Title
Efficient Statistical Models For Detecting And Analyzing Human Genetic Variations

Author
Wang, Zhanyong

Publication Date
2014

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Efficient Statistical Models For Detecting And Analyzing Human Genetic Variations

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Computer Science

by

Zhanyong Wang

2014

Abstract of the Dissertation

Efficient Statistical Models For Detecting And Analyzing Human Genetic Variations

by

Zhanyong Wang

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2014

Professor Eleazar Eskin, Chair

In recent years, the advent of genotyping and sequencing technologies has enabled human genetics to discover numerous genetic variants. Genetic variations between individuals can range from Single Nucleotide Polymorphisms (SNPs) to differences in large segments of DNA, which are referred to as Structural Variations (SVs), including insertions, deletions, and copy number variations (CNVs). Genetic variants play an important role in regulating human diseases and traits.

I first propose an efficient genotyping method which can accurately report the genotypes of thousands of individuals over a high-density SNP map at low cost. This method utilizes pooled sequencing technology and imputation. A probabilistic model, CNVeM, is then developed to detect CNVs from High-Throughput Sequencing (HTS) data. I demonstrate by experiment that CNVeM can estimate the copy numbers and boundaries of copied regions more precisely than previous methods.

Genome wide association studies (GWAS) have discovered numerous individual SNPs involved in genetic traits. However, it is likely that complex traits are influenced by interaction of multiple SNPs. I propose a two-stage statistical model, TEPAA, to reduce computational time greatly while maintaining almost

identical power to the brute force approach which considers all possible combinations of SNPs. The experiment on the Northern Finland Birth Cohort data shows that TEPAA achieved 63 times speedup.

Another drawback of GWAS is that rare causal variants will not be identified. Rare causal variants are likely to have been introduced in a population recently and are likely to be in shared Identity-By-Descent (IBD) segments. I propose a new test statistic to detect IBD segments associated with quantitative traits. I make a connection between the proposed statistic and linear models so that it does not require permutations to assess the significance of an association. In addition, the method can control for population structure by utilizing linear mixed models.

The dissertation of Zhanyong Wang is approved.

Nelson Freimer

Jason Ernst

Stott Parker

Eleazar Eskin, Committee Chair

University of California, Los Angeles

2014

To my parents He Wang and Xihua Cui and my wife Caiyun Liu

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my advisor, Professor Eleazar Eskin, for his support and insightful guidance during my Phd study. Professor Eskin guided me to the area of human genetics and he taught me the statistics, genetics and paper writing skills that are needed in the Phd study. Professor Eskin is friendly and patient to students and he has enthusiastic attitude toward research. From him, I not only learnt the knowledge and skills, but also how to do research. He encouraged me to collaborate with other scientists while developing my independent working ability. This will benefit my whole career.

I would also like to thank my Phd committee, Professor Jason Ernst, Professor Stott Parker and Professor Nelson Freimer. They provided valuable feedback and comments on my dissertation. I would also like to thank Professor Eran Halperin from Tel Aviv University, Professor Sagi Snir from University of Haifa and Professor Jose Lozano from Universyt of the Basque Country. We collaborated on several projects and they provided great support and help while I develop the statistical models and write the paper. Their expertise in statistics and human genetics enhanced my understanding of this area and inspired many research ideas. I would also like to thank Professor Jason Cong, Professor Reiman Glenn and Professor Jake Lusis for the collaboration projects, and my collaborators from their labs, Marco Vitanza, Yu-Ting Chen, Peng Wei, Mete Civelek and Brian Parker. I would thank Professor Laxmi Parida for her professional guidance while I did my internship at IBM Research.

I would like to give many thanks to my fellow graduate students in the Zarlab for their help and collaboration during the study. Especially, I would like to thank Wenyun Yang, Farhad Hormozdiari, Dan He, Serghei Mangul, Nathaniel Parrish, Buhm Han, Emrah Kostem, Nicholas Furlotte, Jae-Hoon Sul, Eun Yong Kang, Joanne Joo, Dat Duong and Michael Bilow.

Finally, I would like to thank my family. My parents and parents-in-law are really supportive for my study. I would like to especially thank my wife, Caiyun Liu. She supports me to pursue a Phd degree at the first beginning and is always supportive. She accompanied me for the whole phd study and always encouraged me during my hard time. Without the family's love, I would never make my Phd study so enjoyable and unforgettable.

Vita

2002–2006	Bachelor of Engineering in Computer Science, Shandong University, Shandong, China
2006–2008	Master of Philosophy in Computer Science, City University of Hong Kong, Hong Kong
2008–2009	Research Assistant, Computer Science Department, City University of Hong Kong, Hong Kong
2009	Research Assistant, Li Ka Shing Faculty of Medicine, Hong Kong University, Hong Kong
2009–2014	Doctoral Student and Research Assistant, Computer Science Department, University of California, Los Angeles, California
2012–2013	Teaching Assistant, Computer Science Department, University of California, Los Angeles, California

Publications

Zhanyong Wang, Jae Hoon Sul, Sagi Snir, Jose A Lozano and Eleazar Eskin, "Gene-Gene Interactions Detection Using a Two-Stage Model." *Proceedings of 18th International Conference on Research in Computational Molecular Biology (RECOMB 2014)*, Pittsburgh, United States, April 2-5, 2014.

Dan He, **Zhanyong Wang**, Laxmi Parida and Eleazar Eskin,"Inheritance Path based Pedigree Reconstruction Algorithm for Complicated Pedigrees." *5th ACM*

Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB 2014), Newport Beach, United States, 2014.

Zhanyong Wang*, Farhad Hormozdiari*, Wen-Yun Yang, Eran Halperin and Eleazar Eskin, "CNVeM: Copy Number Variation detection Using Uncertainty of Read Mapping." *Journal of Computional Biology.* 20(3):224-36, 2013.

Dan He, **Zhanyong Wang**, Buhm Han, Laxmi Parida, and Eleazar Eskin, "IPED: Inheritance Path-based Pedigree Reconstruction Algorithm Using Genotype Data." *Journal of Computational Biology.* 20(10): 780-791,2013.

Wen-Yun Yang, Farhad Hormozdiari, **Zhanyong Wang**, Dan He, Bogdan Pasaniuc and Eleazar Eskin, "Leveraging reads that span multiple single nucleotide polymorphisms for haplotype inference from sequencing data." *Bioinformatics.* 29(18):2245-2252, 2013.

Dan He, **Zhanyong Wang**, Laxmi Parida and Eleazar Eskin, "IPED: Inheritance Path based Pedigree Reconstruction Algorithm using Genotype Data." *Proceedings of 17th International Conference on Research in Computational Molecular Biology (RECOMB 2013)*, Beijing, China, April 7-10, 2013.

Zhanyong Wang*, Farhad Hormozdiari*, Wen-Yun Yang, Eran Halperin and Eleazar Eskin, "CNVeM: Copy Number Variation detection Using Uncertainty of Read Mapping." *Proceedings of 16th International Conference on Research in Computational Molecular Biology (RECOMB 2012)* , Barcelona, Spain, April 21-24, 2012.

Farhad Hormozdiari*, **Zhanyong Wang***, Wen-Yun Yang and Eleazar Eskin,

"Efficient Genotyping of Individuals using Overlapping Pool Sequencing and Imputation." *Conference Record of the Forty Sixth Asilomar Conference on Signals,Systems and Computers(Asilomar)*, pp.1023-1027, Pacific Grove, United States, Nov 4-7, 2012.

Jun Li, Panwen Wang, Xiaorong Liu, Ee Lyn Lim, **Zhanyong Wang**, Meredith Yeager, Maria Wong, Pak Chung Sham, Stephen Chanock and Junwen Wang, "GWASdb: a database for human genetic variants identified by genome-wide association studies." *Nuclear Acids Research*. 40:D1047-54, 2012.

Zhi-Zhong Chen, Lusheng Wang and **Zhanyong Wang**, "Approximation Algorithms for Reconstructing the Duplication History of Tandem Repeats." *Algorithmica*. 54(4):501-29, 2009.

Lusheng Wang, **Zhanyong Wang** and Wanling Yang, "Linked region detection using high-density SNP genotype data via the minimum recombinant model of pedigree haplotype inference." *BMC Bioinformatics*. 10:216, 2009.

Wanling Yang, **Zhanyong Wang**, Lusheng Wang, Pak-Chung Sham, Peng Huang and Yu Lung Lau, "Predicting the number and sizes of IBD regions among family members and evaluating the family size requirement for linkage studies." *European Journal of Human Genetics*. 16(12): p1535, 2008.

Guohui Lin*, **Zhanyong Wang***, Lusheng Wang, Yu-Lung Lau and Wanling Yang, "Identification of linked regions using high-density SNP genotype data in linkage analysis." *Bioinformatics*. 24(1):86-93, 2008.

Lusheng Wang, **Zhanyong Wang**, Zhizhong Chen, "Approximation Algorithms

for Reconstructing the Duplication History of Tandem Repeats." *Proceeding of 13th International Computing and Combinatorics Conference(COCOON 2007)*, Banff, Canada, July 16-19, 2007.

Wangsen Feng, **Zhanyong Wang**, and Lusheng Wang, "Identification of Distinguishing Motifs." *Proceeding of 18th Annual Symposium on Combinatorial Pattern Matching (CPM 2007)*, London, Canada, July 9-11, 2007.

* These authors contributed equally to the work.

CHAPTER 1

Introduction

Deoxyribonucleic acid (DNA) in the cells is the carrier of genetic information of all known living organisms. The DNA sequence is composed of a particular order of repeating units called nucleotides. There are four types of nucleotides in the DNA sequence, which are denoted as 'A', 'C', 'G' and 'T'. It is the sequence of these four nucleotides in DNA that encodes genetic information. Individuals within one species share as much as 99.9% of their DNA sequences. Completed in 2003, the Human Genome Project (HGP) determined the common sequence of the 3 billion nucleotides that make up the human genome [Con01] (In this article, we use DNA sequence and genome interchangeably).

Differences between the DNA sequences within one species are called genetic variations, or genetic variants. Genetic variation in the genome is present in many forms, including single nucleotide polymorphisms (SNPs, locations in the DNA sequence which are polymorphic in the population), small insertion-deletion polymorphisms and chromosomal structural variations (SVs), including insertions, deletions, and copy number variations (CNVs).

Genetic variations play an important role in regulating the human diseases, such as cancer, diabetes and so on. Individuals who carry a mutation at a certain variant may have high probability to develop the disease than those without the mutation. Also, human traits, such as height, hair color are also affected by the genetic variants. Thus, it is of crucial importance to study the role of genetic variants in human genome. The process of identifying genetic variants that are

associated with a certain trait or disease is referred to as an association study, which is fundamental in understanding diseases and searching for treatments.

The first step to conduct an association study is to collect the genotypes for a group of individuals over a set of genetic variants. The advent of high throughput sequencing technologies have ushered in a new era of genetic variant discovery. For the first time, we are able to collect thousands of individuals' genetic data at hundreds of thousands genetic markers, and perform a genome-wide association study (GWAS). SNPs have been a main interest in the field of genetics for the last decade, and they contain significant amounts of information for GWAS. High-Throughput Sequencing(HTS) technologies are rapidly decreasing the cost of obtaining genetic information. The cost for utilizing one of these technologies consists of a sample preparation step and a sequencing step of the prepared sample. The dramatic increase in the efficiency of the sequencing technology makes the cost of the sequencing step negligible for small target regions. Thus the main remaining cost is the sample preparation step. Using overlapping sequencing pools, where samples are mixed together into pools which are prepared and sequenced together has been shown to reduce cost significantly for collecting information on genetic variants that only occur in a few of the samples. These methods utilize ideas from compressed sensing. In Chapter 2, I extend this approach to utilize additional information from reference genetic variation datasets which provide the correlation structure between genetic variants. Utilizing this information, we can significantly increase the efficiency of overlapping pool sequencing.

CNVs are another important mediator for diseases and traits. The development of HTS technologies has also provided great opportunities to identify CNV regions in mammalian genomes. In a typical experiment, millions of short reads obtained from a genome of interest are mapped to a reference genome. The mapping information can be used to identify CNV regions. One important challenge in analyzing the mapping information is the large fraction of reads that can be

mapped to multiple positions. Most existing methods either only consider reads that can be uniquely mapped to the reference genome, or randomly place a read to one of its mapping positions. Therefore, these methods have low power to detect CNVs located within repeated sequences. In Chapter 3, I propose a probabilistic model, CNVeM, that utilizes the inherent uncertainty of read mapping. It uses maximum likelihood to estimate locations and copy numbers of copied regions, and implements an expectation-maximization (EM) algorithm. One important contribution of our model is that it can distinguish between regions in the reference genome that differ from each other by as little as 0.1%. As our model aims to predict the copy number of each nucleotide, it can predict the CNV boundaries with high resolution. We apply our method to simulated datasets and achieve higher accuracy compared to CNVnator, the state of art CNV detector. Moreover, we apply our method to real data from which we detected known CNVs. To our knowledge, this is the first attempt to predict CNVs at nucleotide resolution, and to utilize uncertainty of read mapping.

I have further extended the approach to apply it to cancer data. Recent studies have reported that CNVs are an important factor leading to cancer. In order to obtain the DNA-sequence of the cancer cells using HTS technologies, a biopsy is first conducted on the patient, where tumor tissue specimens were collected from the random sites of the tumor. However, analysis of tumor CNVs can be confounded by the presence of contaminating cells from normal surrounding stromal tissue, which have normal copy numbers. Another challenge is also the large fraction of reads that can be mapped to multiple positions. Most existing methods have low power to detect CNVs of tumor cells. In Chapter 4, I propose a probabilistic model, CNVmix, that utilizes the inherent uncertainty of read mapping to infer CNVs from tumor samples mixed with stromal cells. I propose a method to estimate the proportion of stromal cells in the contaminated samples. Then the information is utilized to estimate locations and copy numbers of CNV

regions.

GWAS studies have discovered numerous loci involved in genetic traits. Virtually all studies have reported associations between individual SNP and traits. However, current studies on certain complex diseases have also suggested that some SNPs influence diseases through interactions [WAP00, BSW05, YIS04]. One approach to detect interactions of SNPs is the brute force approach which performs a pairwise association test between a trait and each pair of SNPs. The brute force approach is often computationally infeasible because of the large number of SNPs collected in current GWAS studies. In Chapter 5, I propose a two-stage model, Threshold-based Efficient Pairwise Association Approach (TEPAA), to reduce the number of tests needed while maintaining almost identical power to the brute force approach. In the first stage, our method performs the single marker test on all SNPs and selects a subset of SNPs that achieve a certain significance threshold. In the second stage, we perform a pairwise association test between traits and pairs of the SNPs selected from the first stage. The key insight of our approach is that we derive the joint distribution between the association statistics of a single SNP and the association statistics of pairs of SNPs. This joint distribution allows us to provide guarantees that the statistical power of our approach will closely approximate the brute force approach. We applied our approach to the Northern Finland Birth Cohort data and achieved 63 times speedup while maintaining 99% of the power of the brute force approach.

Another drawback of GWAS is that rare causal variants will not be identified as they are rare in the population and the statistical power is low. Rare causal variants are likely to have been introduced in a population recently and are likely to be in shared Identity-By-Descent (IBD) segments. Recently, many methods have been developed to detect the IBD segments between a pair of individuals. These methods are able to detect very small shared IBD segments between a pair of individuals up to 2 centimorgans in length. This IBD information can be used

to identify recent rare mutations associated with phenotypes of interest. Previous approaches for IBD association were applicable to case/control phenotypes. In Chapter 6, I propose a novel and natural statistic for the IBD association testing, which can be applied to quantitative traits. A drawback of the statistic is that it requires a large number of permutations to assess the significance of the association, which can be a great computational challenge. We make a connection between the proposed statistic and linear models so that it does not require permutations to assess the significance of an association. In addition, our method can control for population structure by utilizing linear mixed models.

CHAPTER 2

Efficient Genotyping of Individuals using Overlapping Pool Sequencing and Imputation

2.1 Backgroud

In the past few years GWAS studies have successfully detected single SNPs associated with many diseases [MC09, MCC09]. Most of the associated SNPs have been collected using genotyping technologies [GSL05, MDL04]. Genotype chips typically collect SNPs with minor allele frequency (MAF) of at least 0.01; these SNPs are known as 'common SNPs' [MDL04]. However, the recent studies have shown that rare variants, or SNPs with MAF lower than 0.01, may play an important role in diseases [MCC09, EFG10]. Since rare SNPs outnumber common SNPs, one possibility is to increase the number of SNPs collected by the genotype chips. However, this will increase the cost of genotyping and is limited to collecting only previously discovered SNPs. Another approach is to apply imputation methods. In these methods, a standard genotype chip is used for genotyping. Then one of the existing computational methods [HDM09, MHM07] is used to infer the ungenotyped SNPs. However, the imputation methods may have error rates as high as 5% and are also limited to genotypes on previously discovered SNPs.

High throughput sequencing (HTS) technologies, where millions of fragments of DNA are obtained in each run of a sequencing machine [SJ08, Met08, Mar08], have the advantage that they can collect rare variants [BTL11, Ban10, LCY11,

ZBG12]. Although HTS costs are decreasing, compared to the cost of genotyping they are expensive. The cost of HTS technologies consists of a sample preparation step and a sequencing step of the prepared samples. Recent advances in sequencing technologies have dramatically decreased the cost of the sequencing step. Thus, the main cost is in the sample preparation step. Many studies require a large number of individuals to be sequenced in order to have sufficient statistical power to implicate variations in disease. However, due to cost constraints, it is impractical to sequence each individual separately because of the sample preparation costs. To reduce the sample preparation costs, one strategy is to use overlapping pools where multiple individuals (samples) are grouped into one pool and are sequenced together. The cost is reduced because only one sample preparation is necessary per pool. This reduces total number of sample preparation steps necessary for the study. In this strategy the pools are designed such that each individual sample is present in more than one pool. Utilizing the knowledge of which individual is in which pool and the results of the sequencing of the pools, in principle, it is possible to infer the genotypes of each individual.

In the past few years a number of studies have investigated the overlapping pool problem, which consists of two main subproblems. The first subproblem is to determine the design matrix, which indicates how individuals should be pooled together so that the detection of rare and common SNPs is possible with high accuracy [PP09, HHS08]. The second subproblem is to recover the sequence of each individual given the design matrix and the results of the sequencing. This problem is known as the decoding problem.

Prabhu et al. [PP09] introduced an elegant method to compute the design matrix using error-correcting codes. This method is able to recover a single rare-allele carrier from multiple pools. Using this design matrix the method can detect which individual carries a rare SNP with only $\log(N)$ pools where N is the number of individuals. However, this method fails for common SNPs.

Shental et al. [SAZ10] introduced a method to solve the decoding problem for rare SNPs. This method utilizes compressed sensing (CS) [CRT05], where they minimize sequencing errors and the predicted minor allele frequency for each SNP. However, this approach is also not applicable for common SNPs.

Recently, Golan et al. [GER12] utilized a pooling strategy where each individual is present in only one pool. In this strategy, different individuals have different abundance levels, which further reduces the cost of sample preparation.

Most similar to our approach, He et al. [HZP11] developed a likelihood method that solves the decoding problem using linear programming. They incorporated sequencing errors and the results of imputation of the common variants into their model. Imputation information provides information on the genotypes of common SNPs, although this information may be inaccurate. The key idea behind this approach is to combine the imputation information with the results of the sequencing to obtain more accurate genotypes. The method detects rare SNPs with high accuracy. This approach is also among the first methods to use the overlapping pooling approach to genotype common SNPs.

In this work, we present an approach for solving the decoding problem where the design matrix, the results of sequencing, and imputation information is given. We propose two methods to solve the decoding problem. The first method is based on compressed sensing (CS) [CRT05] which is an active research topic in many fields. The second method is a likelihood-based approach where we compute the maximum a posteriori (MAP) estimate. To solve both objective functions we use the proximal gradient descent algorithm which is an extension of the gradient descent algorithm. We use simulated data to illustrate the accuracy of each method. In our experiments we show that the MAP model has lower error rate than the CS model.

2.2 Methods

2.2.1 Problem Statement and Notations

Consider the scenario where a set of N individuals are to be sequenced and the length of the genome is L. We can denote the sequence of these individuals by a matrix

$$G = \{0, 1, 2\}^{N \times L}. \tag{2.1}$$

The element g_{ij} stands for the number of minor alleles at i-th individual's j-th genetic locus. We aim to reconstruct the matrix G utilizing HTS technologies. However, as we mentioned above, it is infeasible to sequence each individual separately in practice due to budget constraints, especially when N is large. We design a pooling schema to mix the samples into T pools. The design schema is represented by a matrix

$$A = \{0, 1\}^{T \times N} \tag{2.2}$$

where $a_{ij} = 1$ if and only if the j-th individual appears in the i-th pool. Under an error-free model, the number of minor alleles at each locus in each pool is given as

$$Y = AG. \tag{2.3}$$

Our objective is to reconstruct G from this equation. However, we do not observe Y directly, but only observe an estimate of Y from the sequencing data. From the results of sequencing the pools, we can estimate the number of minor alleles for each SNP in each pool using the read counts at each position.

In principle, we can obtain the genotypes by finding a solution to the set of equations $AX = Y$. However, as in our design schema $T < N$, the solution is not unique. We need other constraints or external data to accurately reconstruct the matrix G. One possible constraint is that for SNPs that are rare, the column vector corresponding to each SNP will contain mostly zeros. This idea is the

basis of most previous overlapping pool methods [HHF11, SAZ10]. Since the allele frequency can be inferred from the sequencing results, this constraint can be utilized to reconstruct the columns of G corresponding to rare SNPs.

For common SNPs, there is not enough information in only the sequencing data. However, the data can be augmented using information from an imputation method applied to genotype data collected from microarrays on a subset of the SNPs. Imputation methods can be utilized to infer the unmeasured common SNPs, where nearby SNPs are used to impute ungenotyped variants using the linkage disequilibrium(LD) structure of the genome. However, this process is inevitably noisy, especially when imputing SNPs of low allele frequencies or SNPs in regions of low LD. In addition, imputation methods can not infer genotypes for rare variants. Combining a pooling sequencing approach, we could provide more accurate genotypes for imputed SNPs and rare variants.

Denote the matrix imputed from genotyped common SNPs to be

$$M = \{0, 1, 2\}^{N \times L} \tag{2.4}$$

For the positions j that are neither genotyped nor imputed for individual i, we set the element $M_{ij} = 0$. We can represent the true genotype matrix G as a sum of imputed genotypes and a residual error matrix E, with $e_{ij} \in \{-2, -1, 0, 1, 2\}$. We note that for common variants, the residuals represent the errors in the imputation and for rare variants, since the corresponding column of M is all zeros, the genotypes of the rare variants are captured in the residuals E. Thus, we have

$$G = M + E. \tag{2.5}$$

2.2.2 Compressed Sensing (CS)

As rare variants appear only in a few individuals and the imputation precision is over 95%, we can assume that the difference between G and M is sparse. To solve

the formula $Y = AG$, it is natural that we introduce an $L1$ penalty. Then the optimization problem becomes:

$$\begin{aligned} \underset{G}{\text{minimize}} \quad & ||Y - AG||_F^2 + \lambda||G - M||_1 \\ \text{subject to} \quad & G_{ij} = \{0, 1, 2\},\ i = 1, \ldots, N; j = 1 \ldots L. \end{aligned} \tag{2.6}$$

A preliminary method. As we have an imputed matrix M and the difference between G and M is sparse, it is feasible that we enumerate all possible differences for each individual at one SNP. We enumerate all possible locations of the differences and mutate corresponding loci in M to recover $\hat{G}$ and find the one which minimize $||Y - A\hat{G}||_F^2 + \lambda||G - M||_1$.

A proximal gradient method. The disadvantage of the previous method is that enumerating all possible differences makes the method to be intractable for large set of inputs. We propose an alternative procedure to solve the objective function (2.6). In this method we relax the condition that the genotype for each individual at each given position is $\{0, 1, 2\}$. We assume the genotype to be a real number between zero and two (i.e. $0 \leq G_{ij} \leq 2$) and solve the objective function with the relaxed constraint. After obtaining the solution, we round the solution to an integer value. The main intuition behind this method is to use a gradient descent method. In the gradient descent G_k is the value of matrix G computed in the k-th step. In the first step of gradient descent we set G_1 equal to the imputation matrix, or any random matrix. In the k-th run of the method we set G_k equal to $G_{k-1} - t\nabla f$, where f is the objective function we want to optimize. We keep iteratively updating the value of G until we achieve convergence of the objective function. However, in the proximal gradient method after each step we project the computed value to the space which contains the L_1 regularizer. We utilize a constant step function $t \geq 0$, then we initialize the value of G_1 to the imputation matrix M. In the k-th step we set G_k equal to $G_{k-1} - t\nabla f$, where

$f = \frac{1}{\lambda}||Y - AG||_F^2$. Considering that our objective function has the L_1 regularizer we have to project G_k using the $\mathbf{prox}_t$ function. We keep improving the value of G until the value of objective function converge to an optimal solution. The pseudocode of the method is shown in Algorithm 1.

$$\mathbf{prox}_t(G_{ij}) = \begin{cases} G_{ij} - t & G_{ij} \geq M_{ij} + t \\ M_{ij} & M_{ij} + t \leq G_{ij} \leq M_{ij} - t \\ G_{ij} + t & G_{ij} \leq M_{ij} - t \end{cases}$$

Algorithm 1: Calculate G to minimize the Equation (2.6)

Require: $f = \frac{1}{\lambda}||Y - AG||_F^2$

pick a constant step $t \geq 0$

$G_0 \leftarrow M$

while Not converged **do**

$G_k \leftarrow \mathbf{prox}_t(G_{k-1} - t\nabla f)$

$k \leftarrow k + 1$

end while

2.2.3 Maximum a Posteriori

The difficulty of the CS method is to select the correct λ as different values of λ result in different G's, thus recovering the original G depends on the correct choice of λ. In this section we use the generative model of the data to obtain the desired objective function. We introduce a new variable $G' = \frac{G}{2}$ so that G'_{ij} is the probability that the i-th individual has a minor allele at the j-th SNP. Moreover, we have two matrices, $C_{T\times L}$ and $D_{T\times L}$, which represent the major and minor allele counts observed from the HTS data, respectively. C_{ij} indicates the major allele count for the j-th SNP in the i-th pool. Let ϵ indicates the sequencing error rate.

The probability of observing a minor allele for the j-th SNP in the i-th pool

is $\frac{\sum_{k=1}^{N} A_{ik} G'_{kj}}{\sum_{k=1}^{N} A_{ik}}$ when the sequencing error rate ϵ is zero. The denominator value $\sum_{k=1}^{N} A_{ik}$ is the normalization constant and A is the design matrix. In the case where the sequencing error is not zero, the probability of minor allele is $(1 - \epsilon)\frac{\sum_{k=1}^{N} A_{ik} G'_{kj}}{\sum_{k=1}^{N} A_{ik}} + \epsilon(1 - \frac{\sum_{k=1}^{N} A_{ik} G'_{kj}}{\sum_{k=1}^{N} A_{ik}})$. The first part represents the scenario where the sequenced allele is the minor allele. Incorporating the assumption that the sequencing error rate is ϵ, the probability of observing the minor allele will then shrink by a factor of $1 - \epsilon$. The second part represents the scenario where the sequenced allele is the major allele. Then a minor allele will only be observed in the case that sequencing error occurs, with a probability of ϵ. We can calculate the likelihood of observing the data as follows:

$$P(C, D|G) = \prod_{j=1}^{L} \prod_{i=1}^{T} ((1 - 2\epsilon)\frac{\sum_{k=1}^{N} A_{ik} G'_{kj}}{\sum_{k=1}^{N} A_{ik}} + \epsilon)^{C_{ij}} \times ((2\epsilon - 1)\frac{\sum_{k=1}^{N} A_{ik} G'_{kj}}{\sum_{k=1}^{N} A_{ik}} + 1 - \epsilon)^{D_{ij}}. \tag{2.7}$$

Let the imputation error be denoted as ϵ_{Im}. As the imputation error rate is less than 5%, the difference between G and M should be small. Given the imputation matrix M, it is natural to approximate the prior of the G as follow:

$$P(G|M) \propto (1 - \epsilon_{Im})^{N - |G-M|_1} (\frac{\epsilon_{Im}}{2})^{|G-M|_1}. \tag{2.8}$$

Considering the (2.7) and (2.8) one can compute the posteriori probability of the data according to Bayes rule:

$$P(G|C, D, M) \propto P(C, D|G) \times P(G|M). \tag{2.9}$$

Maximizing the posterior probability of the genotype matrix G in (2.9) is equal

to maximizing the following log probability with respect to G:

$$\begin{aligned} & |G - M|_1 \log \frac{\epsilon_{Im}}{2(1-\epsilon_{Im})} \\ + & \sum_{j=1}^{L} \sum_{i=1}^{T} \left(C_{ij} \log((1-2\epsilon) \frac{\sum_{k=1}^{N} A_{ik} G'_{kj}}{\sum_{k=1}^{N} A_{ik}} + \epsilon) \right. \\ + & \left. D_{ij} \log((2\epsilon - 1) \frac{\sum_{k=1}^{N} A_{ik} G'_{kj}}{\sum_{k=1}^{N} A_{ik}} + 1 - \epsilon) \right). \end{aligned} \tag{2.10}$$

We use the similar proximal gradient method as mentioned in section 2.2.2 to solve this maximum a posteriori (MAP) objective function.

2.3 Results

In order to assess the performance of our method, we designed a simulated framework where we can measure the accuracy of our method. We simulated the genotype of 50 individuals. For simplicity, we assess the accuracy on one SNP in each simulation. Since minor allele frequency (MAF) is a crucial factor that will affect the accuracy of pooling sequencing, we evaluated our method under various MAFs ranging from 1% up to 30%. For each SNP, whether the genotype for each individual is homozygous or heterozygous is randomly determined according to the pre-selected MAF. These genotypes serve as the true genotype G.

We also simulate the imputation matrix M. According to the current technology, the genotyping error rate is as low as 0.5% and the imputation error rate is 5%. We analyze two cases where the SNP is either genotyped or imputed. If the SNP is genotyped, the genotype of this SNP in the imputation matrix M is obtained from the corresponding cell in matrix G, with a chance of 0.5% of having an error. If the SNP is imputed, the genotype of this SNP in the imputation matrix M is obtained from the corresponding cell in matrix G, with a chance of 5% of having an error.

We use a random design matrix A. We simulate 15 pools and the probability

for any individual to appear in any pool is 50%. For each pool in order to simulate the read count of each locus we assume the number of reads generated from each position follows a Poisson distribution. Given that the sequencing coverage is m, on average each position is covered by m reads. Thus, the number of reads (K) at one SNP follows a Poisson distribution $K \sim Pois(m)$. If we take each sequencing event as a Bernoulli trial, then the number of reads carrying the minor allele follows a binomial distribution $Binom(K, MAF)$. Using this distribution, we simulate the number of reads carrying minor alleles and major alleles, respectively. The read counts are then considered as the output from pooling sequencing.

From the read counts, we reconstruct the matrix Y and use the methods in Section 2.2.2 and Section 2.2.3 to recover the matrix $\hat{G}$. The methods are tested under various MAF ranging from 1% up to 30%. Each scenario is repeated 50 times for genotyped SNPs and imputed SNPs respectively. For genotyped SNPs, our method always achieves almost 100% accuracy. Here we only demonstrate the accuracy for imputed SNPs. The results for imputed SNPs are shown in Figure 2.1. The accuracy depends on the MAF, especially for the MAP method. Both of our methods provide improvement in imputation accuracy. When the MAF is lower than 10%, the MAP method has higher than 99% accuracy.

2.4 Discussion

Many studies require thousands of individuals to be sequenced in order to have enough power to detect the SNPs involved in disease. This motivated the need for efficient methods for genotyping individuals.

One approach of efficient genotyping is through overlapped sequencing pools. A problem with traditional approaches is that they are unable to genotype common variants. However, for many cohorts that are currently being sequenced, genotype data collected from microarrays is already available and imputation

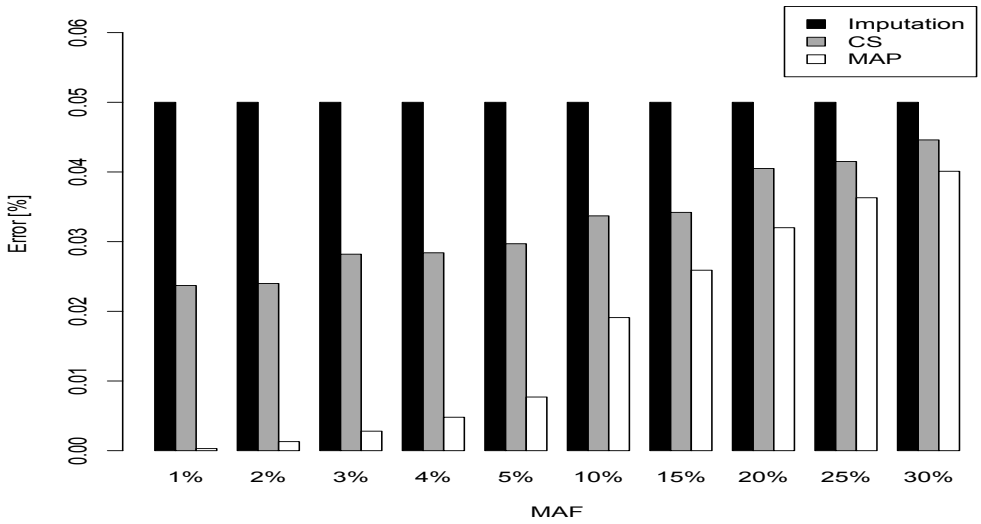


Figure 2.1: The error rate computed for each given method on the simulated data sets. We range the minor allele frequency (MAF) from 1% up to 30%. CS represents the results of the compressed sensing method proposed in this work and MAP represents the results from the maximum a posteriori method. MAP has the lowest error rare among all methods and as expected the error rate increases as the MAF increases.

methods are applied to obtain genotypes at uncollected SNPs. Unfortunately, these methods can only infer genotypes at previously known SNPs. Our approach utilizes imputation information in conjunction with sequencing pools to both infer the rare variants as well as improve the accuracy of the imputed genotypes.

We proposed two methods to solve the decoding problem for overlapping pools. The first method is based on the sparsity of rare variants and low error rate of imputation methods, where we use the compressed sensing technique to formulate the problem. The second method is based on a maximum likelihood approach. In this method we used the generative model of data to obtain the objective function. We simulated data sets for 50 individuals where these individuals are randomly

pooled into 15 pools reducing the sample preparation costs by a factor of 3. Using this simulation framework we illustrate the fact that both of our methods tend to have low error rate. Moreover, the MAP method tends to outperform the CS method. In all of our experiments MAP method had lower error rate compared to the CS method.

We note that our method is very accurate for MAF less than 10%. For higher MAF, our performance is still better than just using the imputed genotypes. One way to further increase the accuracy is to use a larger number of pools which increases the total cost of sequencing.

CHAPTER 3

Copy Number Variation Detection Using Uncertainty of Read Mapping

3.1 Background

Genetic variations between individuals can range from single nucleotide differences to differences in large segments of DNA. Variations on the nucleotide level are referred to as Single Nucleotide Polymorphisms (SNPs) and on the segment level are referred to as Structural Variations (SVs), including insertions, deletions, and copy number variations (CNVs). SVs and in particular CNVs, in which a large region of genome is deleted or duplicated, play an important role in the genetics of complex diseases and traits [IFR04, TSB05]. Many recent studies have shown a correlation between CNVs and different genomic disorders, ranging from brain related diseases (such as autism, schizophrenia and idiopathic learning disability [SLM07]) to cancers (e.g. non-small cell lung cancer [CHR05]).

Common methods to detect CNVs were until recently based on whole genome array comparative genome hybridization (ArrayCGH). In ArrayCGH, both a genome of interest (donor genome) and a reference genome are hybridized to a tiling array and the intensity ratio of the two genomes (donor/reference) provides an estimate of the copy number gain or loss [Car07, CLC08]. Although a powerful method to detect the presence of CNVs and to estimate copy numbers, the ArrayCGH approach is unable to identify the boundaries of CNVs with high resolution.

The development of high-throughput sequencing (HTS) technologies provides

great opportunities for detecting CNV regions. With HTS technologies, whole genome shotgun sequencing of one or more individuals becomes possible. Methods to detect the CNVs from short reads generated by HTS technologies can be categorized by two main ideas. The first category of methods divides the genome into small windows and the number of reads mapped to each specific window (read depth) is used as a proxy for the copy number of that window [AKM09, SKA10, SMA10, CGJ09, YXM09, Con10]. Alkan et. al [AKM09] used a set of fixed regions which are unique among all primates as control windows and calculated the average read depth for those regions. Then they scaled the results to predict the copy number of other windows. Simpson et al. [SMA10] used the same idea of splitting the genome into windows while incorporating read depth and heterozygous SNPs information (in inbred mouse) into a Hidden Markov Model (HMM). Adjacent windows with same copy number state are combined into one CNV region. Abyzov et al. [AUS11] developed a method for CNV discovery from statistical analysis of read depth. The method is based on the established mean-shift approach [CM02], which is a popular method in computer vision. This approach is able to detect the presence of large CNVs and the copy numbers. However, the resolution of this approach is limited by the size of the windows, which is typically at least one kilobase.

In the second strategy, "paired-end" reads, where "paired-end" refers to the two ends of the same segment of a DNA molecule, are used to detect CNVs. A short gap appears between the two paired-end reads and the distance of this gap is roughly fixed and known. The second class of approaches utilizes discordant paired-end reads, which are the reads mapped to the reference genome in an unexpected way [MFD10, HFE10, HAE09]. Discordant reads may indicate the presence of CNVs. Read depth information is then used to compute the copy number for each candidate CNV region [SKA10, AKM09]. Medvedev et al. [MFD10] introduced the idea of using both the read depth as well as the discordant reads

to detect CNVs. This method first clusters the discordant reads to identify the CNV boundary, after which it builds a "donor graph" representing the genome as segments of sequences connected by edges. Moreover, it uses maximum flow to estimate the most likely copy numbers for the donor genome. One limitation of this strategy is that it only detects CNVs in regions which are not repeat-rich. This may reduce the applicability of this method given the existence of many repeat-rich regions in the genome. Also, the CNVs may have complex structure. For example, there exist multiple copies of CNVs in the reference genome. This method can not detect variation within different copies.

Another important challenge for CNV detection lies in the uncertainty of read mapping. All of the mentioned methods use read depth information. The read depth is obtained by mapping the short reads to the reference genome and then calculating the number of reads within a region. However, a read can be mapped to multiple locations, although the read originated from one specific locus in the donor genome. This mapping uncertainty can be due to short read length, sequencing errors, and the presence of repetitive regions. With few exceptions [HHF11], most studies either consider all possible locations or randomly pick one mapping location, or even discard all such reads. These methods have difficulty in detecting CNVs with high accuracy, especially CNVs in repeat-rich regions.

In this work, we show that handling the uncertainty of read mapping can help us in predicting the copy number of CNVs, especially in repeat-rich regions. We propose a probabilistic model, CNVeM, that utilizes the uncertainty of read mapping. We use maximum likelihood to estimate locations and copy numbers of copied regions, and implement an expectation-maximization (EM) algorithm. One important contribution of our model is that we distinguish between similar copies of a region in the reference genome. We can predict exactly which copy of a region is duplicated or deleted utilizing the differences between copies and

handling uncertainty of read mapping.

In our model, we predict the copy number for each nucleotide and adjacent nucleotides with same copy number are then combined to form a full CNV region. In this way, we can detect the boundaries precisely and are able to predict small CNVs. To our knowledge, this is the first attempt to detect CNVs at nucleotide resolution and to distinguish between similar regions in the reference genome.

3.2 Methods

3.2.1 A Motivating Example

One important contribution of our method is that we distinguish between regions in the reference genome that differ from each other by a single nucleotide. Figure 3.1 illustrates an example. The reference genome has two nearly identical copies of a CNV region, represented as A and B. They only differ by one nucleotide as indicated in the figure, where the nucleotide is 'C' in region A and 'T' in region B. In the donor genome, region B is copied twice as B1 and B2. Reads $\{r1, r2, \ldots, r6\}$ are obtained from the donor genome as shown in the lower part of Figure 3.1 and then mapped to the reference genome as shown in the upper part of Figure 3.1. As shown in the figure, reads $\{r1, r3, r5\}$ can be mapped to both region A and B in the reference. However, read $\{r2\}$ can only be mapped to region A and reads $\{r4, r6\}$ can only be mapped to region B. If we assign a read to one of multiple mapping positions randomly following the traditional strategy, we would determine the copy number of both region A and B to be 1.5. However, in CNVeM, we use the EM algorithm to find the optimal solution. In each iteration, we assign a read to different mapping positions according to the distribution of copy numbers of those positions, and update the copy number of each position. Upon convergence, the EM algorithm assigns reads $\{r1, r3, r5\}$ to region A with probability 1/3 and to region B with probability 2/3. We correctly predict the

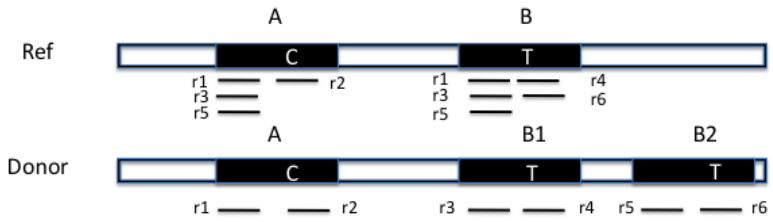


Figure 3.1: Similar copies of a CNV region exist in the reference genome. 'C' and 'T' are the only different nucleotide between region A and B. Reads $\{r1, r2, \ldots, r6\}$ are obtained from the donor genome as shown in the lower part of the figure. Furthermore, these reads can be mapped to the reference genome as shown in the upper part of the figure.

copy number of region A to be 1 and copy number of region B to be 2.

3.2.2 The Generative Model

We use short read information from HTS technologies to detect copy number variants. Let $\mathcal{G} = (g_1, g_2, \ldots, g_K)$ be K continuous nucleotides in the reference genome, where g_i is the i^{th} nucleotide. We assign the copy number of each nucleotide in the reference genome to be 1. The donor genome is also composed of these nucleotides. However, large regions of the genome can be either deleted or duplicated and thus the copy number can be changed. For each nucleotide g_i, we denote the copy number as C_i in the donor genome. If $C_i < 1$, we call it a copy loss. If $C_i > 1$, we call it a copy gain. $\mathcal{C} = (C_1, C_2, \ldots, C_K)$ can be interpreted as the copy number vector of the donor genome. For most nucleotides, the copy numbers are the same in the donor genome and in the reference genome. So one can assume that the length of donor genome is the same as the length of the reference genome, i.e. $\sum_{i=1}^{K} C_i = K$. We define vector $(\frac{C_1}{K}, \frac{C_2}{K}, \ldots, \frac{C_K}{K})$ to be the normalized copy number vector of the donor genome.

Using HTS technology, millions of short reads are sampled from the donor genome. We assume that a read r_j of length l is generated by randomly picking a position i from $\mathcal{G}$ according to distribution $\mathcal{C}/K$, and then copying l consecutive positions starting from position i. The copying process is error-prone, with known probability ϵ for a sequencing error rate at any position of the read. This process is repeated until we have a set of N reads $\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$. The objective is to infer $\mathcal{C} = (C_1, C_2, \ldots, C_K)$ from $\mathcal{R}$. Since the reads are mapped to the reference genome, mapping information is utilized to infer CNVs.

In our model, each read r_j is sequenced starting from one position in the donor genome. As we assume that the donor genome is obtained from the reference genome by alternating the copy number of some regions, each position in the donor genome "originates" from a nucleotide in the reference genome. Consequently, each read originates from a position in the reference genome. If a region in the reference genome is duplicated in the donor genome, any read generated from the duplicated segments of the donor genome originates from a unique position in the reference genome. $\mathcal{Z} = (Z_1, Z_2, \ldots, Z_N)$ is the origin for each read in the reference genome, where $Z_j \in \{1, 2, \ldots, K\}$. We then define the following likelihood model of all reads given copy number $\mathcal{C}$ and reference genome $\mathcal{G}$:

$$P(\mathcal{R}|\mathcal{C}, \mathcal{G}) = \prod_{j=1}^{N} P(r_j|\mathcal{C}, \mathcal{G}) = \prod_{j=1}^{N} \sum_{i=1}^{K} P(r_j, Z_j = i|\mathcal{C}, \mathcal{G}). \tag{3.1}$$

Here the first equality follows from the fact that the probability that read set $\mathcal{R}$ is composed of independent probabilities of all the reads, and the second equality follows from the fact that the read probability is equal to the marginalization of read mapping uncertainty, i.e., $P(r) = \sum_i P(r, Z = i)$.

The interpretation of the above probability definition $P(r_j, Z_j = i|\mathcal{C}, \mathcal{G})$ is straightforward: the probability of j-th read originating from i-th position of the reference genome, given the copy numbers and reference genome. We can further

expand this probability as follows:

$$P(r_j, Z_j = i|\mathcal{C}, \mathcal{G}) = P(Z_j = i|\mathcal{C})P(r_j|Z_j = i, \mathcal{G}). \tag{3.2}$$

The equality follows from the fact that the read origin Z is independent of reference genome $\mathcal{G}$ and the sequence of read r is independent of copy number $\mathcal{C}$. We define the first term $P(Z_j = i|\mathcal{C}) = C_i/K$ to be the probability of read r_j originating from position i. For each position i and read r_j, we have a probability $P(r_j|Z_j = i, \mathcal{G})$, which stands for the probability of observing read sequence r_j given that the origin of read r_j is position i. We can write $P(r_j|Z_j = i, \mathcal{G})$ as

$$P(r_j|Z_j = i, \mathcal{G}) = \prod_{x=1}^{l} \gamma(g_{i+x-1}, r_j^x) \tag{3.3}$$

and

$$\gamma(g_{i+x-1}, r_j^x) = \begin{cases} \epsilon/3 & \text{if } r_j^x \neq g_{i+x-1} \\ 1-\epsilon & \text{otherwise} \end{cases}$$

where r_j^x stands for the x-th nucleotide of read r_j, and the l consecutive nucleotides starting from position i in the reference genome are g_i, g_{i+1}, ..., g_{i+l-1}. In practice, for each read r_j, the probability $P(r_j|Z_j = i, \mathcal{G})$ will be close to zero for all but a few positions, which are reported by the mapping methods.

We also take the prior probability of the donor genome into consideration. As we assume the donor genome sequence can be obtained by either deleting or duplicating large regions of nucleotides from the reference genome, adjacent positions will have similar copy numbers in the donor genome. Then, in our probabilistic model, it is natural to assume that the copy number of the current nucleotide is only dependent on the previous nucleotide. We have $P(\mathcal{C}) = P(C_1, C_2, \ldots, C_K) = P(C_1)\prod_{i=2}^{K} P(C_i|C_{i-1})$.

Using Bayes rule, we can get the posterior probability of $\mathcal{C}$ given the read set

$\mathcal{R}$ and reference genome $\mathcal{G}$:

$$\begin{aligned} P(\mathcal{C}|\mathcal{R},\mathcal{G}) &\propto P(\mathcal{R}|\mathcal{C},\mathcal{G})P(\mathcal{C}) \\ &\propto \left(\prod_{j=1}^{N}\sum_{i=1}^{K}\frac{C_i}{K}P(r_j|Z_j=i)\right)\times\left(P(C_1)\prod_{i=2}^{K}P(C_i|C_{i-1})\right). \end{aligned} \tag{3.4}$$

3.2.3 Optimization

Maximizing the posterior probability of copy number $\mathcal{C}$ in (3.4) is equal to maximizing the following log probability with respect to $\mathcal{C}$:

$$\sum_{j=1}^{N}\left(\log\sum_{i=1}^{K}\frac{C_i}{K}P(r_j|Z_j=i)\right)+\log\left(P(C_1)\prod_{i=2}^{K}P(C_i|C_{i-1})\right). \tag{3.5}$$

In this section, we illustrate a lower-level description of our method. In order to make the above objective function simpler we eliminate the constraint $\sum_{i=1}^{K}C_i = K$ by introducing a penalty function $g(\mathcal{C}) = K - \sum_{i=1}^{K}C_i$, which prevents the $C_i's$ from growing unbounded (the above objective function will have a higher value if the $C_i's$ grow larger). Incorporating the penalty function, our objective function now becomes

$$\sum_{j=1}^{N}\left(\log\sum_{i=1}^{K}\frac{C_i}{K}P(r_j|Z_j=i)\right)+\log\left(P(C_1)\prod_{i=2}^{K}P(C_i|C_{i-1})\right)+\delta\left(K-\sum_{i=1}^{K}C_i\right). \tag{3.6}$$

where δ is a penalty function coefficient (We set $\delta = \frac{N}{K}$ in our experiments, from which we achieve best results). We optimize the objective function (3.6) through an expectation-maximization (EM) algorithm. The algorithm iteratively applies the following two steps until convergence.

Expectation-step: Estimate the posterior probability of each read origin under the current estimate of $\mathcal{C}^{(t)}$:

$$\begin{aligned} P(Z_j=i|r_j) &= \frac{1}{P(r_j)}P(r_j|Z_j=i)P(Z_j=i|\mathcal{C}^{(t)},\mathcal{G}) \\ &= \frac{P(r_j|Z_j=i)\mathcal{C}_i^{(t)}}{\sum_{k=1}^{K}P(r_j|Z_j=k)\mathcal{C}_k^{(t)}}. \end{aligned} \tag{3.7}$$

We can then calculate the expected value of the log objective function, with respect to the posterior probability of $\mathcal{Z}$ using the current estimate of $\mathcal{C}^{(t)}$:

$$\begin{aligned} Q(\mathcal{C}|\mathcal{C}^{(t)}) &= \sum_{j=1}^{N}\sum_{i=1}^{K} P(Z_j = i|r_j) \log\left[\frac{C_i}{K} P(r_j|Z_j = i)\right] + \log P(\mathcal{C}) + \delta(K - \sum_{i=1}^{K} C_i) \\ &= \sum_{j=1}^{N}\sum_{i=1}^{K} \log\left(\frac{C_i}{K}\right)^{P(Z_j=i|r_j)} + \sum_{j=1}^{N}\sum_{i=1}^{K} \log P(r_j|Z_j = i)^{P(Z_j=i|r_j)} \\ &\quad + \log P(\mathcal{C}) + \delta\left(K - \sum_{i=1}^{K} C_i\right). \end{aligned} \tag{3.8}$$

Maximization-step:

We find the vector $\mathcal{C}^{(t+1)}$ that maximizes the above function:

$$\mathcal{C}^{(t+1)} = \arg\max_{\mathcal{C}} Q(\mathcal{C}|\mathcal{C}^{(t)}). \tag{3.9}$$

In each iteration of the EM algorithm, both $C^{(t)}$ and $P(r_j|Z_j = i)$ are fixed values, so $P(Z_j = i|r_j)$ is a fixed value within the each iteration. Furthermore,

$$\sum_{j=1}^{N}\sum_{i=1}^{K} \log P(r_j|Z_j = i)^{P(Z_j=i|r_j)}$$

is also a fixed value within one single iteration. Then, maximizing the above function reduces to finding

$$\begin{aligned} \mathcal{C}^{(t+1)} &= \arg\max_{\mathcal{C}} \left(\sum_{j=1}^{N}\sum_{i=1}^{K} \log\left(\frac{C_i}{K}\right)^{P(Z_j=i|r_j)} + \log P(\mathcal{C}) + \delta(-\sum_{i=1}^{K} C_i)\right) \\ &= \arg\max_{\mathcal{C}} \log\left(P(\mathcal{C}) \times \prod_{i=1}^{K}\left(\frac{C_i}{K}\right)^{\sum_{j=1}^{N} P(Z_j=i|r_j)} \times e^{\delta(-\sum_{i=1}^{K} C_i)}\right) \\ &= \arg\max_{\mathcal{C}} \log\left(P(\mathcal{C}) \times \prod_{i=1}^{K}\left(\left(\frac{C_i}{K}\right)^{d_i} \times e^{-\delta C_i}\right)\right) \\ &= \arg\max_{\mathcal{C}} \log\left(P(C_1)\left(\frac{C_1}{K}\right)^{d_1} \times e^{-\delta C_1} \times \prod_{i=2}^{K}\left(P(C_i|C_{i-1})\left(\frac{C_i}{K}\right)^{d_i} \times e^{-\delta C_i}\right)\right) \end{aligned} \tag{3.10}$$

where

$$d_i = \sum_{j=1}^{N} P(Z_j = i|r_j).$$

We solve the M-step using dynamic programming. Denote the objective function in the M-step to be

$$f \;=\; \log\left[P(\mathcal{C}) \times \prod_{i=1}^{K}\left(\left(\frac{C_i}{K}\right)^{d_i} \times e^{-\delta C_i}\right)\right]. \tag{3.11}$$

Then we define $f(k, x)$ to be the maximum function value for the first k positions when the copy number of kth position is $C_k = x$. Now we design the dynamic programming solution as indicated in Equation (3.12).

$$f(k,x) = \begin{cases} \log[P(C_k = x) \times (\frac{C_k}{K})^{d_k} \times e^{-\delta C_k}] & \text{if } k = 1 \\ \max_{C_{k-1}} \quad \{f(k-1, C_{k-1}) + \log[P(C_k|C_{k-1})]\} & \\ \qquad\qquad + \log[(\frac{C_k}{K})^{d_k} \times e^{-\delta C_k}] & \text{otherwise} \end{cases} \tag{3.12}$$

We prove that the above dynamic programming solution returns the global optimal solution for the objective function in (3.11) as follows.

Lemma 3.2.1. *The objective function in (3.11) is solved optimally using the dynamic programming mentioned in (3.12) .*

Proof. We recall $f(i, x) = \max_{C_1, C_2, \cdots C_i - 1} \log[P(C_1, C_2, \cdots C_{i-1}, C_i = x) \times \prod_{j=1}^{i} (\frac{C_i}{K})^{d_i} \times e^{\delta C_i}]$ where $d_j = \sum_{l=1}^{N} P(Z_l = j|r_l)$. Moreover, $f(i, x)$ is the maximum value of the copy number for the first $i - 1$ positions and the copy number of position i is x ($C_i = x$). Using the above definition we drive $f(i + 1, y)$:

$$
\begin{aligned}
f(i+1,y) &= \max_{C_1,C_2,\cdots C_i} \log\left[P(C_1,C_2,\cdots C_i,C_{i+1}=y)\times\prod_{j=1}^{i+1}\left(\frac{C_j}{K}\right)^{d_j}\times e^{-\delta C_j}\right] \\
&= \max_{C_1,C_2,\cdots C_i} \log\left[P(C_1,C_2,\cdots C_i,C_{i+1}=y)\times\frac{C_{i+1}}{K}e^{\delta C_{i+1}}\times\prod_{j=1}^{i}\left(\frac{C_j}{K}\right)^{d_j}\times e^{-\delta C_j}\right] \\
&= \max_{C_1,C_2,\cdots C_i} \log\Bigg[P(C_1,C_2,\cdots C_i)P(C_{i+1}=y|C_1,C_2,\cdots C_i)\times\frac{C_{i+1}}{K}e^{-\delta C_{i+1}} \\
&\qquad\qquad \times\prod_{j=1}^{i}\left(\frac{C_j}{K}\right)^{d_j}\times e^{-\delta C_j}\Bigg] \\
&= \max_{C_1,C_2,\cdots C_i} \log\Bigg[P(C_1,C_2,\cdots C_i)P(C_{i+1}=y|C_i)\times\frac{C_{i+1}}{K}e^{-\delta C_{i+1}} \\
&\qquad\qquad \times\prod_{j=1}^{i}\left(\frac{C_j}{K}\right)^{d_j}\times e^{-\delta C_j}\Bigg] \\
&= \max_{C_1,C_2,\cdots C_i} \log\Bigg[P(C_1,C_2,\cdots C_i)\times\prod_{j=1}^{i}\left(\frac{C_j}{K}\right)^{d_j}\times e^{-\delta C_j}P(C_{i+1}=y|C_i) \\
&\qquad\qquad \times\frac{C_{i+1}}{K}e^{-\delta C_{i+1}}\Bigg] \\
&= \max_{C_1,C_2,\cdots C_i} \log\Bigg[P(C_1,C_2,\cdots C_i)\times\prod_{j=1}^{i}\left(\frac{C_j}{K}\right)^{d_j}\times e^{-\delta C_j}P(C_{i+1}=y|C_i) \\
&\qquad\qquad \times\frac{C_{i+1}}{K}e^{-\delta C_{i+1}}\Bigg] \\
&= \max_{C_i}\max_{C_1,C_2,\cdots C_{i-1}} \log\Bigg[P(C_1,C_2,\cdots C_i)\times\prod_{j=1}^{i}\left(\frac{C_j}{K}\right)^{d_j} \\
&\qquad\qquad \times e^{-\delta C_j}P(C_{i+1}=y|C_i)\times\frac{C_{i+1}}{K}e^{-\delta C_{i+1}}\Bigg] \\
&= \max_{C_i}\max_{C_1,C_2,\cdots C_{i-1}} \log\left[P(C_1,C_2,\cdots C_i)\times\prod_{j=1}^{i}\left(\frac{C_j}{K}\right)^{d_j}\times e^{-\delta C_j}\right] \\
&\qquad\qquad + \log\left[P(C_{i+1}=y|C_i)\times\frac{C_{i+1}}{K}e^{-\delta C_{i+1}})\right] \\
&= \max_{C_i}\left[f(i,C_i)+\log(P(C_{i+1}=y|C_i)\times\frac{C_{i+1}}{K}e^{-\delta C_{i+1}})\right].
\end{aligned}
$$

□

The maximum value of the objective function in the M-step is then $\max_x f(K, x)$. Using a backtracking process, we find the vector $C = (C_1, C_2, \ldots, C_K)$ that maximizes function f in the M-step. By iteratively running the E-step and M-step, we achieve local optimal solution.

3.2.4 Implementation

This optimization process requires an initial input of copy numbers. Different initial inputs will affect the convergence time. To achieve better performance, it is important to start with a "*good*" initial guess. In order to obtain a good initial input, we split the genome into non-overlapping bins of 300 bp. All nucleotides within one bin share the same copy number. Using a similar model as in (3.1), we get a initial guess of copy numbers by optimizing the objective function (3.13).

$$P(\mathcal{R}|\mathcal{C},\mathcal{G}) \;=\; \prod_{j=1}^{N} P(r_j|\mathcal{C},\mathcal{G}) = \prod_{j=1}^{N} \sum_{i=1}^{\lceil K/300 \rceil} \frac{C_i \times 300}{K} P(r_j | Z_j \in i\text{-th bin}) \tag{3.13}$$

where $P(r_j|Z_j \in i\text{-th bin}) = \frac{1}{300}\sum_{s=1}^{300}\prod_{x=1}^{l}\gamma(g_{i\times 300+s+x-1}, r_j^x)$. Similarly, we can optimize function (3.13) by the EM algorithm. As proved in [HH06], the likelihood function (3.13) is concave. The EM algorithm will converge to global optimal solution and it will be a good initial guess for the objective function in formula (3.6).

After obtaining a solution using a standard EM algorithm, we conduct our extended EM algorithm introduced in section 3.2.3. We summarize our method in Algorithm 2.

3.2.5 GC-bias Correction

One of the short comings of the HTS technologies is the existence of different biases in the sequencing process. Some biases are due to the environment while others are due to chemical reactions (DNA amplifications, GC content). Studies

Algorithm 2: The complete algorithm of CNVeM

Input: Read mapping information, allowing reads map to multiple locations.

Output: Copy number variations compared to reference genome.

Initialization: Choose an initial configuration of copy numbers $\mathcal{C}^{(0)}$.

STAGE ONE:

Optimize the function in (3.13) using a standard EM algorithm based on bins. We get an initial solution of copy numbers for each bin.

STAGE TWO:

2.1 Use the output from STAGE ONE as an initial guess.

2.2 For each read r_j with $j \in \{1, 2, \ldots, N\}$, consider all mapping positions, calculate the posterior probability of each position according to the joint probability in formula (3.2). Then map the read to multiple locations fractionally according to the posterior probability.

2.3 Calculate the total number of reads mapped to each position.

2.4 Update the copy numbers of all nucleotides using the dynamic programing in formula (3.12).

2.5 Repeat Steps 2.3-2.4 until it converges.

show that both Sanger and HTS sequencing have bias toward high GC regions. GC-bias can influence the number of reads generated from a position and thus the reads are no longer uniformly generated. There have been a number of papers [AKM09, AUS11, SKA10, YXM09] which deal with GC-bias in CNV calling. In this work, we adapted the idea mentioned in [AUS11, YXM09] to correct for GC-bias. In equation (3.10), d_i is the number of reads mapped to position i. We correct this bias by updating the definition of d_i to be $d_i^c = d_i \times \frac{\overline{DOC}_{global}}{\overline{DOC}_{gc}}$, where d_i^c is the corrected number of reads mapped to position i, d_i is the original number of reads mapped to position i, $\overline{DOC}_{global}$ is the average depth of coverage (DOC) over all positions, and $\overline{DOC}_{gc}$ is the average DOC over all positions where the reads have the same GC content as in the reads mapped to position i.

3.3 Results

3.3.1 Simulation Results

In order to assess our method, we carried out experiments on simulated datasets. We developed a simulation framework, in which a donor genome is obtained by altering the copy number of some regions in the reference genome.

Experiment on a simulated mouse chromosome

We first tested CNVeM on a simulated mouse genome. We obtained the masked reference chromosome 17 of Mus Musculus. After pruning all the 'N's, the length of the chromosome 17 reduced to 58Mb. This can be used as the "template sequence". We then duplicate segments of the sequence to generate a reference genome. The lengths of the duplicated segments are chosen from the range $[1000, 10000]$. We allow nucleotides to mutate with probability 1% in the duplication process. The copy numbers of these segments are then altered to generate

Table 3.1: The results on the simulated mouse chromosome 17 under different sequencing depth and mutation rates between duplicated segments. No. of predicated CNVs are the number of regions CNVeM reports as CNVs. False discovery rate is the ratio between number of false positives and number of predicted CNVs, while false negative rate is the ratio between number of false negatives and number of true CNVs. It is obvious that CNVeM reports false positive regions due to that fact that it calls more CNVs than implanted in the donor genome.

mutation rate between duplicated segments	Depth of Coverage	No. of Predicted CNVs	No. of Correct CNVs	False Discovery Rate	False Negative Rate
1%	30X	102	100	2.0%	0
	15X	102	100	2.0%	0
	5X	105	100	4.8%	0
0.5%	30X	102	100	2.0%	0
	15X	105	100	4.8%	0
	5X	109	100	8.3%	0
0.1%	30X	101	97	4.0%	3.0%
	15X	107	98	8.4%	2.0%
	5X	116	96	17.2%	4.0%

the donor genome. The copy numbers are chosen from the set $\{0, 1, 2, 3, 4, 5\}$. In each experiment, we simulated 100 copy number variations between the reference genome and donor genome. To generate a read, we randomly picked a position from the donor genome and copied 36 consecutive bases starting from this position. The copying process is repeated until we have the desired coverage. All reads are then mapped to the reference genome using mrsFast [HHA10], allowing reads to map with two mismatches. In addition to detecting the existence of copy number variants, CNVeM especially aims to distinguish which copy is duplicated

or deleted in the donor genome, while others have the same number of copy occurrences compared to the reference genome. Simulations are performed using various depth of coverage settings. A CNV is considered to be detected correctly when it overlaps with the true CNV region, meanwhile the predicted copy numbers should be the same as the true copy numbers. The results are shown in the first row of Table 3.1.

We also compared our reported CNVs to true CNVs by base pairs. The overlap is calculated by intersecting the coordinates of predicted CNVs with those of true CNVs. The results in the first row of Table 3.2 indicate high accuracy of CNVeM in predicting the break points.

Furthermore, we simulated the duplicated segments under different mutation rates to assess the power of our method in locating the copy variation origin. All results are summarized in Table 3.1 and Table 3.2. We see that both the mutation rate between duplicated segments and sequencing depth can affect the accuracy of our program. The smaller the mutation rate, the more similar the duplicated sequence, and the more difficult to distinguish which segment has copy number variation in the donor sequence. We have higher false discovery rate when the read depth is lower and the difference between duplicated copies is smaller, but we manage to recall almost all copy number variations.

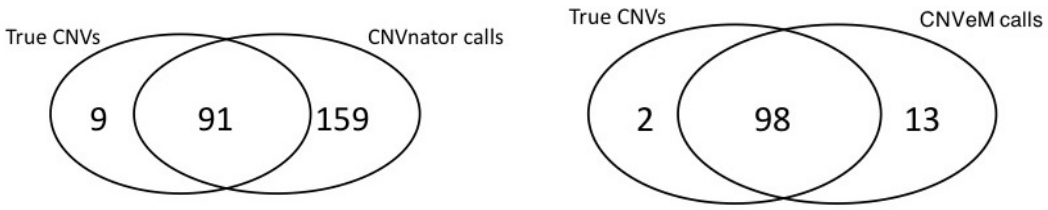
The key observation in comparing the two tables (Table 3.1 and Table 3.2) is that the false negative rate in predicting the correct quantitive copy number is always lower than the false negative rate in calling the breakpoints of CNVs, moreover the false discovery rate of quantitative value for CNV is always higher than the false discovery rate in breakpoint calling. This illustrates that CNVeM is robust in detecting the existence of CNVs and determining the break points of CNVs. To achieve high sensitivity in CNV calling, CNVeM inevitably reports false positive regions. However, most of these false positive regions are short and thus we have low false discovery rate in break points calling.

Table 3.2: Measuring the accuracy of CNV break points by base pairs under different sequencing depth and mutation rates between duplicated segments. False discovery rate is the ratio between length of false positive regions and total length of predicted CNVs, while false negative rate is the ratio between length of false negative regions and total length of true CNVs.

mutation rate between duplicated segments	Depth of Coverage	Length of Predicted CNVs(bp)	Length of over-lap(bp)	False Discovery Rate	False Negative Rate
1% (504000bp)	30X	506755	502183	0.9%	0.3%
	15X	506162	501291	1.0%	0.5%
	5X	507703	495074	1.8%	2.5%
0.5% (493000bp)	30X	492271	488114	0.9%	1.0%
	15X	500460	488387	2.4%	0.9%
	5X	501139	483830	3.5%	1.9%
0.1% (492000bp)	30X	469821	452120	3.8%	9.1%
	15X	465518	433495	6.9%	11.9%
	5X	462193	417340	9.7%	15.2%

Comparing CNVeM with CNVnator on GC-biased data

In this section we compare CNVeM with the CNVnator [AUS11], which is the state of art CNV detector. Using a similar framework, we generated a reference genome and donor genome from chromosome 17 of Mus Musculus. We set the mutation rate between duplicated segments to be 0.1%. Reads are then simulated from the donor genome, allowing GC-bias [AUS11, YXM09]. In order to make the comparison fair for CNVnator, we used Bowtie [LTP09] to do the mapping with option '-best -M 1'. With this option, Bowtie returns the best mapping for each read and in the case of ties it will randomly pick one mapping location for a read. This step is due to the fact CNVnator assumes there exists one mapping location for each read. However, for CNVeM, we use mrsFAST [HHA10] to return all possible mapping positions for each read. Figure 3.2 illustrates the intersection of CNVs found by CNVeM and CNVnator on the simulated dataset, where 100 CNVs are implanted to the donor genome. CNVeM finds 111 CNVs which includes 98 of the true CNVs. This indicates that CNVeM has 13 false positives and 2 false negatives. However, CNVnator finds 250 CNV regions among which 91 regions are true CNVs. CNVnator fails to find 9 regions which are true CNVs. Moreover, CNVnator reports 159 false positives. This results from the fact that CNVnator randomly places a read to one of its multiple mapping positions, and thus affects the read depth (RD) information, from which CNVnator determines the copy variation status. All the results indicate that CNVeM has lower false discovery rate and false negative rate compared to CNVnator. Another disadvantage of CNVnator is that it can only determine the CNV to be a copy gain or copy loss, instead of recalling the exact quantitive copy number as in CNVeM.



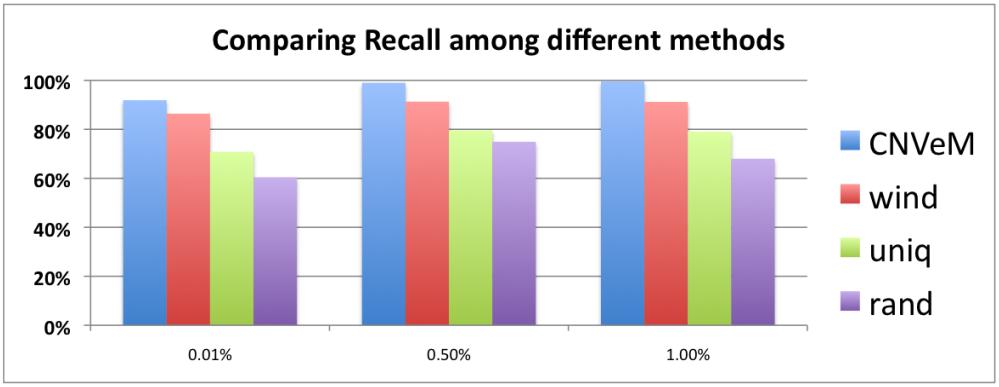
(a) CNVnator results (b) Our CNV results

Figure 3.2: Intersection of two CNV detection results with true CNVs. (a) We illustrate the Venn diagram of the CNVnator calling with the true CNV regions. (b) We illustrate the intersection between the CNVeM calls and the true CNV regions. This figure indicates that we have less false positives and false negatives than CNVnator.

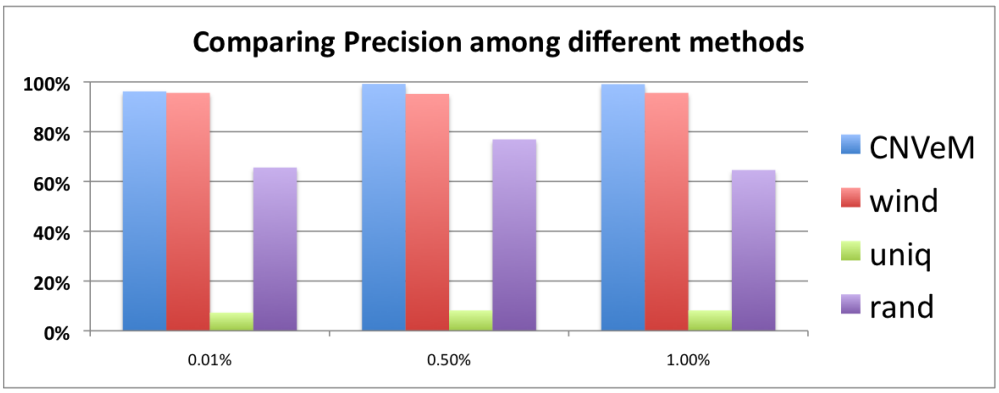
Comparison between different strategies dealing with read mapping uncertainty

When handling reads that can be mapped to multiple positions, existing methods either discard those reads, or randomly place the read to one of the multiple mapping positions. CNVeM considers all possible mapping positions, and a read can be placed to one of the positions with a probability. We compared the performance of these different strategies. Furthermore, we consider the popular strategy which divides the genome into bins. All nucleotides within one bin have the same copy number. We develop a method 'wind' using the same EM framework as in section 3.2.3 for the bin strategy.

We run these methods on the same simulated datasets. Following the same process mentioned above, we generated the reference genome and donor genome from chromosome 17 of Mus Musculus, with mutation rate between duplicated segments set to be 0.1%, 0.5%, and 1%, respectively. Reads are simulated at $30X$ coverage. The results plotted in Figure 3.3 illustrate that CNVeM has highest recall and precision at different mutation rates.



(a) Recall



(b) Precision

Figure 3.3: Comparison between several strategies dealing with read mapping uncertainty. The x-axis represents the mutation rate between duplicated segments. The shorthands *CNVeM*, *wind*, *uniq* and *rand* represent the results from *CNVeM*, the results from *wind* which divides the genome into bins, the results from only considering reads mapped to unique positions, and results from placing a read to one of multiple mapping positions randomly, respectively.

Time and memory usage

When dealing with HTS technology which generates tens of gigabytes of data per day, not only the accuracy of the method becomes important, but memory and time usage become important factors. The time and memory usage is estimated for the CNV calling process, and we assume the mapping is done in a separate step.

Our program takes 30 minutes to detect all the CNVs in the simulated dataset on masked chromosome 17 of the mouse genome, where we had 30X coverage (having around 50 million reads). All the experiments were run on a 64-bit AMD Opteron processor, furthermore, our program used 2Gb of memory at the peak of usage. In order to run CNVeM on the whole genome sequencing data, the memory usage increases linearly with the size of the genome.

3.3.2 Results on Real Data

We used the data published by Sanger Institute [SSS09], where chromosome 17 of mouse strain A/J is deeply sequenced using Illumina technology to test our method on real data. The data contains 112 million (56 million pair-end) reads and the length of each read is 36bp. This results in a 42X coverage. We aligned the reads to the masked chromosome 17 using mrsFast [HHA10], allowing up to 2 mismatches. Out of these 112 million reads, 39 million reads mapped uniquely to the genome. However, 4 million reads mapped to more than one position in the genome. We supply the mapping information of both uniquely and non-uniquely mapped reads to CNVeM, and managed to detect 44 copy gain regions and 355 copy loss regions. Among those 44 copy gain regions, 28 regions have been reported by Sudbery et al. [SSS09], and 15 regions out of these 355 copy loss regions have been reported by Sudbery et al. [SSS09] as copy loss regions. Sudbery et al. also reported 416 deletion regions. We checked the coordinates of those deletion regions and discovered that 415 of them have overlap with the rest of copy loss regions reported by CNVeM. Furthermore, we apply CNVnator on this real data where it manages to detect 42 copy gain regions and 264 copy loss regions. Comparing the CNVeM calls with those of CNVnator, we see 26 copy gain regions overlap, and 86 copy loss regions are found by both methods.

3.4 Discussion

CNV regions have been shown to be correlated with many diseases ranging from cancers to learning disabilities [CHR05, SLM07]. Two main strategies exist to improve CNV detection, either to improve the technology from which we gather data from individuals, or to design better algorithms. The shift from ArrayCGH to HTS is a good indicator of improvements in the data gathering process, as current studies suggest that the use of HTS results in higher power in detecting CNV breakpoints and quantifying the true copy number for each region.

It has been shown previously that we can use both the depth of coverage (DOC) and paired-end information to detect CNVs accurately [MFD10]. We have shown that correct usage of DOC improves the accuracy of CNV detection greatly. In this work we have presented a probabilistic model for detecting CNVs, based on an expectation-maximization (EM) method. Our method incorporates all available mapping information in the CNV prediction. It not only has higher accuracy in detecting the CNVs but also can detect which of the paralog regions in the genome is copied or deleted. All previous methods fail to distinguish paralog regions as they either discard all multiple mapping reads (reads mapped to multiple positions) or randomly assign a read to one of the mapping positions.

Another main contribution of this work is that we can predict the CNV breakpoints in base-pair resolution. Unlike previous methods which define CNV for each bin (segment of fixed or variable length), our objective function is defined for each base-pair. In other words we are predicting the CNV for each base-pair. This helps us to detect the breakpoint of each CNV with high accuracy.

Although we mention that using DOC can improve the accuracy of CNV detection, we do not deny the fact that paired-end mapping has valuable information. Our future work is to incorporate paired-end reads information into our probabilistic model.

CHAPTER 4

Copy Number Variation Detection from Tumor Samples Contaminated by Stromal Cells

4.1 Background

Many recent studies have shown a correlation between CNVs and cancers [CHR05, IFR04, TSB05]. Historically, two key techniques have been used to detect CNVs in tumor genomes: array comparative genome hybridization (ArrayCGH) and loss of heterozygosity (LOH) [CLC08, Car07, RIF06, Zha10]. These techniques, although powerful to detect the presence of CNVs, are unable to identify the boundaries of CNVs with high resolution.

The development of high-throughput sequencing (HTS) technologies provides great opportunities to detect CNV regions with high resolution in tumor genomes. With HTS technologies, whole genome shotgun sequencing becomes possible. Millions of reads are obtained from fragments of the DNA molecules. The reads are mapped to the reference genome and the mapping information is utilized to call CNVs.

Recent studies have proposed methods to detect CNVs using short reads generated from HTS technologies. One approach is to split the genome into small windows and use the number of reads mapped to each specific window (*read depth*) as a proxy for the copy number of that window [AKM09, SKA10, SMA10, CGJ09, YXM09]. However, the resolution of this approach is limited by the size of the windows, which is typically at least one kilobase. Another approach is to use

"paired-end" reads, where "paired-end" refers to the two ends of the same segment of a DNA molecule, to detect CNVs [MFD10, HFE10, HAE09]. One limitation of this strategy is that it can not detect CNVs in repeat-rich regions, where a short segment of the DNA sequence appears repetitively. This may reduce the applicability of this method given the existence of many repeat-rich regions in the human genomes.

In Chapter 3, we proposed a statistical method, CNVeM, to detect CNVs in the donor genome. However, CNVeM cannot be applied to detect CNVs in tumor genome directly. One challenge in detecting tumor CNVs comes from the specimen collecting process, prior to applying either ArrayCGH, LOH or HTS technologies. In a typical experiment, tumor tissue samples were cut from random sites of the tumor at biopsy. However, tumor cells are surrounded by stromal cells, which are the normal connective tissue cells in the organs. Tumor samples are easily contaminated by the stromal cells in the specimen collection process. This heterogeneity in tumor samples contributes to the complexity of CNV detection. Liu et al. proposed an HMM model to infer CNVs using SNP arrays from tumor samples mixed with stromal cells [LLS10]. However, their method suffers from low resolution from the inherent limitation of array techniques.

In this study, we extended CNVeM and proposed a new probabilistic model, CNVmix, that estimates the copy numbers for each nucleotide based on the read mapping information from tumor samples. CNVmix is able to incorporate the contaminating genomes. We proposed a method to estimate the proportion of contaminating genomes in the tumor samples. Most mammals, including human, are diploid. One diploid cell contains two sets of genomes, each inherited from one parent. CNVs in diploid tumor cells appear in many forms, such as hemizygous deletion where a region of one genome is deleted, homozygous deletion where a region of both genomes is deleted and amplification where a region of one or both genomes is duplicated. We utilize the hemizygous deletion regions to estimate the

proportion of contaminating cells in the tumor samples. We identify hemizygous deletion regions by detecting regions in which the read depth is lower than expected and the *alleles* (types of nucleotides at a SNP position) of heterozygous SNPs are imbalanced. In hemizygous regions, tumor genomes have copy number 1, while stromal genomes have normal copy number 2. Utilizing the ratio between two alleles at heterozygous SNPs within hemizygous deletion regions, we estimate the proportion of contaminating genomes as in section 4.2.2. With the estimated contamination rate, we develop the generative model of observing the read set from the contaminated tumor samples. The CNVs in the tumor genomes are estimated by optimizing the parameters in the generative model.

Another important challenge for detecting CNVs in tumor genome lies in the uncertainty of read mapping. Similar with CNVeM, the new model CNVmix also utilizes the uncertainty of read mapping to detect tumor CNVs. We can detect the CNV boundaries and copy numbers of CNVs precisely and are able to predict small CNVs.

We apply our method to simulated datasets and achieve higher accuracy compared to existing methods. To our knowledge, this is the first attempt to predict tumor CNVs using HTS outputs from contaminated tumor samples.

4.2 Methods

4.2.1 The generative model

We use short read information from the HTS technologies to detect copy number variants in diploid tumor genome. We use same notation as the generative model of CNVeM in Chapter 3. One difference is that we aim to detect CNVs in tumor genome, which is diploid, so we assign the copy number of each nucleotide in the diploid reference genome to be 2. For most nucleotides, the copy numbers are the

same in the tumor genome and in the reference genome. So one can assume that the length of tumor genome is the same as the length of the reference genome, i.e. $\sum_{i=1}^{K} C_i = 2K$. We define vector $(\frac{C_1}{2K}, \frac{C_2}{2K}, \ldots, \frac{C_K}{2K})$ to be the normalized copy number vector of the tumor genome.

Using HTS technology, millions of short reads are generated from the tumor samples. As the samples are contaminated by stromal cells, which have same copy numbers with the reference genome, reads can originate from either the tumor genome or the stromal genome. Denote the proportion of stromal cells in tumor samples to be ρ ($0 \leq \rho \leq 1$). The probability of a read originating from the stromal genome is ρ.

In our model, each read r_j originates from one position in either the stromal genome or the tumor genome. Let $\mathcal{H} = \{H_1, H_2, \ldots, H_N\}$ be the source of each read, where $H_j \in \{0, 1\}$ represents whether read r_j originates from the stromal genome ($H_j = 0$) or the tumor genome ($H_j = 1$). $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_N\}$ is the true origin of each read, where $Z_j \in \{1, 2, \ldots, K\}$, and we then define the following likelihood model of all reads given copy number $\mathcal{C}$, contamination rate ρ and reference genome sequence $\mathcal{G}$

$$P(\mathcal{R}|\rho, \mathcal{C}, \mathcal{G}) = \prod_{j=1}^{N} P(r_j|\rho, \mathcal{C}, \mathcal{G}) = \prod_{j=1}^{N} \sum_{h=0}^{1} \sum_{i=1}^{K} P(r_j, H_j = h, Z_j = i|\rho, \mathcal{C}, \mathcal{G}). \quad (4.1)$$

The interpretation of the above probability definition $P(r_j, Z_j = i, H_j = h|\rho, \mathcal{C}, \mathcal{G})$ is straightforward: the probability of j-th read generated from i-th position in the stromal genome (tumor genome), given the contamination rate, copy numbers and reference genome sequence. We can further expand this probability as follows:

$$P(r_j, Z_j = i, H_j = h|\rho, \mathcal{C}, \mathcal{G}) = P(Z_j = i, H_j = h|\rho, \mathcal{C}) P(r_j|Z_j = i, \mathcal{G}). \quad (4.2)$$

The equality follows from the fact that the read position Z and source H are independent of the reference genome sequence $\mathcal{G}$ and the sequence of read r is

independent of copy number $\mathcal{C}$ and contamination proportion ρ. We define the first term to be the probability for read r_j originating from position i of the stromal genome (tumor genome) as follows:

$$P(Z_j = i, H_j = h|\rho, \mathcal{C}) = \begin{cases} 1/K \times \rho & \text{if } h = 0 \\ C_i/2K \times (1-\rho) & \text{if } h = 1. \end{cases} \tag{4.3}$$

As in Equation (3.3), we denote the probability of observing the sequence of read r_j given that the true origin of read r_j is position i to be $P(r_j|Z_j = i, \mathcal{G})$. The prior probability of the tumor genome is also defined as $P(\mathcal{C}) = P(C_1, C_2, \ldots, C_K) = P(C_1)\prod_{i=2}^{K} P(C_i|C_{i-1})$.

Using Bayes rule, we can get the posterior probability of $\mathcal{C}$ given the read set $\mathcal{R}$, contamination rate ρ and the reference genome sequence $\mathcal{G}$:

$$\begin{aligned} P(\mathcal{C}|\rho, \mathcal{R}, \mathcal{G}) &\propto P(\mathcal{R}|\rho, \mathcal{C}, \mathcal{G})P(\mathcal{C}) \\ &\propto \left(\prod_{j=1}^{N}\sum_{i=1}^{K}[\frac{2\rho + (1-\rho)C_i}{2K}]P(r_j|Z_j = i)\right) \times \left(P(C_1)\prod_{i=2}^{K} P(C_i|C_{i-1})\right). \end{aligned} \tag{4.4}$$

4.2.2 Estimation of contamination rate ρ

Due to the existence of stromal cells in the tumor samples, the detection of CNVs becomes more difficult. Accurate estimation of the proportion of stromal cells in the mixed sample plays an important role in detecting CNVs in tumor genomes. We proposed a method to estimate the contamination rate from the read mapping information. We utilize the hemizygous deletion regions to estimate the proportion of contaminating cells in the tumor samples. In hemizygous deletion regions, tumor genomes have copy number 1, while stromal genomes have normal copy number 2. Hemizygous deletion regions are predicted to be copy loss regions using the method in section 4.2.3 no matter contaminating cells exist or not.

After identifying copy loss regions using the method in section 4.2.3, allele

frequency information is used to extract hemizygous deletions from the copy loss regions. Each heterozygous SNP has two alleles, denoted as 'A' and 'B'. In hemizygous deletion regions, the alleles of heterozygous SNPs are imbalanced as one of the alleles is deleted. This information can be used as a signal to indicate whether a copy loss region is hemizygous deletion region or not. After identifying hemizygous deletion regions, we denote the B allele frequency (BAF) of a heterozygous SNP in the hemizygous deletion region as b and apply the following strategy similar to [LLS10] to estimate the contamination rate $\hat{\rho}$.

$$\hat{\rho} = \frac{B_T - bn_T}{B_T - bn_T - (B_N - bn_N)} \tag{4.5}$$

where B_T,n_T is the B allele copy number (0 or 1) and total copy number (1) of the heterozygous SNP in hemizygous deletion region in the tumor genome; $B_N = 1$, $n_N = 2$ is the B allele copy number and total copy number of the heterozygous SNP in stromal genome. The estimates from all heterozygous SNPs in hemizygous deletion regions are averaged to approximate the true contamination rate. In the uncommon case that no hemizygous deletion region is identified in the sample, higher copy numbers can also be used with the same formula to estimate the contamination rate ρ.

4.2.3 Optimization

Maximizing the posterior of copy number $\mathcal{C}$ in Equation (4.4) is equal to maximizing the following log probability with respect to $\mathcal{C}$:

$$\sum_{j=1}^{N}\left(\log\sum_{i=1}^{K}[\frac{2\rho+(1-\rho)C_i}{2K}]P(r_j|Z_j=i)\right)+\log\left(P(C_1)\prod_{i=2}^{K}P(C_i|C_{i-1})\right). \tag{4.6}$$

Similar as in Chapter 3, we incorporate a penalty function coefficient δ in order

to eliminate the constraint $\sum_{i=1}^{K} C_i = 2K$. Our objective function now becomes

$$\sum_{j=1}^{N}\left(\log\sum_{i=1}^{K}[\frac{2\rho+(1-\rho)C_i}{2K}]P(r_j|Z_j=i)\right)+\log\left(P(C_1)\prod_{i=2}^{K}P(C_i|C_{i-1})\right)$$
$$+\delta\left(2K-\sum_{i=1}^{K}C_i\right) \tag{4.7}$$

We optimize the objective function (4.7) through an expectation-maximization (EM) algorithm.

Expectation-step:

$$\begin{aligned} Q(\mathcal{C}|\mathcal{C}^{(t)}) &= \sum_{j=1}^{N}\sum_{i=1}^{K}\log\left([\frac{2\rho+(1-\rho)C_i}{2K}]\right)^{P(Z_j=i|r_j)}+\sum_{j=1}^{N}\sum_{i=1}^{K}\log P(r_j|Z_j=i)^{P(Z_j=i|r_j)} \\ &+ \log P(\mathcal{C})+\delta\left(2K-\sum_{i=1}^{K}C_i\right) \end{aligned} \tag{4.8}$$

Maximization-step:

$$\begin{aligned} \mathcal{C}^{(t+1)} &= \arg\max_{\mathcal{C}}\log\left[P(C_1)\left([\frac{2\rho+(1-\rho)C_i}{2K}]\right)^{d_1}\times e^{-\delta C_1}\right. \\ &\left.\times\prod_{i=2}^{K}\left(P(C_i|C_{i-1})\left([\frac{2\rho+(1-\rho)C_i}{2K}]\right)^{d_i}\times e^{-\delta C_i}\right)\right] \end{aligned}$$

where $d_i = \sum_{j=1}^{N} P(Z_j = i|r_j)$.

We solve the M-step using dynamic programming. Denote the objective function in the M-step to be

$$f = \log\left[P(\mathcal{C})\times\prod_{i=1}^{K}\left(\left([\frac{2\rho+(1-\rho)C_i}{2K}]\right)^{d_i}\times e^{-\delta C_i}\right)\right]. \tag{4.9}$$

Then we define $f(k, x)$ to be the maximum function value for first k positions when the copy number of kth position is $C_k = x$. Now we design the dynamic

programming solution indicated as follows:

$$f(k,x) = \begin{cases} \log[P(C_k = x) \times ([\frac{2\rho+(1-\rho)C_k}{2K}])^{d_k} \times e^{-\delta C_k}] & \text{if } k = 1 \\ \max_{C_{k-1}}\{f(k-1, C_{k-1}) + \log[P(C_k|C_{k-1})]\} & \\ \quad + \log[([\frac{2\rho+(1-\rho)C_k}{2K}])^{d_k} \times e^{-\delta C_k}] & \text{otherwise} \end{cases}$$

Similar to the proof in Chapter 3, it can be proved that the above dynamic programming solution returns the global optimal solution for objective function in Equation (4.9). The maximum value of the objective function in the M-step is then $\max_x f(K, x)$. Using a backtrack process, we find the vector $C = (C_1, C_2, \ldots, C_K)$ that maximizes the function f in the M-step. By iteratively running the E-step and M-step, we achieve local optimal solution.

4.3 Results

In order to assess our method, we carried out experiments on simulation datasets. We developed a simulation framework, in which the tumor genome is obtained by altering the copy number of some regions from the reference genome.

4.3.1 Experiment on a simulated human chromosome 17

We tested our method on a simulated human genome. We obtained the human reference chromosome 17 from Feb. 2009 assembly of human genome (hg19, GRCh37 Genome Reference Consortium Human Reference 37). After pruning all the 'N's, the length of the chromosome 17 reduced to 40Mb. We utilize the similar framework as in section 3.3.1 to generate the reference genome and tumor genome. The only difference is that the reads are generated from both reference genome and tumor genome. This is to simulate the fact that tumor samples are contaminated by the stromal cells. Simulations are performed on different contamination rate to assess the power of CNVmix in detecting tumor CNVs from contaminated sam-

ples. A CNV is considered to be detected correctly when it overlaps with the true CNV region, meanwhile the predicted copy numbers should be the same with the true copy numbers. The results are shown in Table 4.1. We also compared our

Table 4.1: The results on the simulated human chromosome 17 under different proportion of contamination cells. No. of predicated CNVs are the number of regions CNVmix reports as CNVs. False discovery rate is the ratio between number of false positives and number of predicted CNVs, while false negative rate is the ratio between number of false negatives and number of true CNVs. It is obvious that CNVmix reports false positive regions due to that fact that it calls more CNVs than implanted in the tumor genome.

True Contamination Rate	Estimated Contamination Rate	No. of Predicted CNVs	No. of Correct CNVs	False Discovery Rate	False Negative Rate
0%	0	100	100	0	0
20%	21.2%	101	100	1.0%	0
40%	38.8%	103	100	2.9%	0
50%	52.7%	116	100	13.8%	0
60%	58.5%	123	99	19.5%	1.0%
80%	84.2%	249	93	62.6%	7.0%

reported CNVs to true CNVs by base pairs. The overlap is calculated by intersecting the coordinates of predicted CNVs with those of true CNVs. The results in Table 4.2 indicate high accuracy of CNVmix in predicting the break points.

From Table 4.1 and Table 4.2, we observe that both false discovery rate and false negative rate increase as the contamination rate increases. Nonetheless, our method has sufficient power to detect CNVs from mixed tumor samples, even tumor samples that are contaminated by as much as 60% normal cells. When the contamination rate is higher than 50%, there is a dramatic rise in terms of

false discovery rate, while the increase in false negative rate is more smooth. This phenomenon indicates that in order to achieve high sensitivity, our method inevitably reports false positive regions. However, most of these false positive regions are short, and thus we have low false discovery rate in break points calling. The false negative rate in predicting the correct quantitive copy number is always lower than the false negative rate in calling the breakpoints of CNVs. Moreover the false discovery rate of quantitative value for CNVs is always higher than the false discovery rate in breakpoint calling. This illustrates that CNVmix is robust in detecting the existence of copy variation and determining the break points of CNVs.

Table 4.2: Measuring the accuracy of CNV break points by base pairs under different proportion of contaminating cells. The total length of true CNVs is 491000bp. False discovery rate is the ratio between length of false positive regions and total length of predicted CNVs, while false negative rate is the ratio between length of false negative regions and total length of true CNVs.

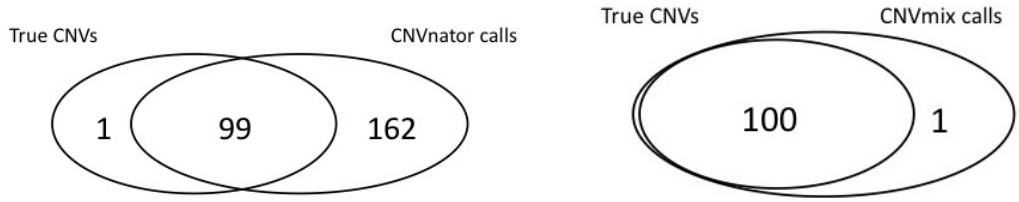
True Contamination Rate	Estimated Contamination Rate	Length of Predicted CNVs(bp)	Length of overlap(bp)	False Discovery Rate	False Negative Rate
0%	0	484100	478084	2.6%	1.2%
20%	21.2%	491299	482355	1.8%	1.8%
40%	38.8%	493397	482293	2.3%	1.8%
50%	52.7%	493397	482293	6.6%	1.5%
60%	58.5%	525477	477592	9.1%	2.7%
80%	84.2%	872751	445754	48.3%	9.2%

4.3.2 Comparison of our method with CNVnator

In this section we compare our method with the CNVnator [AUS11], which is the state-of-the-art CNV detector. Using a similar simulation framework, we generated the genomes in stromal cells and tumor cells from chromosome 17 of human. We set the contamination rate to be 20% and reads are then simulated at $30X$ coverage. We use the same parameter configuration as in section 3.3.1 to compare CNVmix and CNVnator. Figure 4.3.2 illustrates the intersection of CNVs found by CNVmix and CNVnator on the simulated dataset, where 100 CNVs are implanted in the tumor genome. CNVmix finds 101 CNVs which include all of the true CNVs. This indicates that CNVmix has 1 false positive and no false negatives. However, CNVnator finds 261 CNV regions among which 99 regions are true CNVs. CNVnator fails to identify one CNV region. Moreover, CNVnator reports 162 false positives. This results from fact that CNVnator mistakes contaminating cells for tumor cells. Meanwhile, CNVnator randomly places a read to one of its multiple mapping positions, and thus affects the read depth (RD) information, from which CNVnator determines the copy variation status. All the results indicate that CNVmix has lower false discovery rate and false negative rate than CNVnator.

4.4 Discussions

In this work we present a probabilistic model for detecting CNVs from HTS outputs, based on an Exception-Maximization (EM) method. Our method incorporates all read mapping information. It has higher accuracy in detecting the CNVs compared to previous methods, as they either discard multiple mapping reads or randomly place a multiple mapping read to one of the mapping positions. Considering the fact that tumor samples are easily contaminated by stromal cells, we incorporate the contamination rate in our model and proposed a method to



(a) CNVnator results

(b) Our CNV results

Figure 4.1: Intersection of two CNV detection results with true CNVs. (a) We illustrate the venn diagram of the CNVnator calling with the true CNV regions. (b) We illustrate the intersection between the CNVmix calls and the true CNV regions. This figure indicates that we have less false positives and false negatives than CNVnator.

estimate the contamination rate. The simulation results indicate that our method estimates the contamination rate accurately.

This model can identify hemizygous deletions, homozygous deletions and amplifications accurately according to the simulation results. However, it cannot identify copy neutral LOH region, where one genome is deleted but the other genome is duplicated. Our model depends on the read depth (RD) information to detect CNVs while the copy number of copy neutral LOH is still 2 and RD signal does not reflect the variation. A signal of BAF band centered at 0 or 1 indicates the presence copy-neural LOH. One future direction of this model is to incorporate BAF information for detection of copy neural LOH region.

CHAPTER 5

Gene-Gene Interactions Detection Using A Two-stage Model

5.1 Background

Genome-wide association studies (GWAS) attempt to discover genetic variation associated with disease traits. To perform GWAS, studies collect genetic variation of individuals and their disease status or disease related traits. GWAS studies typically collect single nucleotide polymorphisms (SNPs) because technologies allow for very cost-efficient collection of SNPs. Since SNPs are so prevalent in the genome, they are likely to be correlated with other genetic variations. Current GWAS studies collect about a million SNPs in thousands of individuals. The standard approach for identifying associations between SNPs and traits is that for each SNP, we compare the average trait value of individuals who have one allele of a SNP and that of individuals who have the other allele of the SNP. If the difference between the two average trait values is above a certain threshold, we declare that the SNP is significantly associated with the trait. We refer to computing the difference in the average trait values for each SNP as the "single marker test", and it has successfully identified many individual SNPs associated with several complex diseases [CSS93, BKK94, AHK00, SVL07, Con07].

Current studies on certain complex diseases have also suggested that some SNPs influence diseases through interactions [WAP00, BSW05, YIS04]. In an extreme scenario, two SNPs may not have any effect on a disease independently,

but they may affect the disease when both are present. To detect an interaction of SNPs, one needs to consider the association between a trait and a pair of SNPs. One approach to find such associations is to divide individuals into two groups: one group of individuals who have a certain combination of alleles for a pair of SNPs and the other group of individuals who have different combinations of alleles for the pair of SNPs. We then compute the difference in the average trait value between the two groups to determine whether the pair of SNPs is significantly associated with the trait. Finding an association between a trait and a pair of SNPs is called the "pairwise association test", and recently, several different methods have been proposed [EMM06, ZHZ10, YHW09, MCG06, LHC04].

One major challenge in discovering pairs of SNPs associated with a trait is that it requires enormous computation. One needs to compute associations between a trait and $4\times\binom{M}{2}$ pairs of SNPs where M is the number of SNPs available for testing. When M is close to one million as in current GWAS, an exhaustive pairwise search that considers all pairs of SNPs considers 2000 billion pairs of SNPs, which is a computationally challenging task. As the number of SNPs in GWAS keeps increasing with the improvement of technologies to collect SNPs, the exhaustive search becomes even more computationally infeasible.

In this work, we present a Threshold-based Efficient Pairwise Association Approach (TEPAA) for detecting associations between traits and pairs of SNPs using a two-stage model. In the first stage, our method performs the single marker test on all individual SNPs and selects a subset of SNPs that exceed a certain significance threshold (called "the first stage threshold") for further consideration. In the second stage, individual SNPs that are selected in the first stage are paired with each other, and we perform the pairwise association test on those pairs. In this method, there exists a trade-off between the probability of the method detecting a pair of SNPs associated with a trait (called "statistical power of the method") and the computational burden (or cost). Intuitively, statistical power

increases as we include more SNPs in the second stage, which means higher cost. The first stage threshold determines this trade-off, and we derive the analytical power of our method which allows us to determine the threshold and to control this trade-off. The key insight of our approach is that we derive the joint distribution between the association statistics of single SNPs and the association statistics of pairs of SNPs. This joint distribution allows us to provide guarantees that the statistical power of our approach will closely approximate the brute force approach. We can accurately compute the analytical power of our two stage model at any first stage threshold and compare it to the power of the brute force approach. Hence, we are able to choose as few SNPs as possible in the first stage while achieving almost the same power as the brute force approach.

While recently developed methods such as TEAM [ZHZ10, ZPX09] significantly reduce the computational burden of searching for pairs of associated SNPs, to our knowledge very few methods are feasible to apply to full size human GWAS datasets. The SIXPAC method developed by Pe'er and Prabhu utilizes a novel randomization technique that requires $10\times$ to $100\times$ fewer tests than a brute-force approach to find long-range interactions using standard two-locus test [PP12]. However, their method only handles case-control data and can not apply to quantitative traits. Wan et al. developed an approach BOOST, which designed a Boolean representation of data and used a screening stage to filter out most non-significant SNP interactions [XCQ10]. However, their method can not apply to quantitative traits either.

The only existing method that is feasible on a full size human GWAS dataset to detect SNP pairs associated with quantitative traits is FastEpistasis [SXB10]. FastEpistasis is a brute-force approach which conducts pairwise associations for all pairs of SNPs, or SNP pairs specified by users. The advantage of FastEpistasis is that their method is parallelled and utilizes high-performance computer architectures with multiple cores. Our method utilizes a two-stage strategy and

greatly reduced the number of pairwise association tests with little power loss.

We note that in this work, we are only considering pairs of SNPs which are far apart from each other. There is another class of methods which consider multiple SNPs close to each other [WKE10, WLC11, LLK13]. These problems are completely different and characterized by very different challenges. For example, the computational burden which is the focus of our method is different because the number of pairs of SNPs near each other is significantly smaller than the total number of pairs of SNPs. In addition, neighboring SNPs are typically correlated with each other, referred to as in linkage disequilibrium (LD). Pairs of SNPs far from each other are typically independent or unlinked which is an observation that we leverage in our approach.

5.2 Results

5.2.1 Overview of the Two-stage Model TEPAA

We present a two-stage model, TEPAA, for detecting associations between traits and pairs of SNPs. In this first stage, the association statistics for all SNPs are computed. Any SNPs which have a statistic higher than a pre-determined threshold then advance to the second stage in which all pairs of these SNPs are evaluated. The first stage threshold is important in determining power and cost of our method because it controls the number of SNPs to be selected in the first stage. For a truly associated pair of SNPs to be identified using our approach, both SNPs must advance to the second round and thus must have association statistics higher than the first stage threshold. Clearly, the more stringent the threshold, the smaller the number of SNPs in the second stage and the smaller number of pairs of SNPs which must be evaluated speeds up this method. On the other hand, more stringent thresholds increase the chance that at least one of the pair of truly associated SNPs will not be more significant than the first

stage threshold and the pair will not be identified by the method. Hence, there is a trade-off between power and cost, which is determined by the first stage threshold.

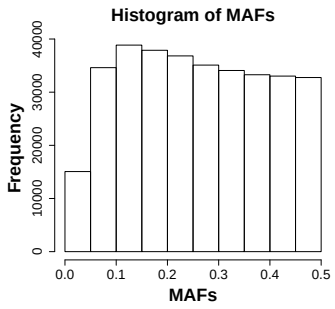
Our method chooses the first stage thresholds such that the two-stage model loses only a small amount of power but increases computational efficiency dramatically compared to the exhaustive search. To find such thresholds, we first derive the analytical power and cost of both the brute force approach and the two-stage model. This analysis allows us to choose the threshold that yields the desired power and cost, and hence it allows us to control the trade-off between the two. To derive the analytical power of our two stage model, we use the framework of Multivariate Normal Distribution(MVN) to model the association statistics [HKE09, KE13, KLE11]. We use a MVN to approximate the joint distribution between the association statistic of single SNP and the association statistic of pairs of SNPs. The non-centrality parameters (NCPs) of statistics are considered to be the mean vector in the MVN and correlations among statistics as a covariance matrix in the MVN. The NCPs and correlations can be calculated from the data and thus we obtained all the parameters of the MVN. The details of the analysis are discussed in Section 5.3.4.

From our analysis, we observe that the thresholds which control the power loss of the two stage approach depend on the minor allele frequency (MAF) of the SNPs. In particular, more common SNPs can be filtered out with less significant thresholds than rare SNPs. In order to efficiently implement TEPAA using MAF dependent thresholds for each pair, we group the SNPs into bins based on their MAFs to apply the correct thresholds to each possible pair. After disregarding rare variants with $MAF < 0.05$, we categorize all common SNPs into 9 bins according to their MAF, with step size 0.05. Each pair of SNPs would have two thresholds, one for each SNP in the first stage. In total, we have $\binom{9}{2}+9$ categories of SNP pairs. We pre-compute the first stage thresholds for each combination of

two MAFs in order to achieve 1% power loss, while achieving high cost savings. We sort the SNPs within each bin by their association statistics and use binary search to rapidly obtain the set of SNPs above a single threshold to efficiently implement the first stage of our method.

5.2.2 Application of TEPAA to the NFBC Data

We applied TEPAA to the Northern Finland Birth Cohort (NFBC) data to demonstrate the utility of our two stage model and the cost saving on a real data. The Northern Finland Birth Cohort Data contains $5,326$ individuals, and $331,476$ SNPs are genotyped. The histogram of all SNPs' MAFs is shown in Fig. 5.1(a). As described in detail in Section 5.3.6, we categorize all common SNPs into 9 bins according to their MAFs. The number of SNP pairs in each category is shown in Fig. 5.1(b). The first stage thresholds of TEPAA are pre-computed for each category in order to have the power loss at 1% using the methods described in Section 5.3.6. The cost saving for each category is summarized in Table 5.1. Based on Fig. 5.1(b) and Table 5.1, the estimated overall cost saving is 63.2 times, which is the ratio between total number of pairwise association tests of brute force approach and that of TEPAA.



(a) The distribution of all SNPs' MAFs

MAF of SNP A \ MAF of SNP B	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.05	1.13	5.2	5.86	5.70	5.54	5.30	5.13	5.00	4.97	4.94
0.1	-	5.97	13.45	13.07	12.72	12.16	11.78	11.48	11.40	11.34
0.15	-	-	7.57	14.71	14.32	13.69	13.26	12.93	12.84	12.77
0.2	-	-	-	7.15	13.91	13.31	12.89	12.56	12.48	12.41
0.25	-	-	-	-	6.77	12.95	12.54	12.22	12.14	12.07
0.3	-	-	-	-	-	6.19	11.99	11.69	11.61	11.55
0.35	-	-	-	-	-	-	5.81	11.32	11.25	11.18
0.4	-	-	-	-	-	-	-	5.52	10.96	10.90
0.45	-	-	-	-	-	-	-	-	5.44	10.83
0.5	-	-	-	-	-	-	-	-	-	5.38

(b) The number of SNP pairs in each category. Numbers are shown in factor of 100 millions.

Figure 5.1: The Distribution of all SNPs' MAFs and number of SNP pairs in each category.

For all SNPs in each bin, we calculate the association statistics and sort the SNPs in descending order of their statistics. We perform our analysis using the dominant model which is standard for analysis of epistatic interactions. We note that the basic approach of TEPAA can be extended to other models such as recessive or additive as well.

We compare the performance of the brute force approach and TEPAA to detect the SNP pairs associated with the phenotype "CRP" (C-reactive protein) on a machine with 2.3 GHz AMD Opteron Processor. Since it is impractical to run the brute force on the whole chromosome, the CPU time of the brute force approach is estimated from one single chromosome by scaling, which is estimated to be $1,542$ hours for phenotype "CRP". The CPU time of TEPAA is 24.5 hours for the same phenotype. We achieved 62.9 times of cost saving, which verifies our analysis of the cost savings of TEPAA when achieving 1% of power loss. However, both brute-force approach and two-stage model report no significant SNP interactions under the significance threshold 10^{-12}. This is understandable since this data set contains only $5,326$ individuals. In the next section, we show that the brute force approach and TEPAA have similar power when there exists significant SNP interactions.

5.2.3 TEPAA Controls Power Loss in Simulated Data

To demonstrate that TEPAA has only 1% power loss using the pre-computed first stage thresholds, we perform simulations where we implant a significant SNP-SNP interaction to the NFBC data and then detect the SNP pair using TEPAA.

We created phenotype data using the phenotype "CRP" (C-reactive protein) in the NFBC data as a starting point. To simulate the significant SNP pairs, we randomly sample the MAF of each SNP from $[0.05, 0.5)$. The alleles of each individuals at these two simulated SNPs are then sampled according to the MAF. The

phenotypes of the individuals with causal alleles at the SNP pairs are increased by a selected effect size so that the pairs has 50% power in the brute-force approach. Then we apply both the brute-force approach and the two-stage approach to the simulated dataset. The first stage significance thresholds in the two-stage approach are selected in order to obtain 1% power loss.

We generated $10,000$ simulated SNP pairs and applied both approaches. The power for each approach is calculated as the proportion of experiments that the approach detected the implanted SNP pairs among all 10000 experiments. The power of brute-force approach is 51% while the power of TEPPA is 50.8%. The practical power loss is 0.4%. We note that the power loss is lower than we expected because the thresholds are chosen for MAF frequency bins to be conservative and valid for all members of that bin.

5.3 Methods

5.3.1 Association Test between One SNP and Traits

We first illustrate the method to detect association between traits and one SNP. A traditional approach to identify the association is that for each SNP, we compare the average trait value of individuals who carry the causal allele at the SNP and that of the individuals who do not have the causal allele at the SNP of interest. If the difference between those two values is above a certain threshold, we declare that the investigated SNP has a significant association with the trait. This approach is referred to as "single marker test" and has been successful in many association studies. We analyze the power of the "single marker test" as follows.

Assume we are investigating SNP A, with minor allele frequency (MAF) to be p_A and the causal allele is the minor allele (for the case where the causal allele is

the major allele, the analysis is similar). Let N be the number of individuals and y_i be the trait value of individual i. Then the number of individuals with the minor allele at SNP A can be denoted as $N_A = N \cdot p_A$ and the number of individuals without the minor allele at SNP A can be denoted as $N_{\neg A} = N \cdot p_{\neg A} = N \cdot (1 - p_A)$ We use x_i^A to denote the allele of individual i at SNP A. y_i is any real number and $x_i^A \in \{0, 1\}$. We set $x_i^A = 1$ when the allele of individual i at SNP A is the minor allele and $x_i^A = 0$ otherwise.

We assume that a trait value of individual i follows the normal distribution with a certain mean μ and a variance σ^2. If the minor allele affects the trait, the mean trait value (μ) of individuals with the minor allele will increase by a certain value β_A (effect size). Now, we can obtain the distribution of y_i as

$$y_i \sim N(\mu + x_i^A \beta_A, \sigma^2) \tag{5.1}$$

Let $\bar{Y}_A$ be the average trait value of individuals who have the causal allele at SNP A and $\bar{Y}_{\neg A}$ be the average trait value of individuals who do not carry the causal allele at SNP A. Then we can derive the distributions of $\bar{Y}_A$ and $\bar{Y}_{\neg A}$ as follows:

$$\bar{Y}_A = \frac{\sum_{i:x_i^A=1} y_i}{N_A} \sim N(\mu + \beta_A, \frac{\sigma^2}{N \cdot p_A}), \bar{Y}_{\neg A} = \frac{\sum_{i:x_i^A=0} y_i}{N_{\neg A}} \sim N(\mu, \frac{\sigma^2}{N \cdot p_{\neg A}}) \tag{5.2}$$

We normalize the difference between $\bar{Y}_A$ and $\bar{Y}_{\neg A}$ to obtain the following statistic S_A, which is normally distributed with mean $\lambda_A \sqrt{N}$ (the non-centrality parameter) and unit variance.

$$S_A = \frac{\bar{Y}_A - \bar{Y}_{\neg A}}{\sqrt{\frac{\sigma^2}{N \cdot p_A \cdot (1-p_A)}}} \sim N(\lambda_A \sqrt{N}, 1), \ \textit{where } \lambda_A = \frac{\beta_A \sqrt{p_A(1-p_A)}}{\sigma} \tag{5.3}$$

Given the significance level α and the observed value of the test statistic S_A, the SNP is deemed as significant, or statistically associated with the trait, if $|S_A| \geq \Phi^{-1}(1 - \alpha/2)$, where Φ^{-1} is the quantile function of the standard normal distribution. For simplicity, we use the notation $T = \Phi^{-1}(1-\alpha/2)$ as the per-SNP threshold.

We declare all those SNPs with statistic $|S_A| > T$ to be associated with trait. So the per-causal-SNP power of a putative causal SNP A, which is the probability of $|S_A| > T$, can be calculated as

$$P_1(\mathrm{A}) = P(|S_A| > T) = \Phi\left(-T + \lambda_A\sqrt{N}\right) + 1 - \Phi\left(T + \lambda_A\sqrt{N}\right) \tag{5.4}$$

The average power $\overline{P_1}$ is obtained by averaging per-causal-SNP powers over all putative causal SNPs.

5.3.2 The Brute-force Approach for Pairwise Association Test

Current studies on complex disease have also suggested that some SNPs influence traits in pairs. Only when both causal alleles appear on a pair of SNPs, the trait value is increased. To detect the interaction of SNPs that influence the trait, we need to consider the association between a trait and a pair of SNPs (pairwise association test). We analyze the power of the brute force approach which calculates the association between a trait and all pairs of SNPs as follows.

We assume there exists a SNP pair AB, composed of SNP A and SNP B, that influence a trait. Assume the causal alleles are minor alleles at both SNPs. Our statistic is the difference between the average trait value of individuals who have minor alleles on both SNPs and that of individuals who do not have minor allele on at least one of the two SNPs A and B. Here we assume the two SNPs have same (positive) direction of effect. We use the same notation as in section 5.3.1. The expected number of individuals who have minor alleles at both SNPs can be computed as $N_{AB} = N \cdot p_A \cdot p_B$ and the expected number of individuals who do not have minor alleles at both SNPs can be computed as $N_{\neg AB} = N \cdot (1 - p_A \cdot p_B)$. If an individual carries the causal alleles at both SNPs A and B, the mean of trait value is increased or decreased by the effect size of the SNP pairs, which is

denoted as β_{AB}. Then we can write the distribution of y_i as

$$y_i \sim N(\mu + x_i^A x_i^B \beta_{AB}, \sigma^2) \tag{5.5}$$

Let $\bar{Y}_{AB}$ be the average trait value of individuals with causal alleles at both SNPs and let $\bar{Y}_{\neg AB}$ be the average trait value of individuals without causal alleles at both SNPs. For simplicity, let $\sum_{11}$ denote $\sum_{i:x_i^A=1 \wedge x_i^B=1}$, and similarly for $\sum_{10}, \sum_{01}, \sum_{00}$ for different alleles of SNPs A and B. We can calculate $\bar{Y}_{AB}$ and $\bar{Y}_{\neg AB}$ as

$$\begin{aligned} \bar{Y}_{AB} &= \frac{1}{N_{AB}} \sum_{11} y_i \sim N(\mu + \beta_{AB}, \frac{\sigma^2}{N p_A p_B}), \\ \bar{Y}_{\neg AB} &= \frac{1}{N_{\neg AB}} \sum_{00,01,10} y_i \sim N(\mu, \frac{\sigma^2}{N(1 - p_A p_B)}) \end{aligned} \tag{5.6}$$

We normalize the difference between $\bar{Y}_{AB}$ and $\bar{Y}_{\neg AB}$ to obtain the following statistic S_{AB}, which is normally distributed with mean $\lambda_{AB}\sqrt{N}$ (the non-centrality parameter) and unit variance.

$$S_{AB} = \frac{\bar{Y}_{AB} - \bar{Y}_{\neg AB}}{\sqrt{\frac{\sigma^2}{N p_A p_B (1 - p_A p_B)}}} \sim N(\lambda_{AB}\sqrt{N}, 1), \ \textit{where} \ \ \lambda_{AB} = \frac{\beta_{AB}\sqrt{p_A p_B (1 - p_A p_B)}}{\sigma} \tag{5.7}$$

According to [PP12], we set the per-SNP-pair significance level $\alpha = 10^{-12}$. The per-SNP-pair statistic threshold is then $T_2 = -\Phi^{-1}(\alpha/2) = 7.13$. The per-causal-SNP-pair power of a putative causal SNP pair AB can be estimated as

$$P_{BF}(\text{AB }) = \Phi\left(-T_2 + \lambda_{AB}\sqrt{N}\right) + 1 - \Phi\left(T_2 + \lambda_{AB}\sqrt{N}\right) \tag{5.8}$$

The average power $\overline{P_{BF}}$ is obtained by averaging per-causal-SNP-pair powers over all putative causal SNP pairs.

Assuming the total number of SNPs is M, we define the cost of brute-force method to be the total number of SNP pairs needed for association analysis, that is, $C_{BF}(M) = \binom{M}{2}$.

5.3.3 Two Stage Model

In the brute force approach, the total number of SNP pairs to be considered is $\binom{M}{2}$ and we need to compute the statistic S_{AB} for all these pairs. Considering the number of SNPs involved in current GWAS, the computational burden makes this strategy infeasible.

We propose a two-stage model to reduce the number of tests needed while maintaining similar power with the brute force approach. In the first stage, we propose two statistic thresholds T_a and T_b and perform the single marker test on all SNPs. In the second stage, we pair all SNPs that are significant under threshold T_a with those significant SNPs under threshold T_b. Then we perform a pairwise association test between traits and all those pairs. The SNP pairs which pass the per-SNP-pair statistic threshold T_2 are considered to be statistically associated with the trait.

The analysis of single marker test in the first stage is quite similar to that of the one SNP association test in Section 5.3.1. We derive the similar equations with (5.1), (5.2) and (5.3) except that the effect size of SNP A becomes $p_B\beta_{AB}$, when the pair of SNP A and SNP B is the causal SNP pair. So the statistic S_A of SNP A becomes

$$S_A = \frac{\bar{Y}_A - \bar{Y}_{\neg A}}{\sqrt{\frac{\sigma^2}{N \cdot p_A \cdot (1-p_A)}}} \sim N(\lambda_A\sqrt{N}, 1),\ \textit{where}\ \lambda_A = \frac{p_B\beta_{AB}\sqrt{p_A(1-p_A)}}{\sigma} \tag{5.9}$$

The analysis of SNP B is the same except that we switch p_A and p_B in the equations.

Assume a pair of SNPs A and B are putatively associated with a trait. The underlying effect size β_{AB} could either be positive or negative. Here we first analyze the case where the true effect size is positive. To find such positive pairwise association in our model, S_A must be no less than T_a, S_B must be no less than T_b (or vice versa, but here we only analyze one case since we will show in Section 5.3.6 that the other case is not necessary) and S_{AB} must be at least T_2.

Hence, we need to consider three statistics and three thresholds to compute the analytical power of the two-stage model. Under the assumption that we are aware the effect size is positive, the per-causal-SNP-pair power of a putative causal SNP pair AB can be denoted as

$$P_2^+(\text{AB}) = P(S_A \geq T_a, S_B \geq T_b \; and \; S_{AB} \geq T_2) \tag{5.10}$$

However, considering the fact that whether the effect size is positive or negative is hidden from us, we also need to calculate the probability where S_{AB} is less than $-T_2$, that is,

$$P_2^-(\text{AB}) = P(S_A \leq -T_a, S_B \leq -T_b \; and \; S_{AB} \leq -T_2) \tag{5.11}$$

So, the per-causal-SNP-pair power of a putative causal SNP pair AB is

$$P_2(\text{AB}) = P_2^+(\text{AB}) + P_2^-(\text{AB}) \tag{5.12}$$

The analysis for the case where the true effect size is negative is exactly the same except that the non-centrality parameters for S_A, S_B and S_{AB} are negative.

To calculate the value of $P_2(\text{AB})$ in Equation (5.12), we need to take into account correlations between statistics. The two statistics S_A and S_{AB} are correlated because both involve SNP A. Similarly, we have a correlation between S_B and S_{AB}. We assume SNPs are independent, and hence there is no correlation between S_A and S_B. The average power $\overline{P_2}$ is obtained by averaging per-causal-SNP-pair powers over all putative causal SNP pairs. Computing the analytical power of the two-stage model is complicated as a result of the correlations between statistics. We estimate the power using a multivariate normal distribution (MVN) framework as in Section 5.3.4.

Denote the per-SNP significance level corresponding to the statistic thresholds T_a and T_b in the first stage to be α_A and α_B, respectively. Then we have $\alpha_A =$

$2\Phi(-T_a)$ and $\alpha_B = 2\Phi(-T_b)$. The cost of the two stage model can be computed as $C_{TS}(M, \alpha_A, \alpha_B) \approx M^2\alpha_A\alpha_B$.

Let's measure the cost saving by the ratio between cost of brute-force method (C_{BF}) and that of the two-stage model (C_{TS}):

$$\frac{C_{BF}(M)}{C_{TS}(M, \alpha_A, \alpha_B)} = \frac{\binom{M}{2}}{M^2\alpha_A\alpha_B} \approx \frac{1}{2\alpha_A\alpha_B} \tag{5.13}$$

And we define the power loss to be

$$1 - \frac{\overline{P_2}}{\overline{P_{BF}}} \tag{5.14}$$

For a given dataset, there exists a trade-off between the power loss and cost saving. The trade off is controlled by the two thresholds T_a and T_b. We carefully design the thresholds to achieve high cost saving while maintaining low power loss. The details of the algorithm is summarized in Section 5.3.6.

5.3.4 Estimating the Two Stage Power Using the MVN

In this section, we provide an approach to compute the power of the two stage model in Equation (5.12). The distribution of association statistics S_A, S_B and S_{AB} has been derived in Section 5.3.2 and 5.3.3. We aim to compute the power in Equation (5.12) for any given thresholds T_a, T_b and T_2.

For many widely used statistical tests, the statistics over multiple markers asymptotically follow a Multivariate Normal Distribution(MVN) [SM05, Lin05]. To derive the analytical power of our two stage model, we use the framework of MVN proposed by [HKE09]. This method creates a MVN using the non-centrality parameters (NCPs) of statistics as a mean vector in the MVN. The NCPs of S_A, S_B, and S_{AB} are already derived in Equations (5.7) and (5.9). So the mean vector is $(\lambda_A\sqrt{N}, \lambda_B\sqrt{N}, \lambda_{AB}\sqrt{N})$. The covariance matrix in the MVN will be

the correlations among statistics. We assume SNPs are independent of each other, so the correlation between S_A and S_A is 1, and the correlation between S_A and S_B is 0. The covariance matrix is as follows:

$$\begin{pmatrix} 1 & 0 & Cor(S_A, S_{AB}) \\ 0 & 1 & Cor(S_B, S_{AB}) \\ Cor(S_A, S_{AB}) & Cor(S_B, S_{AB}) & 1 \end{pmatrix}$$

We only need to compute the correlation between S_A (or S_B) and S_{AB} to derive the complete MVN. To find a correlation between two statistics, S_A and S_{AB}, we use the following formula where $\mathrm{Var}(X)$ denotes the variance of X and $\mathrm{Cov}(X, Y)$ denotes the covariance between X and Y,

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y) \tag{5.15}$$

In our model, $X = S_A$ and $Y = S_{AB}$, and $\mathrm{Var}(S_A) = \mathrm{Var}(S_{AB}) = 1$. Then we can compute $\mathrm{Cov}(S_A, S_{AB})$ as

$$\mathrm{Cov}(S_A, S_{AB}) = (1/2)\mathrm{Var}(S_A + S_{AB}) - 1 \tag{5.16}$$

Hence, we need to derive $\mathrm{Var}(S_A + S_{AB})$ to find the covariance or the correlation between statistics. The covariance and the correlation are equivalent in this case because variances of statistics are 1.

Using Equations (5.7) and (5.9), we can write $S_A + S_{AB}$ as

$$S_A + S_{AB} = \sqrt{N/\sigma^2}\left(\theta_A\left(\bar{Y}_A - \bar{Y}_{\neg A}\right) + \theta_{AB}\left(\bar{Y}_{AB} - \bar{Y}_{\neg AB}\right)\right) \tag{5.17}$$

where $\theta_A = \sqrt{p_A(1 - p_A)}$ and $\theta_{AB} = \sqrt{p_A p_B(1 - p_A p_B)}$.

We then decompose $\bar{Y}_A$, $\bar{Y}_{\neg A}$ and $\bar{Y}_{AB}$ in Equation (5.17) in terms of alleles of SNPs A and B (x_i^A and x_i^B). Substituting Equations (5.1), (5.2), (5.5) and (5.6) into Equation (5.17) and rearranging common terms, we have

$$S_A + S_{AB} = \sqrt{\frac{N}{\sigma^2}} \left[P \sum_{11} y_i + Q \sum_{10} y_i - R \sum_{01} y_i - S \sum_{00} y_i \right] \tag{5.18}$$

where

$$P = \frac{\theta_A}{Np_A} + \frac{\theta_{AB}}{Np_Ap_B}, \quad Q = \frac{\theta_A}{Np_A} - \frac{\theta_{AB}}{N(1-p_Ap_B)}$$

$$R = \frac{\theta_A}{N(1-p_A)} + \frac{\theta_{AB}}{N(1-p_Ap_B)}, \quad S = \frac{\theta_A}{N(1-p_A)} + \frac{\theta_{AB}}{N(1-p_Ap_B)}$$

Note that Equation (5.18) consists of independent terms: each $\sum_{ab} y_i$ term represents a sum of trait values of disjoint individuals, where $ab = 11$, 10, 01 and 00, respectively. Hence, if we take the variance of $S_A + S_{AB}$, covariances among all terms are 0, and $\mathrm{Var}(S_A + S_{AB})$ is a sum of variances of $\sum_{ab} y_i$ terms. Also, note that $\mathrm{Var}(y_i) = \sigma^2$, and hence $\mathrm{Var}\left(\sum_{11} y_i\right)$ is a sum of σ^2 over individuals who have minor alleles at both SNPs A and B. We can then compute the variance of $S_A + S_{AB}$ as

$$\begin{aligned} &\frac{N}{\sigma^2}\left[P^2\mathrm{Var}\Big(\sum_{11} y_i\Big) + Q^2\mathrm{Var}\Big(\sum_{10} y_i\Big) + R^2\mathrm{Var}\Big(\sum_{01} y_i\Big) + S^2\mathrm{Var}\Big(\sum_{00} y_i\Big)\right] \\ &= N\left[P^2Np_Ap_B + Q^2Np_A(1-p_B) + R^2N(1-p_A)p_B + S^2N(1-p_A)(1-p_B)\right] \end{aligned} \tag{5.19}$$

We can also compute $\mathrm{Var}(S_B + S_{AB})$ similarly using Equation (5.19) by exchanging p_A and p_B.

Up to now we obtained all parameters for the MVN framework. Then, we can compute the power as the area outside of the significance threshold under the MVN we created. Fig. 5.2 helps to illustrate the ideas. We can see that in the three dimension space of the MVN framework for statistics S_A, S_B and S_{AB}, the two cubes on the corners correspond to the significance region. Using the MVN, we can compute the power of our two stage model for any given thresholds T_a, T_b and T_{AB} by summing up the volume of these two cubes under the MVN. This method yields a very accurate estimate of power when there exist correlations among statistics, and hence it provides an appropriate framework to compute the analytical power of our model.

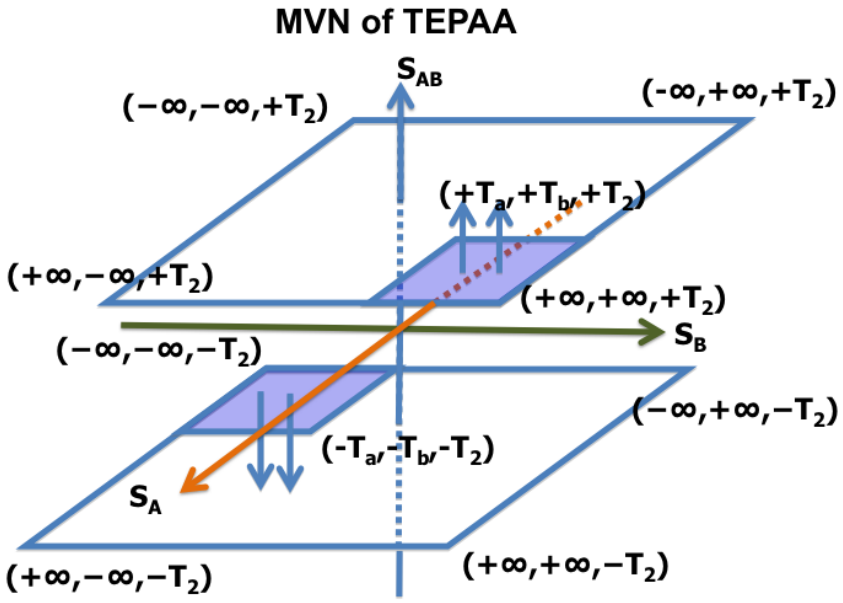


Figure 5.2: The volume of the two cubes under the MVN is the power of our two stage model.

5.3.5 Another Strategy to Computer the parameters of MVN

In Section 5.3.4, we proposed a complicated but step-by-step inference to compute the covariance of two test statistics in the MVN framework.

Now we study SNP pair AB from another direction. First we make a virtual SNP C. The allele of SNP C is exactly the same with the value of the SNP pair AB. The minor allele frequency of SNP C is denoted as $p_C = p_{AB} = p_A p_B$. The statistic S_C will be equivalent to statistic S_{AB}. Instead of computing $Cor(S_A, S_{AB})$ in the covariance matrix of the MVN, now we can compute $Cor(S_A, S_C)$.

The genotype of SNP A and SNP C are binary values under dominant model. The pearson correlation r_{AC} between the genotypes of SNP A and SNP C is then

$$r_{AC} = \frac{p_C(1-p_A)}{\sqrt{p_C(1-p_C)p_A(1-p_A)}} = \frac{p_A p_B(1-p_A)}{\sqrt{p_A p_B(1-p_A p_B)p_A(1-p_A)}} \tag{5.20}$$

Under the case where SNP pair AB are the causal SNP pair of the phenotype,

we are not observing SNP C but instead indirectly observing SNP A. Using the theory of indirect association study, the correlation $Cor(S_A, S_C)$ between the test statistic S_A and S_C is equal to r_{AC}. Similarly, we can compute the correlation $Cor(S_B, S_{AB})$. The correlation computed from the formula (5.20) will be exactly same with our calculation in Section 5.3.4. We prove it as follows.

Proof. In Section 5.3.4, we compute the correlation between S_A and S_{AB} as in Equation (5.16). And the variance of $S_A + S_{AB}$ is computed as in Equation (5.18).

Now let us simplify Equation (5.18) as follows.

$$
\begin{aligned}
& N\left[P^2Np_Ap_B + Q^2Np_A(1-p_B) + R^2N(1-p_A)p_B + S^2N(1-p_A)(1-p_B)\right] \\
= \;& N\left[P^2Np_Ap_B + Q^2Np_A(1-p_B) + R^2N(1-p_A)\right] \\
= \;& (\sqrt{\frac{1-p_A}{p_A}} + \sqrt{\frac{1-p_Ap_B}{p_Ap_B}})^2p_Ap_B + (\sqrt{\frac{1-p_A}{p_A}} - \sqrt{\frac{p_Ap_B}{1-p_Ap_B}})^2(p_A - p_Ap_B) \\
& +(\sqrt{\frac{p_A}{1-p_A}} + \sqrt{\frac{p_Ap_B}{1-p_Ap_B}})^2(1-p_A) \\
= \;& \frac{p_Ap_B}{p_A} - 2p_Ap_B + 2\frac{\sqrt{1-p_A}\sqrt{1-p_Ap_B}\sqrt{p_Ap_B}}{\sqrt{p_A}} + 1 \\
& +1 - p_A - 2\frac{\sqrt{1-p_A}\sqrt{p_Ap_B}\sqrt{p_A}}{\sqrt{1-p_Ap_B}} + \frac{p_Ap_Bp_A}{1-p_Ap_B} \\
& -\frac{p_Ap_B}{p_A} + p_Ap_B + 2\frac{\sqrt{1-p_A}\sqrt{p_Ap_B}p_Ap_B}{\sqrt{p_A}\sqrt{1-p_Ap_B}} - \frac{p_Ap_Bp_Ap_B}{1-p_Ap_B} \\
& +\frac{p_A}{1-p_A} + 2\frac{\sqrt{p_A}}{\sqrt{1-p_A}}\frac{\sqrt{p_Ap_B}}{\sqrt{1-p_Ap_B}} + \frac{p_Ap_B}{1-p_Ap_B} \\
& -\frac{p_Ap_A}{1-p_A} - 2\frac{\sqrt{p_A}\sqrt{p_Ap_B}p_A}{\sqrt{1-p_A}\sqrt{1-p_Ap_B}} - \frac{p_Ap_Bp_A}{1-p_Ap_B} \\
= \;& 2 + 2\frac{\sqrt{1-p_A}\sqrt{1-p_Ap_B}\sqrt{p_Ap_B}}{\sqrt{p_A}} - 2\frac{\sqrt{1-p_A}\sqrt{p_Ap_B}\sqrt{p_A}}{\sqrt{1-p_Ap_B}} + 2\frac{\sqrt{1-p_A}\sqrt{p_Ap_B}p_Ap_B}{\sqrt{p_A}\sqrt{1-p_Ap_B}} \\
& +2\frac{\sqrt{p_A}}{\sqrt{1-p_A}}\frac{\sqrt{p_Ap_B}}{\sqrt{1-p_Ap_B}} - 2\frac{\sqrt{p_A}\sqrt{p_Ap_B}p_A}{\sqrt{1-p_A}\sqrt{1-p_Ap_B}} \\
= \;& 2 + 2\frac{\sqrt{1-p_A}\sqrt{1-p_Ap_B}\sqrt{p_Ap_B}}{\sqrt{p_A}} - 2\frac{\sqrt{1-p_A}\sqrt{p_Ap_B}\sqrt{p_A}}{\sqrt{1-p_Ap_B}} + 2\frac{\sqrt{1-p_A}\sqrt{p_Ap_B}p_Ap_B}{\sqrt{p_A}\sqrt{1-p_Ap_B}} \\
+ \;& 2\frac{\sqrt{p_A}\sqrt{p_Ap_B}\sqrt{1-p_A}}{\sqrt{1-p_Ap_B}} \\
= \;& 2 + 2\frac{\sqrt{1-p_A}\sqrt{1-p_Ap_B}\sqrt{p_Ap_B}}{\sqrt{p_A}} + 2\frac{\sqrt{1-p_A}\sqrt{p_Ap_B}p_Ap_B}{\sqrt{p_A}\sqrt{1-p_Ap_B}} \\
= \;& 2 + 2\frac{p_Ap_B(1-p_A)}{\sqrt{p_A(1-p_A)}\sqrt{p_Ap_B(1-p_Ap_B)}}
\end{aligned}
$$

So now we have the correlation between S_A and S_{AB} as

$$
\begin{aligned}
Cor(S_A, S_{AB}) &= (1/2)\mathrm{Var}(S_A + S_{AB}) - 1 \\
&= 0.5 * (2 + 2\frac{p_Ap_B(1-p_A)}{\sqrt{p_A(1-p_A)}\sqrt{p_Ap_B(1-p_Ap_B)}}) - 1 \\
&= \frac{p_Ap_B(1-p_A)}{\sqrt{p_A(1-p_A)}\sqrt{p_Ap_B(1-p_Ap_B)}}
\end{aligned}
$$

Since $p_{AB} = p_A p_B$, this is exactly same with the correlation in the Equation (5.20).

□

Now we conclude that we can just use the correlation in Equation(5.20) to compute the correlation between S_A and S_{AB}. And we obtained all parameters of the MVN using this simple formula to compute the power of TEPAA.

5.3.6 Efficient Pairwise Association Test Using TEPAA

In previous sections, we have illustrated how to calculate the power and cost savings of our two stage model for any given threshold. In this section, we provide a framework, TEPAA, to determine the first thresholds which generate a relatively small number of SNP pairs for pairwise association test in the second stage while losing a small amount of power compared to the brute force approach.

From Equation (5.12) and Section 5.3.4, we can see that the joint distribution between the association statistics of single SNPs and the association statistic of a pair of SNPs depends on the MAFs of the pair of SNPs. MAFs are observable values, so we can categorize all SNP pairs based on the combination of their MAFs. Since MAFs are continuous value, we can discretize the MAFs into bins to have a small number of combinations. After removing rare variants, we can categorize all SNPs into 9 bins, with step size 0.05. In order to detect the pairwise association for all SNP pairs, we break all combinations of SNP pairs into two cases. First we pair SNPs within different bins and this results in $\binom{9}{2}$ categories. The second case is to combine SNPs within one bin. So totally we have $\binom{9}{2}+9$ categories of SNP pairs.

Assuming the power of the brute force approach is 50%, we can calculate the effect size β_{AB} from Equation (5.8). Then for each category of SNP pairs, we can compute the power loss and cost savings from Equations (5.13) and (5.14) with the MVN, given two first stage significance levels α_A and α_B. We do an exhaustive

search over the space $[0, 1)$ with a small step size to find the optimal values of α_A and α_B to achieve best cost saving while maintaining power loss 1%. The values of α_A and α_B are shown in Table 5.1 when there are $5,326$ samples in the dataset.

For SNPs in each bin, we carry out the single marker test and sort the association statistics of single SNPs. Then for each category of SNP pairs, we do a binary search in each involved bin to find all SNPs that pass the pre-computed thresholds. The selected SNPs are then paired for the second stage pairwise association test. Based on the pre-computed values of α_A and α_B, we can estimate the cost savings for each category of SNP pairs as in Table 5.1. We propose a threshold for each bin for each category of SNP pairs, and the bins are disjoint. So, in the calculation of Equation (5.10), we only need to consider the case where $S_A > T_a$ and $S_B > T_b$ and it is not necessary to consider the case $S_A > T_b$ and $S_B > T_a$. We have the same conclusion in the calculation of Equation (5.11).

We summarize the framework of $TEPAA$ as in Algorithm 3.

Although the calculation is based on the assumption that the brute force approach has power 50%, our approach is robust to the effect size. We did simulations for different effect sizes, which generate different power for the brute force approach. The cost saving of TEPAA is stable when achieving 1% power loss under various effect size.

5.4 Discussions

In this work, we proposed a two-stage model to detect SNP pairs associated with trait. The key idea behind our method is that we model the joint distribution between association statistics at single SNPs and association statistics at pairs of SNPs to allow us to apply a two-stage model that provides guarantees that we detect associations of pairs of SNPs with small number of tests while losing very little power. We rapidly eliminate from consideration pairs of SNPs which with

Algorithm 3: Framework of TEPAA

Input: A GWAS data set with genotype and phenotype for each individual.

Output: SNP pairs associated with the phenotype.

1 Remove rare variants, categorize rest SNPs into 9 bins according to MAFs, with step size 0.05.

2 Pre-compute the thresholds for each combination of bins as in Table 5.1, which only depends on the second stage threshold.

3 For SNPs in each bin, we carry out the single marker test and sort the association statistics of single SNP.

4 For each category of SNP pairs, we do a binary search in each involved bin to find all SNPs that pass the pre-computed thresholds in Table 5.1.

5 Pair up the selected SNPs with positive statistics from different bin in Step (4) to perform pairwise association test. Then pair up the selected SNPs with negative statistics in Step (4) to perform pairwise association test.

high probability are not associated with the trait. Using extensive simulations, we show that our approach can reduce computation time by a factor of 60 while only losing approximately 1% of the power obtained by the brute-force approach.

Table 5.1: The threshold for SNP A/SNP B and cost savings in various combination of MAFs to achieve power loss of 1%. Here we assume the MAF of SNP A is smaller than that of SNP B in each pair. The first and second number in each cell is the threshold for SNP A (α_A) and SNP B (α_B), respectively. These two thresholds are scaled by 10^{-2}. The third number in each cell is the cost saving, which is the ratio between cost of brute-force method and that of the two-stage model.

MAF of SNP B

MAF of SNP A	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.1	34/34 /8	8/50 /25	7/58 /25	5/62 /32	2/76 /66	0.82/84 /145	0.26/79 /487	0.10/84 /1190	0.02/90 /5555
0.15	-	14/14 /51	3/24 /139	3/31 /107	2/46 /108	1/58 /172	0.35/54 /529	0.13/62 /1241	0.03/69 /4830
0.2	-	-	5/5 /400	2/9 /556	2/16 /312	1/21 /476	0.47/31 /686	0.19/58 /907	0.05/69 /2899
0.25	-	-	-	3/3 /1100	2/5 /1000	1/7 /1429	1/16 /625	0.26/21 /1831	0.10/42 /2380
0.3	-	-	-	-	1/1 /1e5	1/3 /3333	1/4 /2500	0.62/12 /1344	0.13/16 /4807
0.35	-	-	-	-	-	0.6/0.6 /2.7e4	0.5/1 /2e4	0.1/2 /5e4	0.03/8 /4e4
0.4	-	-	-	-	-	-	0.3/0.3 /1.1e5	0.1/0.6 /1.6e5	0.1/1 /1e5
0.45	-	-	-	-	-	-	-	0.2/0.2 /2.5e5	0.1/0.5 /2e5
0.5	-	-	-	-	-	-	-	-	0.1/0.1 /1e6

CHAPTER 6

Fast Detection of IBD Segments Associated With Quantitative Traits

6.1 Background

Two individuals are identical-by-descent (IBD) at a locus if they have alleles inherited from a recent common ancestor. Several methods have been developed to detect the IBD segments between purportedly unrelated individuals. The current state-of-the-art methods such as GERMLINE [GLS09] and Beagle [BB10, BB11] can detect even small (2 centimorgan) IBD segments shared between individuals from whole genome sequence data. The IBD segments detected by an IBD detection method can be used in various applications such as haplotype phasing [KMF08], imputation [JAS12] and heritability analysis in founder populations [PHT11, ZHS12, BB13].

One promising application of IBD information is in association mapping [PNT07, GKL11, BT12, HKR13]. The traditional approach for association mapping is to perform a statistical test between a single SNP and the observed case/control status or quantitative phenotypes. These single-SNP-based association testing approaches are designed to have high power to detect association for common SNPs (minor allele frequency > 0.01). Unfortunately, rare causal variants will not be identified by these traditional approaches. Association testing based on IBD information is an alternative to standard association testing methods which may have advantages for discovering associations in loci where rare variants play

a role.

Rare causal variants are likely to have been introduced into a population recently. These mutations are initially "private" to the individual in which they occurred, but then are passed on to progeny. IBD segments containing these recently derived rare alleles are likely to be discovered, because these rare alleles actually can help IBD detection algorithms to detect IBD segments between individuals. If the shared IBD segments contains these rare causal alleles, IBD mapping approaches can identify the loci harboring the rare causal mutation through the association mapping between IBD segments and the phenotypes of interest.

Two categories of methods have been proposed to discover IBD segments associated with the phenotype. The first category of methods compare the IBD rate of case/case pairs with the background IBD rate to detect excessive IBD between cases, and is referred to as pairwise methods [PNT07, BT12, HKR13]. The motivation for pairwise methods is that if a rare variant occurred in a relatively recent ancestor, cases are more likely to share an IBD segment containing the causal variant. The second category of methods is referred to as clustering methods [GKL11]. Individuals are divided into clusters based on the IBD information, and then each cluster is tested for association assuming that the cluster tags a rare causal variation.

There are several computational challenges in pairwise methods. The first challenge is computational inefficiency. In pairwise methods, since the statistic is dependent on two individuals sharing IBD segments, it is difficult to analytically obtain the asymptotic distribution of the test statistic. In order to compute the p-value for the test statistic, one needs to approximate the null distribution of the test statistic through permutations, where the vector of phenotype traits is permuted. In the genome-wide association studies (GWAS), the p-value threshold is necessarily low due to multiple testing [BT12]. Thus one must perform a large number of permutations, which can be computationally demanding. The

second challenge is fine-mapping. In GWAS, after one identifies significant loci, it is important to pinpoint the most significant peak within the loci for follow-up studies. However, in the permutation test, the smallest p-value one can estimate is constrained by the number of permutations, often resulting in many SNPs with the same minimal p-values in the region.

Previously proposed pairwise methods are only applicable to case/control data since they explicitly classify each IBD segment as either being shared between two case individuals or otherwise. In this section, we present a IBD association mapping method designed for quantitative traits. In our method, we first construct the IBD graph based on detected IBD information given by IBD detection algorithm at a locus similar to case/control data [HKR13]. Then the test statistic for hypothesis testing can be computed based on the graph representation of the IBD information, which is referred to as the edge-based statistic. Similar to the pairwise method, the asymptotic distribution of the edge-based statistic is not easily obtained. Thus assessing the significance of the association requires permutation testing, which becomes a great burden when we obtain small p-values. However we show that permutation testing is not necessary, by showing the connection between the edge-based statistic and a linear model. We demonstrate the equivalence between the permutation test and the linear model both analytically and empirically on real data. Using the linear model, we can obtain the p-values for each locus very efficiently.

A further advantage of the connection to linear models is that we can include any covariate and/or random effects terms in the model, because the proposed IBD mapping statistic is reduced to a simple linear model. Incorporating study-specific covariates such as age, sex and other environment factors in the model can greatly improve the statistical power of the association mapping. The ability to include random effects term in the model is particularly useful for controlling population structure. In IBD association mapping, if two individuals are closely

related, their genomes are more likely to share an IBD segment at each genomic locus. In addition, if they share the causal variants, their phenotype will also be similar. This causes a correlation between the IBD structure and the phenotype at many loci, which leads to false positive association signals and inflation of p-values. To correct for the population structure caused by the genetic relatedness between individuals, we utilize a mixed model and include a random effect term which follows normal distribution with covariance of kinship matrix reflecting the closeness between individuals. We demonstrate that our method can control the population structure by applying it to the 1966 North Finland Birth Cohort Data for 10 phenotype traits.

6.2 Methods

6.2.1 The IBD graph

Given N individuals, the IBD information at a genomic locus can be represented as an IBD graph with N vertices (Figure 6.1). An edge exists between a pair of vertices if two individuals are IBD at the locus. The value y_i for each vertex i is the trait value of the corresponding individual, and the vector $Y = (y_1, y_2, ..., y_N)$ contains the phenotypes for all individuals.

6.2.2 Edge-based IBD association mapping statistics

Let V be the set of individuals and let E be the set of edges in the IBD graph, that is, all IBD relationships. We define the *edge statistic* for IBD association mapping at genomic locus k as

$$S_k = \sum_{(i,j)\in E} (y_i + y_j) \tag{6.1}$$

The intuition behind this statistic is that, if a genomic locus contains the causal mutation affecting the phenotype of interest, we would expect that individuals

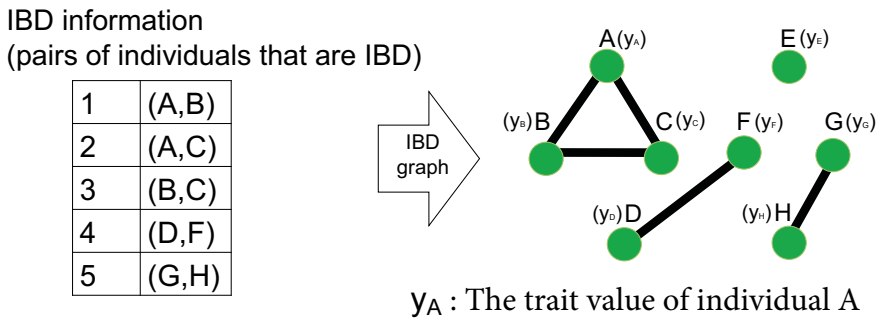


Figure 6.1: An example of IBD graph. IBD detection method provides IBD information as shown in the table. Then we build a graph where vertices are individuals and edges are IBD relationships.

sharing IBD at this locus tend to have higher or lower phenotype values than others not sharing IBD. Each of these values contribute more than once to the statistic S_k and will distinguish the associated genomic locus from other loci. Since the statistics are based on the edges which are dependent on two individuals, asymptotic distribution of S_k is difficult to obtain analytically. In this case, one straightforward way to compute the significance of the association is through permutation.

6.2.3 Permutation Test

To approximate the distribution of S_k under null hypothesis, we can permute the phenotype of all individuals. Let

$$v = (v_1, v_2, ..., v_N), \forall v_i \in Y$$

be the vector of trait values of N individuals, where v_i denotes the phenotype value for i-th individual in the permutation. A single permutation can be thought of as randomly permuting a vector of the trait values. The test statistic, S_k, is a function of v. Let $\hat{v}$ be the vector of observed phenotype vector. The standard permutation test is equivalent to sampling a new v from all possible permutations of $\hat{v}$ assuming

a uniform distribution. Let B be the set of sampled v. The estimated p-value is

$$\hat{p} = \frac{1}{|B|} \sum_{v \in B} \delta(|S_k(v)| \geq |S_k(\hat{v})|) \tag{6.2}$$

where δ is the indicator function. The drawback of this approach is its inefficiency because it requires a large number of permutations to obtain a small p-value. The denominator $|B|$ in Equation (6.2) needs to be large enough to make the value $\hat{p}$ small and thus large number of permutations are required. To assess a p-value p with standard error $p/10$, we need approximately $100/p$ permutations.

6.2.4 IBD-degreetype

To obtain the edge statistic S_k, we sum the trait values involved with each edge in the IBD graph. From the view of vertices, the trait value of each individual contributes d_i times to the statistic S_k, where d_i is the degree of the corresponding vertex in the IBD graph. We introduce a concept called the *IBD-degreetype* which is simply the degree of each individual in the IBD graph. We denote $D = (d_1, d_2, ..., d_N)$ to be the vector of IBD-degreetypes of N individuals. Obtaining the degrees of vertices is equivalent to splitting all edges and counting how many edges are adjacent to each vertex (Figure 6.2). Then we assign these numbers to the vertices. Given this, we consider the IBD-degreetype as conceptually similar to a genotype where the alleles of each individual are analogous to the degree of corresponding vertex in the IBD graph.

The IBD-degreetypes can be used for statistical testing in the IBD association mapping. According to the definition of IBD-degreetype, we can rewrite the test statistic S_k as

$$S_k = \sum_{i \in V} d_i y_i = D^T Y \tag{6.3}$$

We refer to this statistic as the *sum statistic*. The intuition is that individuals sharing IBD segments containing causal variants are likely to have similar (high)

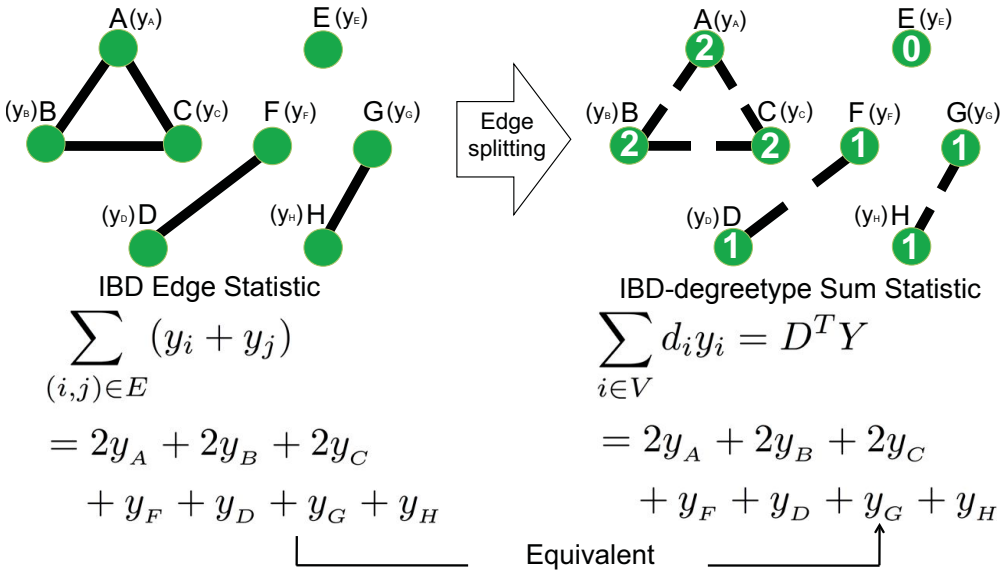


Figure 6.2: Equivalence between two IBD statistics.

trait value and have higher degrees. The trait value of individuals with causal haplotype at this locus can be aggregated by the weighted sum of trait values, where the weight for each individual is the corresponding degree in the IBD graph, which is what Equation (6.3) computes exactly. In the next section, we show how this property could help us to compute the p-value efficiently.

6.2.5 Efficient computation of p-values

The formulation of the statistic in equation (6.3) closely resembles the regression estimator in linear models. We can use this observation with an additional assumptions to obtain p-values analytically which eliminates the need for performing permutation.

If we assume that the phenotype follows a normal distribution with variance σ^2, then we can represent the phenotype using the linear model which includes the IBD-degreetype and the effect of the IBD-degreetype on the phenotype γ

$$y_i = \mu + \gamma d_i + \epsilon_i \tag{6.4}$$

where ϵ_i is normally distributed with mean 0 and variance σ^2, $\epsilon_i \sim N(0, \sigma^2)$. Written using vector notation

$$Y = \mu\mathbf{1} + \gamma D + \mathbf{e} \tag{6.5}$$

where $\mathbf{1}$ is a column vector of "1"s and $\mathbf{e}$ is a random vector where each element is independent and has variance σ^2.

This can be represented using a multivariate normal distribution where the covariance matrix is $\sigma^2\mathbf{I}$ and $\mathbf{I}$ is the identity matrix, $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$. We note that if the region is not involved in the phenotype, then $\gamma = 0$. However, if the region is involved in the phenotype, then $\gamma \neq 0$. We can obtain an estimate of γ, using ordinary least squares (OLS) estimates.

$$\hat{\gamma} = \frac{D^T Y - \frac{1}{N} * \mathbf{1}^T D * \mathbf{1}^T Y}{D^T D - \frac{1}{N} * (\mathbf{1}^T D)^2} \tag{6.6}$$

$$\hat{\mu} = \frac{1}{N}\mathbf{1}^T Y - \frac{1}{N}\hat{\gamma}\mathbf{1}^T D \tag{6.7}$$

The estimated residuals $\hat{\epsilon}_i = y_i - \hat{\mu} - \hat{\gamma} d_i$ can be used to estimate the standard error $\hat{\sigma} = \sqrt{\frac{\hat{\mathbf{e}}^T\hat{\mathbf{e}}}{N-2}}$. Since the studies are large, the association statistic will approximately follow the normal distribution

$$\hat{\gamma} \sim \mathbf{N}\left(\gamma, \frac{\hat{\sigma}^2}{N}\right) \tag{6.8}$$

We note the close relationship between Equation (6.6) and Equation (6.3). Since $\frac{1}{N} * \mathbf{1}^T D * \mathbf{1}^T Y$ and $D^T D - \frac{1}{N} * (\mathbf{1}^T D)^2$ are all constants in Equation (6.6), we denote them as C_1 and C_2 respectively. Now we can derive S_k from $\hat{\gamma}$ by scaling a constant factor C_2 and then shifting a constant factor C_1 as follows

$$S_k = D^T Y = C_2\hat{\gamma} + C_1 \tag{6.9}$$

So S_k will approximately follow the following normal distribution

$$S_k \sim \mathbf{N}(C_2^2\gamma + C_1, \frac{C_2^2\hat{\sigma}^2}{N}) \tag{6.10}$$

Under the null hypothesis where $\gamma = 0$, we can obtain the p-value of S_k using the quantile of the normal distribution without needing to apply permutation test. We declare the investigated locus to be significant if $|\frac{S_k - C_1}{C_2 \hat{\sigma}/\sqrt{N}}| \geq \phi^{-1}(1 - \alpha/2)$, where α is the significance level.

We see that the p-value of S_k is equal to the p-value of $\hat{\gamma}$ since there is a linear transformation between S_k and $\hat{\gamma}$. The permutation test just gives another way to compute the p-value of S_k, where the null distribution of S_k is approximated by permuting the vector Y. So, we can compute the p-value of S_k rapidly using the linear model.

6.2.6 Control for population structure

There are two reasons that population structure affects association mapping. The first is that variants other than the one which is being tested in the statistical test might affect the phenotype. The second is that different individuals might have different total amounts of shared IBD segments. We extend our proposed IBD mapping method to correct for the effect of populations structure due to both reasons. The first challenge is that the results can be confounded by relatedness among the individuals affecting variants outside the locus under consideration. Intuitively, if two individuals are closely related, at each position in their genome they are more likely to share an IBD segment. In addition, their genetic relatedness will cause their phenotypes to be more similar. This causes an apparent correlation between the IBD-degreetype and the phenotype at each position in the genome.

In order to motivate how we address this problem we first consider the standard Fisher polygenic model where each variant in the genome affects the phenotype independently. In this case, the generative model for the phenotype is

$$y_j = \mu + \sum_{i=1}^{M} \beta_i x_{ij} + \epsilon_j \tag{6.11}$$

where the effect of each variant on the phenotype is β_i, the phenotypic mean is μ and ϵ_j is the contribution of the environment on the phenotype which is normally distributed with variance σ_e^2, denoted $\epsilon_j \sim N(0, \sigma_e^2)$. Since most of the variants do not affect the phenotype, $\beta_i = 0$ for most variants. We note that the inherent assumption for this model is the "additive" assumption in that the variants all contribute linearly to the phenotype value and ignore more sophisticated phenomenon which include non additive effects or gene-by-gene interactions.

If we denote the vector of phenotypes Y and vector of effect sizes β, the matrix of genotypes X and the vector of environmental contributions $\mathbf{e}$, then the model for the population can be denoted as

$$Y = \mu\mathbf{1} + X\beta + \mathbf{e} \tag{6.12}$$

where $\mathbf{1}$ is a column vector of 1's, and $\mathbf{e}$ is a random vector drawn from the multivariate normal distribution with mean 0 and covariance matrix $\sigma_e^2\mathbf{I}$, denoted as $\mathbf{e} \sim N(0, \sigma_e^2\mathbf{I})$.

Our IBD statistic Equation (6.3) makes the same assumptions as linear regression which assumes that the phenotype of each individual is independently distributed. Unfortunately, this is not always the case. The reason is due to the discrepancy between the statistical model in Equation (6.3) which is used for testing compared to the true genotype phenotype model in Equation (6.11) which generated the data. If we are considering region k and represent the variants which are not in the region with $i \notin k$, the terms which are missing from the testing model, $\sum_{i \notin K} \beta_i x_{ij}$, are referred to as unmodeled factors. These unmodeled factors correspond to the variants that affect the phenotype in the genome other than the variant which is being tested in the statistical test. After we incorporate the IBD-degreetype, the generative model can be denoted as

$$Y = \mu\mathbf{1} + \gamma_k D_k + \sum_{i \neq k} \beta_i x_i + \mathbf{e} \tag{6.13}$$

If the values for these unmodeled factors are independently distributed, then these factors will increase the amount of variance, but not violate the independently distributed assumption of the statistics. However, if the unmodeled factors are not independently distributed, which is the case when individuals in the sample are related to different degrees. Then this will violate the assumptions of the statistical test in Equation (6.3).

This problem is referred to as "population structure" where differing degrees of relatedness between individuals in the GWAS cause an inflation of the values of the association statistics leading to false positives. Many methods for addressing population structure have been presented over the years including genomic control [DR99] which scales the statistics to avoid inflation, principal component based methods [PPP06] and most recently mixed model methods [KZW08, KSS10, LLL11, ZS12].

The basis of the mixed model approach to correct population structure is the insight that the proportion of the genome shared corresponds to the expected similarity in the values of the unmodeled factors. More precisely, the covariance of the unmodeled factors is proportional to the amount of the genome shared. The amount of genome shared is referred to as the "kinship matrix" and since the genotypes are normalized, the kinship is simply $\mathbf{K} = XX^T/M$ where X is the $N \times M$ matrix of the normalized genotypes. We then add a term to the statistical model to capture these unmodeled factors resulting in the statistical model

$$y = \mu\mathbf{1} + \gamma_k D_k + \mathbf{u} + \mathbf{e} \tag{6.14}$$

where $\mathbf{e} \sim N(0, \sigma_e^2\mathbf{I})$ and $\mathbf{u} \sim N(0, \sigma_g^2\mathbf{K})$. $\mathbf{u}$ represents the contributions of the unmodeled factors and $\mathbf{e}$ represents the effect of non-genetic factors on the phenotype. When performing an association, mixed model methods estimate the maximum likelihood for parameters μ, γ_i, σ_g^2 and σ_e^2 using the likelihood

$$L(N, y, \mu, \sigma_e^2, \sigma_g^2, \mathbf{K}) = (2\pi)^{-\frac{N}{2}} |\sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}|^{-\frac{1}{2}} e^{-\frac{1}{2}(y-\mu)^T(\sigma_g^2\mathbf{K}+\sigma_e^2\mathbf{I})^{-1}(y-\mu)} \tag{6.15}$$

One intuition to explain mixed models is that they decompose the variance of phenotype into a portion corresponding to the genetics ($\mathbf{u}$) and a portion corresponding to the environment ($\mathbf{e}$). The idea behind our method is that we can use the mixed model to obtain the values of the genetic portion and then remove them from the phenotypes obtaining a set of corrected phenotypes which are not affected by population structure. The way this is done is that after the estimates of σ_g^2 and σ_e^2 are estimated, we can then compute the maximum likelihood estimates for $\hat{u}_i$. Our new phenotypes are then $y_i' = y_i - \hat{u}_i$ and can be used in Equation (6.5).

The second challenge comes from the fact that some individuals have more IBD segments than others. If some individuals are closely related to each other, they will have higher IBD-degreetype over the genome and their phenotype will contribute many times to the test statistic S_k, which further increases the variance of our test statistic. We normalize the IBD-degreetype for each individual by subtracting the mean of IBD-degreetype over the genome, which addresses the problem.

6.3 Results

6.3.1 Equivalence between the permutation test and the linear model

The asymptotic distributions of the statistic in Equation (6.1) is difficult to obtain analytically. This is because the statistic is based on the edges that depend on pair of individuals. For this reason, we have to do a permutation test to assess the statistical significance. However, the permutation test is computationally inefficient. If the true p-value is small, which is required in genome-wide association studies, we will need a large number of permutations. For the genome-wide threshold of IBD association testing(6×10^{-6}, [BT12]), more than 10 million permutations are required.

We have shown that the edge statistic in Equation (6.1) is equivalent to the sum statistic in Equation (6.3). We further demonstrated in Equation (6.10) that the sum statistic S_k will approximately follow the normal distribution under the null hypothesis and can be used to determine whether the test statistic is significant. We here show by running experiments on the 1966 North Finland Birth Cohort (NFBC66) dataset to further confirm the equivalence between permutation test and linear model.

We first run Beagle [BT12] to obtain the IBD segments with threshold 10^{-6}, which is a commonly used threshold. Then we build a IBD graph for each genome position. In the IBD graph, each vertex corresponds to one individual, and we connect two vertices with an edge if the two individuals share an IBD segment at this position. The IBD-degreetype is simply the degree of each vertex as defined.

We first compute the test statistic using Equation (6.3). The test statistic is computed for the phenotype body mass index (BMI). Then we permute the phenotype $10,000$ times and compute the corresponding test statistic for each permutation to approximate the null distribution of S_k. The p-value is then estimated using Equation (6.2).

We also compute the p-value for the association between each locus and the phenotype BMI using model in Equation (6.14), which is much faster. The correlation between p-values computed from permutation test and linear model is plotted as in Figure 6.3. We can see that the p-values computed from the two methods are highly correlated. This confirms the correctness of our proposal method that we can use the linear model to compute the p-value for each genome position, instead of doing permutation test, which is computational inefficient.

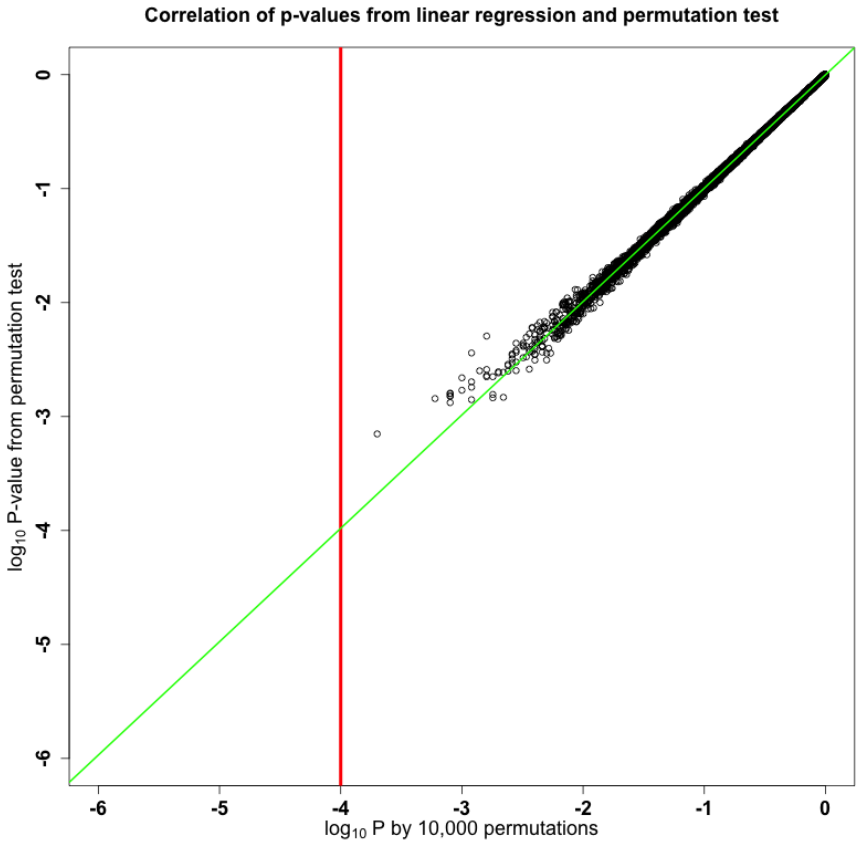


Figure 6.3: The correction between p-values computed from permutation test and linear model. The red vertical line represents the lower bound of p-values that permutation test can approximate given the number of permutations.

6.3.2 Correcting for population structure

We applied our method to the 1966 North Finland Birth Cohort (NFBC66) data to detect the IBD segments associated with the quantitative traits. The NFBC66 data contains genotypes over $330,000$ SNPs for $5,326$ individuals. Ten quantitative phenotypes are collected for each individual. We first applied Beagle [BB10, BB11] to detect the IBD information from the genotypes. The output of Beagle shows the IBD segments shared between individuals across the genome. Then for each genomic locus, we represent the IBD information using an IBD

graph as defined in section 6.1. Each vertex in the graph represents an individual and we build an edge between two vertices if the corresponding individuals are IBD at this locus. The IBD-degreetype is computed for each vertex. We estimate the p-value of each variant using a linear model.

Since the population structure may cause substantial inflation of test statistic and possibly spurious association, we evaluate the performance of our method using the inflation factor. The inflation factor λ is the ratio of median chi-squared test statistics to the median of an expected 1 degree-of-freedom chi-squared distribution [DR99]. An inflation factor greater than 1 indicates the presence of inflation. We first applied the linear model without correction for population structure over all ten phenotypes. From the middle column of Table 6.1, we can see that inflation exists for most phenotypes.

In order to correct for the population structure, we incorporate a random effect term into our linear model. We first compute a pairwise relatedness matrix, the kinship matrix, from genotypes to represent the population structure. Then we estimate the contribution of the population structure to the phenotype using a variance component model, resulting in an estimated covariance matrix of phenotypes. The covariance matrix models the effect of genetic relatedness on the phenotypes. Finally we applied a generalized least square test at each variant to detect the association. The inflation factors are also computed for all ten phenotypes and the results are summarized in the third column of Table 6.1. We can see that we can decrease the inflation factor in most cases. In four cases, the inflation factor increases slightly. We also show the distribution of inflation factors in a box plot as in Figure 6.4. We can see that with population structure correction, the inflation factor is corrected to be close to 1.

Phenotype	without correction	with correction
crp	1.028	1.029
height	1.065	0.904
dia	1.067	1.075
glu	1.074	1.039
hdl	0.977	0.981
ins	0.988	1.000
ldl	1.055	0.975
sys	1.132	1.076
bmi	1.000	0.983
tg	1.011	0.995

Table 6.1: Inflation factors for ten phenotypes from NFBC66 data. Phenotype abbreviations are CRP, C-reactive protein; TG, triglyceride; INS, insulin plasma levels; DBP, diastolic blood pressure; BMI, body mass index; GLU, glucose; HDL, high-density lipoprotein; SBP, systolic blood pressure; LDL, low density lipoprotein.

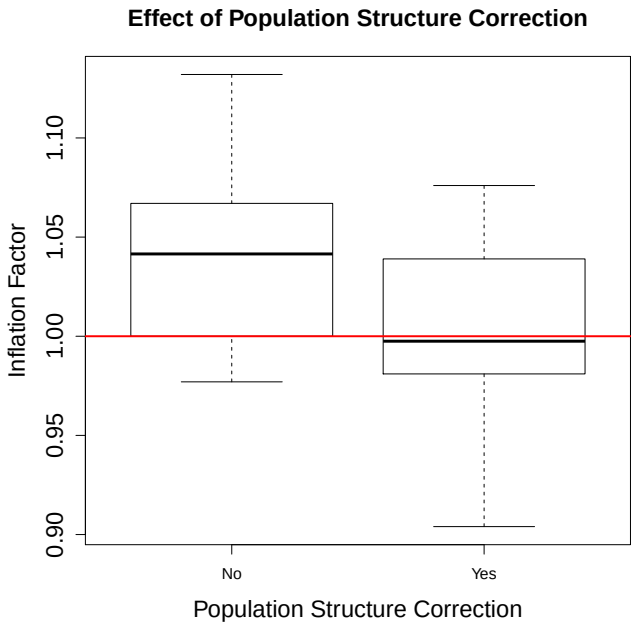


Figure 6.4: A Distribution of inflation factors of IBD mapping statistics on NFBC66 data, without (No) and with (Yes) population structure correction respectively.

6.4 Discussions

In this section, we proposed a test statistic and a fast approach to detect the significant IBD segments associated with quantitative phenotype traits. Previous methods have been proposed to detect significant IBD segments in case-control data, but are not suitable for continuous phenotypes. We proposed a test statistic for continuous traits based on the IBD graph, which is built from the IBD information. In the IBD graph, each vertex represents an individual and the edges between vertices indicate the presence of IBD between the two individuals. Since the asymptotic distribution of the test statistic is hard to derive analytically, we conduct the permutation test to compute the p-value for each SNP. The drawback of this approach is its inefficiency because it requires a large number of permutations to obtain a small p-value. We further proposed a linear model where the

independent variable is the IBD-degree type. We proved the equivalence between the p-value of the coefficient in the linear model and the p-value from the permutation test, both analytically and from simulation. The linear model is a fast approach. However, one more challenge is the population structure, where the differing degrees of relatedness between individuals in the GWAS cause an inflation of the values of association statistics leading to false positives. We incorporated this relatedness into our linear model to correct for the population structure. We applied our method to the North Finland Birth Cohort Data and determined that our method can correct the inflation factor toward 1.

The true utility of the IBD association testing is on detecting significant associations on rare variants that cannot be found using single SNP tests [BT12]. IBD association testing can be conducted without additional cost compared to traditional GWAS on genotype data. The only extra effort is to compute IBD segments from genotype data and build the IBD graph, which is computationally feasible. Our approach is also fast compared to previous IBD association testing methods. After we build the IBD graph, our method has the same computation time as traditional GWAS approaches. Our method connects IBD mapping to linear models. This permits analysis of the statistical power of IBD mapping, which will depend on the effect sizes of the underlying variants and the genetic structure in terms of the relatedness between individuals in the samples. This type of analysis may motivate the development of novel IBD mapping statistics which have higher statistical power than approaches currently being used. We expect that our new method will promote the wide use of IBD association testing and facilitate further research on the power and utility of IBD association testing.

CHAPTER 7

Conclusion

In this dissertation, I presented several methods for detecting and analyzing genetic variants. Genetic variants range from single nucleotide polymorphisms (SNPs) to chromosomal structural variations (SVs). High-throughput Sequencing(HTS) technologies provide great opportunities for both detecting the genetic variants and uncover genetic basis of complex traits and diseases. Although the sequencing cost has decreased dramatically with the development of HTS technologies, it is still infeasible to sequence a large number of individuals in a study due to budget constraints.

I first proposed a strategy to sequence many individuals simultaneously using overlapping pools. Under this strategy, multiple individuals are grouped into one pool and are sequenced together. The cost is reduced because only one sample preparation is necessary per pool. In chapter 2, I presented an approach to recover the genotype of all individuals accurately.

Structure variations, especially CNVs, play an important role in many complex diseases and traits. In chapter 3, I proposed a statistical model to detect the boundaries and copy numbers of CNVs. This method utilized read mapping uncertainty where a read can be mapped to multiple positions in the reference genome. It is the first attempt to predict CNVs at nucleotide resolution, and the first to utilize uncertainty of read mapping. I further extended this approach to detect CNVs from tumor genomes. The challenge of detecting CNVs in tumor genomes lies in the fact that tumor samples are easily contaminated by normal

stromal cells in the sample preparation step. I proposed a method to estimate the contamination rate and incorporated it into the statistical model. In chapter 4, I showed that this method can estimate the contamination rate precisely and we can detect CNVs in tumor genomes with high accuracy.

For some complex diseases, SNPs may also influence the disease through interactions. In an extreme scenario, two SNPs may not have any effect on a disease independently, but they may affect the disease when both are present. The detection of SNP interaction is a great computational challenge since we have to consider all possible pairs of SNPs. I designed a two-stage model to reduce the computational time greatly in chapter 5, and prove that some SNPs do not need to be considered for combinations with other SNPs. This approach achieved 63 times speed up while maintaining 99% of the power of the brute force approach.

GWAS has identified many significant common SNPs associated with diseases and traits. However, rare variants will not be identified in traditional GWAS. Rare causal variants are likely to have been introduced into a population recently and are likely to be in shared Identity-By-Descent (IBD) segments. If the segmental IBD haplotype contains the disease causing mutation, then the individuals who share this particular IBD segment are likely to share the disease as well. In chapter 6, I proposed a new test statistic to detect IBD segments associated with quantitative traits, and made a connection between the proposed statistic and linear models so that it does not require permutations to assess the significance of an association. In addition, the method can control for population structure by utilizing linear mixed models. I applied the method to the 1966 North Finland Birth Cohort (NFBC66) and demonstrated that our method could control for populations structure. Also simulations proved the equivalence between the linear model and permutation test.

References

[AHK00] David Altshuler, Joel N. Hirschhorn, Mia. Klannemark, Cecilla M. Lindgren, Marie-Claude Vohl, James Nemesh, Charles R. Lane, Stephen F. Schaffner, Stacey Bolk, Carl Brewer, Tiinamaija Tuomi, Daniel Gaudet, Thomas J. Hudson, Mark Daly, Leif Groop, and Eric S. Lander. "The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes." *Nature genetics*, **26**(1):76–80, 2000.

[AKM09] Can Alkan, Jeffrey M. Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O. Kitzman, Carl Baker, Maika Malig, Onur Mutlu, S Cenk Sahinalp, Richard A. Gibbs, and Evan E. Eichler. "Personalized copy number and segmental duplication maps using next-generation sequencing." *Nature Genetics*, **41**(10):1061–1067, Oct 2009.

[AUS11] Alexej Abyzov, Alexander E. Urban, Michael Snyder, and Mark Gerstein. "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing." *Genome Research*, **21**(6):974–984, Jun 2011.

[Ban10] Vikas Bansal. "A statistical method for the detection of variants from next-generation resequencing of DNA pools." *Bioinformatics*, **26**(12):i318–i324, Jun 2010.

[BB10] Sharon R. Browning and Brian L. Browning. "High-resolution detection of identity by descent in unrelated individuals." *Am J Hum Genet*, **86**(4):526–39, 4 2010.

[BB11] Brian L. Browning and Sharon R. Browning. "A fast, powerful method for detecting identity by descent." *Am J Hum Genet*, **88**(2):173–82, 2 2011.

[BB13] Sharon R Browning and Brian L Browning. "Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort." *Human genetics*, **132**(2):129–138, 2013.

[BKK94] Rogier M. Bertina, Bobby P. C. Koeleman, Ted Koster, Frits R. Rosendaal, Richard J. Dirven, Hans de Ronde, Pieter A. Van Der Velden, and Pieter H. Reitsma. "Mutation in blood coagulation factor V associated with resistance to activated protein C." *Nature*, **369**(6475):64–67, 1994.

[BSW05] Rachel B. Brem, John D. Storey, Jacqueline Whittle, and Leonid Kruglyak. "Genetic interactions between polymorphisms that affect gene expression in yeast." *Nature*, **436**(7051):701–3, 8 2005.

[BT12] Sharon R. Browning and Elizabeth A. Thompson. "Detecting Rare Variant Associations by Identity-by-Descent Mapping in Case-Control Studies." *Genetics*, **190**(4):1521–31, 4 2012.

[BTL11] Vikas Bansal, Ryan Tewhey, Emily M Leproust, and Nicholas J Schork. "Efficient and cost effective population resequencing by pooling and in-solution hybridization." *PLoS One*, **6**(3):e18353, Mar 2011.

[Car07] Nigel P. Carter. "Methods and strategies for analyzing copy number variation using DNA microarrays." *Nature Genetics*, **39**(7 Suppl):S16–S21, Jul 2007.

[CGJ09] Derek Y. Chiang, Gad Getz, David B. Jaffe, Michael J T. O'Kelly, Xiaojun Zhao, Scott L. Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S. Lander. "High-resolution mapping of copy-number alterations with massively parallel sequencing." *Nature Methods*, **6**(1):99–103, Jan 2009.

[CHR05] Federico Cappuzzo, Fred R. Hirsch, Elisa Rossi, Stefania Bartolini, Giovanni L. Ceresoli, Lynne Bemis, Jerry Haney, Samir Witta, Kathleen Danenberg, Irene Domenichini, Vienna Ludovini, Elisabetta Magrini, Vanesa Gregorc, Claudio Doglioni, Angelo Sidoni, Maurizio Tonato, Wilbur A. Franklin, Lucio Crino, Paul A Bunn, Jr, and Marileila Varella-Garcia. "Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer." *Journal of National Cancer Institute*, **97**(9):643–655, May 2005.

[CLC08] Peng-An Chen, Hsiao-Fei Liu, and Kun-Mao Chao. "CNVDetector: locating copy number variations using array CGH data." *Bioinformatics*, **24**(23):2773–2775, Dec 2008.

[CM02] Dorin Comaniciu and Peter Meer. "Mean shift: A robust approach toward feature space analysis." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**:603–619, May 2002.

[Con01] International Human Genome Sequencing Consortium. "Initial sequencing and analysis of the human genome." *Nature*, **409**(6822):860–921, 02 2001.

[Con07] Wellcome Trust Case Control Consortium. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature*, **447**(7145):661–78, 6 2007.

[Con10] 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing." *Nature*, **467**(7319):1061–1073, Oct 2010.

[CRT05] Emmanuel Candes, Justin Romberg, and Terence Tao. "Stable Signal Recovery from Incomplete and Inaccurate Measurements." *Communications on Pure and Applied Mathematics*, **59**(19):e1207–1223, Oct 2005.

[CSS93] E. H. Corder, A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. W. Small, A. D. Roses, J. L. Haines, and M. A. Pericak-Vance. "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families." *Science*, **261**(5123):921–3, 8 1993.

[DR99] B. Devlin and K. Roeder. "Genomic control for association studies." *Biometrics*, **55**(4):997–1004, 12 1999.

[EFG10] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and et al. "Missing heritability and strategies for finding the underlying causes of complex disease." *Nat Rev Genet*, **11**(6):446–450, Jun 2010.

[EMM06] David M. Evans, Jonathan Marchini, Andrew P. Morris, and Lon R. Cardon. "Two-stage two-locus models in genome-wide association." *PLoS Genet*, **2**(9):e157, 9 2006.

[GER12] David Golan, Yaniv Erlich, and Saharon Rosset. "Weighted pooling–practical and cost-effective techniques for pooled high-throughput sequencing." *Bioinformatics*, **28**(12):i197–i206, Jun 2012.

[GKL11] Alexander Gusev, Eimear E. Kenny, Jennifer K. Lowe, Jaqueline Salit, Richa Saxena, Sekar Kathiresan, David M. Altshuler, Jeffrey M. Friedman, Jan L. Breslow, and Itsik Pe'er. "DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation." *Am J Hum Genet*, **88**(6):706–17, 6 2011.

[GLS09] Alexander Gusev, Jennifer K. Lowe, Markus Stoffel, Mark J. Daly, David Altshuler, Jan L. Breslow, Jeffrey M. Friedman, and Itsik Pe'er. "Whole population, genome-wide mapping of hidden relatedness." *Genome Res*, **19**(2):318–26, 2 2009.

[GSL05] Kevin L Gunderson, Frank J Steemers, Grace Lee, Leo G Mendoza, and Mark S Chee. "A genome-wide scalable SNP genotyping assay using microarray technology." *Nat Genet*, **37**(5):549–554, May 2005.

[HAE09] Fereydoun Hormozdiari, Can Alkan, Evan E. Eichler, and S Cenk Sahinalp. "Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes." *Genome Research*, **19**(7):1270–1278, Jul 2009.

[HDM09] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." *PLoS Genet*, **5**(6):e1000529, Jun 2009.

[HFE10] Dan He, Nicholas Furlotte, and Eleazar Eskin. "Detection and reconstruction of tandemly organized de novo copy number variations." *BMC Bioinformatics*, **11 Suppl 11**:S12, Dec 2010.

[HH06] Eran Halperin and Elad Hazan. "HAPLOFREQ-estimating haplotype frequencies efficiently." *Journal of Computational Biology*, **13**(2):481–500, Mar 2006.

[HHA10] Faraz Hach, Fereydoun Hormozdiari, Can Alkan, Farhad Hormozdiari, Inanc Birol, Evan E. Eichler, and S Cenk Sahinalp. "mrsFAST: a cache-oblivious algorithm for short-read mapping." *Nature Methods*, **7**(8):576–577, Aug 2010.

[HHF11] Dan He, Farhad Hormozdiari, Nicholas Furlotte, and Eleazar Eskin. "Efficient algorithms for tandem copy number variation reconstruction in repeat-rich regions." *Bioinformatics*, **27**(11):1513–1520, Jun 2011.

[HHS08] Iman Hajirasouliha, Fereydoun Hormozdiari, S. Cenk Sahinalp, and Inanc Birol. "Optimal pooling for genome re-sequencing with ultra-high-throughput short-read technologies." *Bioinformatics*, **24**(13):i32–i40, Jul 2008.

[HKE09] Buhm Han, Hyun Min Kang, and Eleazar Eskin. "Rapid and Accurate Multiple Testing Correction and Power Estimation for Millions of Correlated Markers." *PLoS Genet*, **5**:e1000456, 04 2009.

[HKR13] Buhm Han, Eun Yong Kang, Soumya Raychaudhuri, Paul I. W. de Bakker, and Eleazar Eskin. "Fast Pairwise IBD Association Testing in Genome-wide Association Studies." *Bioinformatics*, 2013.

[HZP11] Dan He, Noah Zaitlen, Bogdan Pasaniuc, Eleazar Eskin, and Eran Halperin. "Genotyping common and rare variation using overlapping pool sequencing." *BMC Bioinformatics*, **12 Suppl 6**:S2, Jul 2011.

[IFR04] A John Iafrate, Lars Feuk, Miguel N. Rivera, Marc L. Listewnik, Patricia K. Donahoe, Ying Qi, Stephen W. Scherer, and Charles Lee. "Detection of large-scale variation in the human genome." *Nature Genetics*, **36**(9):949–951, Sep 2004.

[JAS12] Thorlakur Jonsson, Jasvinder K Atwal, Stacy Steinberg, Jon Snaedal, Palmi V Jonsson, Sigurbjorn Bjornsson, Hreinn Stefansson, Patrick Sulem, Daniel Gudbjartsson, Janice Maloney, et al. "A mutation in APP protects against Alzheimer/'s disease and age-related cognitive decline." *Nature*, **488**(7409):96–99, 2012.

[KE13] Emrah Kostem and Eleazar Eskin. "Efficiently Identifying Significant Associations in Genome-wide Association Studies." *J Comput Biol*, 9 2013.

[KLE11] Emrah Kostem, Jose A. Lozano, and Eleazar Eskin. "Increasing Power of Genome-wide Association Studies by Collecting Additional SNPs." *Genetics*, **188**(2):449–60, 4 2011.

[KMF08] Augustine Kong, Gisli Masson, Michael L Frigge, Arnaldur Gylfason, Pasha Zusmanovich, Gudmar Thorleifsson, Pall I Olason, Andres Ingason, Stacy Steinberg, Thorunn Rafnar, et al. "Detection of sharing by descent, long-range phasing and haplotype imputation." *Nature genetics*, **40**(9):1068–1075, 2008.

[KSS10] Hyun Min Kang, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-Yee Y. Kong, Nelson B. Freimer, Chiara Sabatti, and Eleazar Eskin. "Variance component model to account for sample structure in genome-wide association studies." *Nat Genet*, **42**(4):348–54, 4 2010.

[KZW08] Hyun Min Kang, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly, and Eleazar Eskin. "Efficient control of population structure in model organism association mapping." *Genetics*, **178**(3):1709, 2008.

[LCY11] Joon Sang Lee, Murim Choi, Xiting Yan, Richard P Lifton, and Hongyu Zhao. "On optimal pooling designs to identify rare variants through massive resequencing." *Genet Epidemiol*, **35**(3):139–147, Apr 2011.

[LHC04] K. Ljungberg, S. Holmgren, and O. Carlborg. "Simultaneous search for multiple QTL using the global optimization algorithm DIRECT." *Bioinformatics*, **20**(12):1887–95, 8 2004.

[Lin05] Danyu Lin. "An efficient Monte Carlo approach to assessing statistical significance in genomic studies." *Bioinformatics*, **21**(6):781–7, 3 2005.

[LLK13] Jennifer Listgarten, Christoph Lippert, Eun Yong Kang, Jing Xiang, Carl M. Kadie, and David Heckerman. "A powerful and efficient set test for genetic markers that handles confounders." *Bioinformatics*, 4 2013.

[LLL11] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman. "FaST linear mixed models for genome-wide association studies." *Nat Methods*, **8**(10):833–5, 2011.

[LLS10] Zongzhi Liu, Ao Li, Vincent Schulz, Min Chen, and David Tuck. "MixHMM: Inferring Copy Number Variation and Allelic Imbalance Using SNP Arrays and Tumor Samples Mixed with Stromal Cells." *PLoS ONE*, **5**(6):e10909, 06 2010.

[LTP09] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." *Genome Biology*, **10**(3):R25, Mar 2009.

[Mar08] Elaine R Mardis. "The impact of next-generation sequencing technology on genetics." *Trends Genet*, **24**(3):133–141, Mar 2008.

[MC09] Teri A Manolio and Francis S Collins. "The HapMap and genome-wide association studies in diagnosis and therapy." *Annu Rev Med*, **60**:443–456, Feb 2009.

[MCC09] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, and et al. "Finding the missing heritability of complex diseases." *Nature*, **461**(7265):747–753, Oct 2009.

[MCG06] Joshua Millstein, David V. Conti, Frank D. Gilliland, and W. James Gauderman. "A testing framework for identifying susceptibility genes in the presence of epistasis." *The American Journal of Human Genetics*, **78**(1):15–27, 2006.

[MDL04] Hajime Matsuzaki, Shoulian Dong, Halina Loi, Xiaojun Di, Guoying Liu, Earl Hubbell, and et al. "Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays." *Nat Methods*, **1**(2):109–111, Nov 2004.

[Met08] Michael L. Metzker. "Sequencing technologies the next generation." *Nature Reviews Genetics*, **11**(12):31–46, Jan 2008.

[MFD10] Paul Medvedev, Marc Fiume, Misko Dzamba, Tim Smith, and Michael Brudno. "Detecting copy number variation with mated short reads." *Genome Research*, **20**(11):1613–1622, Nov 2010.

[MHM07] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. "A new multipoint method for genome-wide association studies by imputation of genotypes." *Nat Genet*, **39**(7):906–913, Jul 2007.

[PHT11] Alkes L Price, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson. "Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals." *PLoS genetics*, **7**(2):e1001317, 2011.

[PNT07] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. "PLINK: a tool set for whole-genome association and population-based linkage analyses." *Am J Hum Genet*, **81**(3):559–75, 9 2007.

[PP09] Snehit Prabhu and Itsik Pe'er. "Overlapping pools for high-throughput targeted resequencing." *Genome Res*, **19**(7):1254–1261, Jul 2009.

[PP12] Snehit Prabhu and Itsik Pe'er. "Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease." *Genome Research*, **22(11)**:2230–2240, 2012.

[PPP06] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. "Principal components analysis corrects for stratification in genome-wide association studies." *Nat Genet*, **38**(8):904–909, 08 2006.

[10.1038/ng1847.]

[RIF06] Richard Redon, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T Daniel Andrews, Heike Fiegler, Michael H. Shapero, Andrew R. Carson, Wenwei Chen, Eun Kyung Cho, Stephanie Dallaire, Jennifer L. Freeman, Juan R. Gonzlez, Mnica Gratacs, Jing Huang, Dimitrios Kalaitzopoulos, Daisuke Komura, Jeffrey R. MacDonald, Christian R. Marshall, Rui Mei, Lyndal Montgomery, Kunihiro Nishimura, Kohji Okamura, Fan Shen, Martin J. Somerville, Joelle Tchinda, Armand Valsesia, Cara Woodwark, Fengtang Yang, Junjun Zhang, Tatiana Zerjal, Jane Zhang, Lluis Armengol, Donald F. Conrad, Xavier Estivill, Chris Tyler-Smith, Nigel P. Carter, Hiroyuki Aburatani, Charles Lee, Keith W. Jones, Stephen W. Scherer, and Matthew E. Hurles. "Global variation in copy number in the human genome." *Nature*, **444**(7118):444–454, Nov 2006.

[SAZ10] Noam Shental, Amnon Amir, and Or Zuk. "Identification of rare alleles and their carriers using compressed se(que)nsing." *Nucleic Acids Res*, **38**(19):e179, Oct 2010.

[SJ08] Jay Shendure and Hanlee Ji. "Next-generation DNA sequencing." *Nature Biotechnology*, **26**(12):1135–1145, Oct 2008.

[SKA10] Peter H. Sudmant, Jacob O. Kitzman, Francesca Antonacci, Can Alkan, Maika Malig, Anya Tsalenko, Nick Sampas, Laurakay Bruhn, Jay Shendure, 1000 Genomes Project , and Evan E. Eichler. "Diversity of human copy number variation and multicopy genes." *Science*, **330**(6004):641–646, Oct 2010.

[SLM07] Jonathan Sebat, B. Lakshmi, Dheeraj Malhotra, Jennifer Troge, Christa Lese-Martin, Tom Walsh, Boris Yamrom, Seungtai Yoon, Alex Krasnitz, Jude Kendall, Anthony Leotta, Deepa Pai, Ray Zhang, Yoon-Ha Lee, James Hicks, Sarah J. Spence, Annette T. Lee, Kaija Puura, Terho Lehtimki, David Ledbetter, Peter K. Gregersen, Joel Bregman, James S. Sutcliffe, Vaidehi Jobanputra, Wendy Chung, Dorothy Warburton, Mary-Claire King, David Skuse, Daniel H. Geschwind, T Conrad Gilliam, Kenny Ye, and Michael Wigler. "Strong association of de novo copy number mutations with autism." *Science*, **316**(5823):445–449, Apr 2007.

[SM05] S.R Seaman and B Muller-Myhsok. "Rapid Simulation of P Values for Product Methods and Multiple-Testing Adjustment in Association Studies." *American journal of human genetics*, **76(3)**:399 – 408, 2005.

[SMA10] Jared T. Simpson, Rebecca E. McIntyre, David J. Adams, and Richard Durbin. "Copy number variant detection in inbred strains from short read sequence data." *Bioinformatics*, **26**(4):565–567, Feb 2010.

[SSS09] Ian Sudbery, Jim Stalker, Jared T. Simpson, Thomas Keane, Alistair G. Rust, Matthew E. Hurles, Klaudia Walter, Dee Lynch, Lydia Teboul, Steve D. Brown, Heng Li, Zemin Ning, Joseph H. Nadeau, Colleen M. Croniger, Richard Durbin, and David J. Adams. "Deep short-read sequencing of chromosome 17 from the mouse strains A/J and CAST/Ei identifies significant germline variation and candidate genes that regulate liver triglyceride levels." *Genome Biology*, **10**(10):R112, Oct 2009.

[SVL07] Richa Saxena, Benjamin F. Voight, Valeriya Lyssenko, Noël P. Burtt, Paul I. W. de Bakker, Hong Chen, Jeffrey J. Roix, Sekar Kathiresan, Joel N. Hirschhorn, Mark J. Daly, Thomas E. Hughes, Leif Groop, David Altshuler, Peter Almgren, Jose C. Florez, Joanne Meyer, Kristin Ardlie, Kristina Bengtsson Bostrőm, Bo Isomaa, Guillaume Lettre, Ulf Lindblad, Helen N. Lyon, Olle Melander, Christopher Newton-Cheh, Peter Nilsson, Marju Orho-Melander, Lennart Rastam, Elizabeth K. Speliotes, Marja-Riitta R. Taskinen, Tiinamaija Tuomi, Candace Guiducci, Anna Berglund, Joyce Carlson, Lauren Gianniny, Rachel Hackett, Liselotte Hall, Johan Holmkvist, Esa Laurila, Marketa Sjőgren, Maria Sterner, Aarti Surti, Margareta Svensson, Malin Svensson, Ryan Tewhey, Brendan Blumenstiel, Melissa Parkin, Matthew

Defelice, Rachel Barry, Wendy Brodeur, Jody Camarata, Nancy Chia, Mary Fava, John Gibbons, Bob Handsaker, Claire Healy, Kieu Nguyen, Casey Gates, Carrie Sougnez, Diane Gage, Marcia Nizzari, Stacey B. Gabriel, Gung-Wei W. Chirn, Qicheng Ma, Hemang Parikh, Delwood Richardson, Darrell Ricke, and Shaun Purcell. "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels." *Science*, **316**(5829):1331–6, 6 2007.

[SXB10] Thierry Schpbach, Ioannis Xenarios, Sven Bergmann, and Karen Kapur. "FastEpistasis: a high performance computing solution for quantitative trait epistasis." *Bioinformatics*, **26**(11):1468–1469, 2010.

[TSB05] Eray Tuzun, Andrew J. Sharp, Jeffrey A. Bailey, Rajinder Kaul, V Anne Morrison, Lisa M. Pertz, Eric Haugen, Hillary Hayden, Donna Albertson, Daniel Pinkel, Maynard V. Olson, and Evan E. Eichler. "Fine-scale structural variation of the human genome." *Nature Genetics*, **37**(7):727–732, Jul 2005.

[WAP00] Scott M. Williams, Jonathan H. Addy, John A. Phillips, Min Dai, John Kpodonu, James Afful, Harold Jackson, Karen Joseph, Felicia Eason, Mark M. Murray, Pamela Epperson, Adwoa Aduonum, Lee-Jun Wong, Pedro A. Jose, and Robin A. Felder. "Combinations of variations in multiple genes are associated with hypertension." *Hypertension*, **36**(1):2–6, 7 2000.

[WKE10] Michael C. Wu, Peter Kraft, Michael P. Epstein, Deanne M. Taylor, Stephen J. Chanock, David J. Hunter, and Xihong Lin. "Powerful SNP-set analysis for case-control genome-wide association studies." *Am J Hum Genet*, **86**(6):929–42, 6 2010.

[WLC11] Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. "Rare-variant association testing for sequencing data with the sequence kernel association test." *Am J Hum Genet*, **89**(1):82–93, 7 2011.

[XCQ10] Wan Xiang, Yang Can, Yang Qiang, Xue Hong, Fan Xiaodan, Tang Nelson, and Yu Weichuan. "BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies." *The American Journal of Human Genetics*, **87**:325–340, 2010.

[YHW09] Can Yang, Zengyou He, Xiang Wan, Qiang Yang, Hong Xue, and Weichuan Yu. "SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies." *Bioinformatics*, **25**(4):504–11, 2 2009.

[YIS04] E. D. Yanchina, T. V. Ivchik, E. I. Shvarts, A. N. Kokosov, and N. E. Khodzhayantz. "Gene-gene interactions between glutathione-s transferase M1 and matrix metalloproteinase 9 in the formation of hereditary predisposition to chronic obstructive pulmonary disease." *Bulletin of Experimental Biology and Medicine*, **137**(1):64–66, 2004.

[YXM09] Seungtai Yoon, Zhenyu Xuan, Vladimir Makarov, Kenny Ye, and Jonathan Sebat. "Sensitive and accurate detection of copy number variants using read depth of coverage." *Genome Research*, **19**(9):1586–1592, Sep 2009.

[ZBG12] Yuan Zhu, Alan O Bergland, Josefa Gonzlez, and Dmitri A Petrov. "Empirical validation of pooled whole genome population re-sequencing in Drosophila melanogaster." *PLoS One*, **7**(7):e41901, Jul 2012.

[Zha10] Nancy R. Zhang. "DNA copy number profiling in normal and tumor genomes." *Frontiers in Computational and Systems Biology*, **15**:259–281, 2010.

[ZHS12] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. "The mystery of missing heritability: Genetic interactions create phantom heritability." *Proceedings of the National Academy of Sciences*, **109**(4):1193–1198, 2012.

[ZHZ10] Xiang Zhang, Shunping Huang, Fei Zou, and Wei Wang. "TEAM: efficient two-locus epistasis tests in human genome-wide association study." *Bioinformatics*, **26**(12):i217–i227, 2010.

[ZPX09] Xiang Zhang, Feng Pan, Yuying Xie, Fei Zou, and Wei Wang. *COE: A General Approach for Efficient Genome-Wide Two-Locus Epistasis Test in Disease Association Study*, pp. 253–269. Springer Berlin Heidelberg, 2009.

[ZS12] Xiang Zhou and Matthew Stephens. "Genome-wide efficient mixed-model analysis for association studies." *Nat Genet*, **44**(7):821–4, 2012.