indicating the presence of four distinct sub-populations, highlighted in different colors. With an intermediate level of noise contamination, TCP still performed well, producing four distinct patterns. With a high level of noise contamination, TCP achieved some partial organization of the samples, but was not able to reproduce the four distinct patterns exactly. The high SNR patterns were accurately recovered, while the low SNR patterns were prone to be merged with the noises.

4.5.2 Recovery Uniform Pattern Test over Synthetic Irrelevant Read Depth Signal

Real sequencing data usually contain information on millions of genes in each trial, thus non-zero copy numbers are very sparse if the normal copy number (two copies)
is normalized to zero. To evaluate the potential of the proposed TCP in more general cases, artificial-read-depth signals were randomly inserted into each sample of the previously analyzed synthetic data. In this way, the length of the data was expanded to simulate real sequence data. The inserted signals were normalized to follow a normal distribution with a mean value of zero. The modified synthetic data therefore contain a large number of irrelevant, noisy features, completely burying the four distinct patterns. TCP analysis of the synthetic data set containing 100, 1,000 or 10,000 artificial-read-depth signals is shown in Fig. 4.2. The analysis was conducted with different degraded-noise levels (indicated by the SNR values). A feature-selection scheme was applied to remove irrelevant read depths signal before using TCP to group similar uniform pattern together. Feature selection used the Laplacian Score, which measures the importance of a read depth by its power of locality preserving and performs well in unsupervised data analysis [30]. This analytical result of a synthetic data set with addition of 100 noisy features is drawn by heatmap in Fig. 4.2(a). Similar to previous example, each column (x-axis) was the copy number reads while each row (y-axis) represent the profile of a sample. This addition had a minor influence on the performance of TCP compared with the noise level in the sample. As the noise level was increased, the four patterns became less discernable, as patterns with values below the noise level were not identified. Addition of 1,000 read depth signals (Fig. 4.2(b)) produced a similar result to addition of 100 read depth signals (Fig. 4.2(a)), and the performance of TCP was still acceptable even with the addition of read depth 10,000 signals (Fig. 4.2(c)). In each case, the four patterns were largely revealed by TCP in the presence of low or intermediate noise levels, but pattern discrimination was unsatisfactory with a high level of noise.
Figure 4.1: TCP analysis of a synthetic data set. A synthetic data set consisting of 100 samples is created, and Gaussian noise with zero mean value and (a) low, (b) intermediate, or (c) high variance is added to the data. The SNRs are (a) 3.57, (b) 2.82, and (c) 1.55. The raw observation is in the first row. The results of 100 and 300 iterations of TCP are shown in the second and third row, respectively, while the final result of TCP is shown in the last row.

4.5.3 Comparison of TCP with Popular Methods on Synthetic Data

TCP was compared with three popular methods: Sparse non-negative matrix factorization (NMF), hierarchical clustering (HC) and Quick Sort. NMF is an un-
Figure 4.2: Recovery ability of TCP analysis. A synthetic data set consisting of 100 samples is created, and (a) 100, (b) 1,000, and (c) 10,000 artificial-read-depth signals are randomly inserted into the data. The raw data are shown in the top row. In each case, four levels of degraded noise are introduced as indicated by the SNRs, and the data are analyzed by TCP to recover uniform block patterns. In most cases, the four patterns are largely revealed by TCP in the presence of low or intermediate noise levels.

supervised method for finding an appropriate representation of non-negative data. NMF aims to decompose a non-negative data matrix into two non-negative factors, one of which is considered to be an informative but concise representation of the original data, whereas the other is the corresponding loading factor [38]. The first
Figure 4.3: Performance comparison of TCP with the other methods. TCP is compared with the Sparse NMF, Quick Sort and hierarchical clustering methods on synthetic data under contamination with (a) low noise, (b) intermediate noise, and (c) high noise, as demonstrated by the SNRs of the raw data. TCP shows a high level of robustness to noise degradation, recovering the four distinct patterns in the true signal under all three noise levels.

Figure 4.3: Performance comparison of TCP with the other methods. TCP is compared with the Sparse NMF, Quick Sort and hierarchical clustering methods on synthetic data under contamination with (a) low noise, (b) intermediate noise, and (c) high noise, as demonstrated by the SNRs of the raw data. TCP shows a high level of robustness to noise degradation, recovering the four distinct patterns in the true signal under all three noise levels.

factor can be thought of as the basis for a hidden linear space, whereas the second conveys important meta-expression information within that hidden space. NMF has been successfully applied to data classification, dimension reduction and clustering [49, 36, 58, 54]. To cater to real applications, various constraints are imposed
either on the feature or its loading matrix, or even on both. For example, Sparse NMF incorporates NMF with sparseness constraints on both the loading and feature matrices [33]. In hierarchical clustering, the similarity of sample was measured by Euclidean distance. The Quick Sort algorithm is a classic sorting method for organization of samples in ascending (or descending) order. The scheme uses the idea of “divide-and-conquer” to decompose large sample sets into multi-level subsets. Sorting is conducted on each subset and the results are combined to produce an ordered set. The quick sort algorithm was applied on the dataset to have the ordered pattern.

To compare Sparse NMF, Quick Sort, HC and TCP, they were applied to the synthetic data set with contamination by Gaussian noise with zero mean value and three different variances (Fig. 4.3). For each noisy scenario, the figure shows the true data with similar samples positioned together and four distinct patterns. The same data with added noise was consider as raw observation. It was analyzed by Sparse NMF, Quick Sort, hierarchical clustering and TCP to find hidden patterns. With a low level of noise contamination (Fig. 4.3(a)), all four methods accurately revealed the four distinct patterns. With an intermediate level of noise (Fig. 4.3(b)), Sparse NMF did not discriminate the four patterns, whereas the other three did. The Quick Sort did not clearly distinguish one block, as its signal level was similar to the noise level. The HC mistakenly mixed two low value blocks. The TCP method clearly revealed four patterns. Moreover, the recovered pattern had clear boundaries to its background. Such nice contrast was due to the enforced sparsity constraints within TCP model. When the noise level was high (Fig. 4.3(c)), TCP outperformed the other methods, and was the only method that produced identifiable patterns, demonstrating that TCP is highly robust to noise degradation.
4.5.4 Experiment on Empirical Single-cell-sequencing Tumor Data

To further assess the performance of the proposed method on empirical applications, a single-cell-sequencing data set of 100 cells was downloaded from NCBI and tested [59]. The sequencing samples were from high-grade (grade III), triple-negative (\(ER^-\), \(PR^-\), \(HER2^-\)) ductal carcinomas (T10). The sequencing workflow for these 100 single cells consists of flow sorting of single nuclei, whole-genome amplification, library construction and sequenced by Illumina Genome Analyzer [5]. The Illumina runs generated a total of \(1.1 \times 10^9\) reads, \(5.8 \times 10^{10}\) base calls (33.3 Gb downloads in sequence read archive (SRA) format) and were in low coverage. The proposed sequencing data preparation pipeline consists of four steps, reads mapping, duplicates removing, read depth(RD) computing and data normalization. After downloading the sequencing file from NCBI SRA, bowtie2 [41] alignment tool was employed to map the millions of short reads to the hg19 human reference genome [40] and Samtools was employed to remove potential PCR duplicates. RD signal was computed based on bins of variable sizes along the whole genome and normalized by locally-weighted polynomial regression and linear interpolation based on the GC content in each bin [5]. The data have been used to study the evolutionary dynamics and population structure of tumors in order to have a comprehensive view of the evolutionary process occurring in individual tumor cells [59]. The cells have been analyzed by fluorescence-activated cell sorting (FACS), and they consist of 47 diploids or pseudodiploids (2N), 24 hypodiploids (1.7N) and 29 aneuploids (3N or 3.3N). The diploid or pseudodiploid cells, on the whole, have a small number of copy-number variations, whereas the hypodiploids have narrow deletions, while the aneuploids have numerous copy-number duplications [59].

After the preprocessing steps, the normalized RD signals for the 100 cells were obtained. The TCP method was applied to the RD signals to reorganize their arrangement such that similar samples were in consecutive positions. For visual
comparison, the correlative heat map of RD signals and the rearrangement after TCP are shown in Fig. 4.4. Each row represents the RD signals of a single cell. The columns represent contiguous segments of genome. Analysis of this data set by TCP reflected that the cells have three blocks as shown in (Fig. 4.4(c)) that were not apparent in the RD signals (Fig. 4.4(a)), or by simple sorting scheme (Fig. 4.4(b)). The three blocks corresponded to the “ploidy” groups identified by FACS analysis. The hypodiploid (1.7N) subgroup showed narrow deletions and the aneuploid (3N or 3.3N) subgroup showed rich duplications. The first five cells in the TCP output (represented by the top five rows in Fig. 4.4(c)) contained both deletions and duplications, but these variations could not be merged into the hypodiploid or aneuploid subgroups, which is consistent with the ploidy level (2N) of these five cells by FACS. For comparison, the result by hierarchical clustering is shown in Fig. 4.4(d). Hierarchical clustering divides the 100 single-cells into three subgroups. When inspecting the ploidy information in each cluster, it was found that hierarchical clustering misclassified five cells. The misclassified samples were highlighted in arrowed red slack parenthesis in Fig. 4.4(d).

### 4.6 Chapter Summary

With the accumulation of multiple single-cell-sequencing data, accurate and efficient subgrouping of similar cells remains a problem, because of the introduction of noise by the techniques required to accumulate enough start materials, and because of the huge genome lengths. Subgrouping will be helpful to explore gene-expression heterogeneity among single cells, providing important information on cellular evolution. We have now addressed this problem by developing a novel method for continually permutating of the data so that similar samples are positioned adjacently. The permutation is guided by the total variational norm of the recovered matrix. Since the major operation involved is sample permutation, the method, which uses the classical Kuhn-Munkres algorithm, is very efficient, greatly saving computational cost. Experiments on both simulated and real single-cell-sequencing data have demon-
Figure 4.4: TCP analysis of empirical data. (a) The RD signals of 100 single-cell sequencing data; (b) result by quick sort scheme; (c) result after using the TCP to assemble similar samples at contingent positions; (d) result by hierarchical clustering. The data contains three subpopulations of (D) diploids or pseudodiploids; (A) aneuploids; and (H) hypodiploids. The result after TCP produces three blocks and is consistent with the previous characterization of the cells by FACS analysis [59]. From the top to the bottom, the ploidy of the clustered subpopulations in (d) is A, H, D, respectively. Five of the diploids or pseudodiploids are mis-classified to the hypodiploids by HC, as highlighted in red arrow.

Stratified the superiority of the proposed method in recovering the hidden structures of samples, with few manual interventions.
Chapter 5
Conclusions and Future Work

5.1 Conclusions

This thesis has addressed three issues in the detection of copy number variants from next-generation sequencing data: 1) establishing a preprocessing workflow for single-cell sequencing data, 2) having an efficient quadratic model for CNV detection with multiple norms constrained, and 3) obtaining a permutation-based model for recovering SNS block structures.

First, we introduced the development of sequencing technology and emphasized the promising applications of single-cell sequencing technologies. We also provided an extensive review of the important implications and studies on copy number variants. We also categorized and analyzed the current methodologies for detecting CNVs from NGS. Moreover, we established preprocessing workflows for both bulk and single-cell sequencing data in this chapter. The workflows that we built facilitate the analysis of data and ensure that high-quality processed data can be obtained.

Subsequently, we presented an original model for identifying CNVs from raw NGS profiles by formulating the problem into a changing point detection optimization. The new formulation possesses a unique and fundamental characteristic distinct from its peers, in that the two constraints of sparsity and smoothness of the reconstructed copy number are considered. An exact numerical solution for the convex formulation was provided by using the classical ADM framework to guarantee a global optimal solution. Another superior characteristic of the numerical method lies in its efficiency in avoiding the matrix inversion operation, which is commonly used in signal recovery problems.

We demonstrated the ability of the proposed method to separate CNVs from
other variations in a wide range of data types, including synthetic, simulated, and empirical sequencing data. The RD ratios obtained by our method demonstrated sparsity and smoothness, thus accurately identifying CNVs.

Thirdly, we proposed a novel method for continual permutation of the data so that similar samples are positioned adjacently. Such permutation is guided by the total variational norm of the recovered matrix. The proposed TCP model could recover block structures within multiple single-cell sequencing data. It thus helps to accurately and efficiently classify similar cells into subgroups, and could be used to explore gene expression heterogeneity among single cells, providing important information on cellular evolution. Since the major operation involved in the method is sample permutation, this approach, which uses the classical Kuhn-Munkres algorithm, is very efficient, greatly reducing computational cost. Experiments on both simulated and real single-cell sequencing data have demonstrated the superiority of the proposed method in recovering the hidden structures of samples, with few manual interventions.

Finally, both the ADM-CNV and TCP model employs the framework of ADM for numerical solution. The main advantages of ADM lies in its capability of decomposing a large problem into several subproblems. Theoretically, each subproblem is independent and allows for parallel computation. Besides, the subproblem of ADM, in many cases, turns out to be classical problem and thus popular or standard methods could be used to solve it. However, the ADM model converges slowly due to its iterative nature. A sophisticated and adaptive parameter pruning scheme is helpful to accelerate its convergence. This is one aim of our future work.
5.2 Future Work

In future work, we may explore some issues in three particular directions:

1. **Bias correction for SNS.** In Chapter 2, we established a pipeline for preprocessing of single-cell sequencing data. Amplification from a single cell to generate sufficient DNA for sequencing library construction is vital to the reliable sequencing. Unfortunately, current experimental methods are all hindered by non-uniform amplification bias, which results in certain genomic regions having higher coverage than others. We used a simple regression method, locally weighted scatterplot smoothing, to fit the number of short sequences within each bin according GC content to achieve bias correction. One area of further research is to estimate the statistical distribution of GC and thus correcting the GC bias quantitatively.

2. **Hyper-parameter pruning for ADM-CNV.** In Chapter 3, we introduced a direct scheme to estimate the hyper-parameters for ADM-CNV. One area of further research is to explore the other plausibility, such as L-curve and cross-validation based parameter correction. The other question is whether there exists an efficient and fast method to estimate the hyper-parameters from the sequencing data directly. If the answer is yes, we could accelerate and enhance the performance of ADM-CNV in CNV calling.

3. **Removing of un-stratified reads.** As described in Chapter 4, we tested the performance of TCP with unstratified reads on synthetic data. A few popular feature selection methods were used to combine with TCP to reveal subgroup patterns. It would be interesting to incorporate a feature selection scheme into TCP minimization, thus achieving the dynamic identification of subgroup patterns.

4. **Accelerating of the ADM Method for sequence analysis.** We used ADM model to solve the proposed minimization problem. A distinct advantage of ADM
lies in its capability of decomposing a large problem into several subproblems. However, the ADM model converges slowly due to its iterative nature. It would be valuable to design adaptive parameter pruning scheme into minimization, thus accelerating its convergence.
Bibliography


[8] V. Boeva, T. Popova, K. Bleakley, P. Chiche, J. Cappo, G. Schleiermacher,


C. Russ, C. Nusbaum, M. Meyerson, and E. S. Lander. High-resolution mapping


[19] J. Duan, J. Zhang, H. Deng, and Y. Wang. Comparative studies of copy num-
er variation detection methods for next generation sequencing technologies.

number variation from next generation sequencing data with total variation
penalized least square optimization. *IEEE International Conference on Bioin-
formatics and Biomedicine Workshops*, pages 3–12, 2011.

to discover copy number variation from short sequencing reads. *BMC Bioin-

[22] J. Eckstein and D. P. Bertsekas. On the douglasrachford splitting method and
the proximal point algorithm for maximal monotone operators. *Mathematical


[85] M. Sill, S. Kaiser, A. Benner, and A. Kopp-Schneider. Robust biclustering by


detection of copy number variants using read depth of coverage. *Genome

number alteration and loss of heterozygosity in impure and aneuploid tumor
samples using whole-genome sequencing data. *Bioinformatics*, 30(18):2576–
2583, 2014.

mization model for discovering copy number variants in single-cell sequencing

method for identifying recurrent dna copy number aberrations in cancer from

[105] Y. Zhang, Y. Cheung, and W. Su. A total-variation constrained permuta-
tion model for revealing common copy number patterns. *Scientific Reports*,

[106] Y. Zhang, Y. Cheung, B. Xu, and W. Su. Detection copy number variants from
ngs with sparse and smooth constraints. *Computational Biology and Bioin-

Deannncnv: a tool for online detection and annotation of copy number vari-
ations from whole-exome sequencing data. *Nucleic Acids Research*, 43(W1),
2015.

[108] Z. Zhang, K. Lange, R. Ophoff, and C. Sabatti. Reconstructing dna copy num-


CURRICULUM VITAE

Academic qualifications of the thesis author, Mrs. ZHANG Yue:

• Received the degree of Bachelor of Science from the Harbin Normal University, June 2000.

• Received the degree of Master of Science from the Harbin Institute of Technology, June 2003.

September 2017