A Kernel based Study of the Association between Copy Number Variants and Disease-related Traits

by
Nastaran Maus Esfahani

Thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

under the supervision of
Professor Paul Kennedy
and
Professor Daniel Catchpoole

University of Technology Sydney
Faculty of Engineering and Information Technology
October 2022

Certificate of Original Authorship

I, Nastaran Maus Esfahani, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note: Signature removed prior to publication.

[Nastaran Maus Esfahani]

DATE: 16th July, 2023

PLACE: Sydney, Australia

Acknowledgements

To my mother Nasrin, my father Mahdi and my only brother Saeed, who gave me their endless emotional support during my Ph.D. journey.

To my supervisor Prof Paul Kennedy who taught me the problem-solving skill that I benefit from it not only in my study but also in my personal life.

To my co-supervisor A/Prof Daniel Catchpoole, for his positive attitude and encouragement whenever things got too difficult for me to handle.

To my friends who were always there for me no matter how far away they were from me.

To myself for not giving up on my Ph.D. journey for what happened to me at the beginning stages, which should not happen to any woman anywhere in the world, and starting from square one to achieve one of my life goals.

Contents

List of Figures

Figure **Page**

List of Tables

Table **Page**

List of Publications

Listed below are the publications and other outputs associated with the research presented in this thesis.

Maus Esfahani, N., Catchpoole, D., Khan, J., & Kennedy, P. J. (2021). MCKAT: a multi-dimensional copy number variant kernel association test. BMC bioinformatics, 22(1), 1-16. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04494-w

Maus Esfahani, N., Catchpoole, D., & Kennedy, P. J. (2021). SMCKAT, a Sequential Multi-Dimensional CNV Kernel-Based Association Test. Life, 11(12), 1302. https://www.mdpi.com/2075-1729/11/12/1302

List of Abbreviations and Symbols

Abbreviation	Description
AS-PCR	allele-specific PCR
ARMS	alveolar rhabdomyosarcoma
ASD	autism spectrum disorder
CAPS	cleaved amplified polymorphic sequence
CCRET	CNV collapsing random effects test
CKAT	CNV kernel association test
CNVRs	CNV regions
CAST	collapsing and summation test
cAUC	common area under the curve
CXN	copy number
CNP	copy number polymorphisms
CONCUR	copy number profile curve-based association test
CNV	copy number variation
dCAPs	derived CAPS
DR	difference from the Reference
ERMS	embryonal rhabdomyosarcoma
FWER	family-wise error rate
MCKAT	multi-dimensional CNV kernel-based association test
NIH	national Institute of Health
PCR	polymerase chain reaction
Q-Q	quantile-quantile
ROI	region of interest
RMS	rhabdomyosarcoma
SMCKAT	sequential multi-dimensional copy number variant kernel association test
SNP	single nucleotide polymorphism
VOUS	Variants of uncertain significance
WS	weighted-sum

Symbol	Description
X	a copy number variant
$X^{(1)}$	start chromosomal position
$X^{(2)}$	end chromosomal position
$X^{(3)}$	CNV type
$X^{(4)}$	CNV dosage
K	similarity matrix
K_{ij}	similarity between CNV profile i and j
K_s	single-pair CNV Kernel
R_i	list of a subject's CNV
K_w	whole Chromosome CNV Kernel
y_i	status of the phenotype
Z	covariant matrix
G	CNV group
K_{PG}	pair CNV Group Kernel
P	CNV group series
K_{WG}	Whole Genome CNV Group Kernel

Abstract

Copy number variants (CNVs) are the most common form of structural genetic variation, reflecting the gain or loss of DNA segments compared with a reference genome. Studies have shown that CNVs are linked to various disorders like autism, intellectual disability, and schizophrenia. Consequently, the interest in studying a possible association of CNVs to specific disease traits is growing. However, due to the specific multi-dimensional characteristics of the CNVs, methods for testing the association between CNVs and the disease-related traits are still few and underdeveloped. The research presented in this thesis addresses several aspects of research on the association between CNVs and disease related traits , and the broader concepts of the association between CNV sequential order with adverse phenotype, and the association of the CNV and other genetic variation interactions with disease related traits.

This work makes three contributions to knowledge, relating to the significance of CNVs on some chromosomal regions in association with disease related traits. Contribution 1 proposed a multi-dimensional CNV kernel based association test (MCKAT). MCKAT performs better than the state of the art methods and was evaluatd on both simulated and real data. MCKAT can identify chromosomal regions at cytogenetic band level containing CNVs that are significantly associated with disease related traits. MCKAT considers all CNV characteristics in testing the association and can provide strong evidence, small p-values, to accept or reject the association hypothesis. MCKAT is applicable to both frequent and rare CNV data sets.

Contribution 2 is a sequential multi-dimensional CNV kernel based association test (SMCKAT). SMCKAT tests the association between the CNV sequential order and disease related traits. SMCKAT considers not only the CNV characteristics but the CNV sequential order. SMKAT can identify the chromosomal regions that the CNV sequential order is significantly associated with disease related traits.

Based on our knowledge, SMCKAT is the first such method to test the association between the CNV sequential order and disease related traits.

Contribution 3 uses our proposed method to demonstrate that considering the CNV-gene intersection along with the CNV characteristics in testing the association between CNVs and disease related traits is informative and can provide more insights about the disease development. This is because CNVs can affect their intersected genes in different way like changing the gene expression or disturbing their function. Our proposed methods can be used not only in testing the association between the dual effect of CNVs and their intersected with disease related traits but any other genetic variations based on data availability.

Overall, this research confirms that having association tests, specific to CNVs and compatible with CNV characteristics, to identify the chromosomal regions which contain CNVs that are significantly associated with disease related traits are of biological significance. This work provides methods that can help biologists to identify CNV hot spots associated with disease related traits without doing extensive investigations at the individual level to find significant CNVs. The results of these methods may also provide them with a better understanding of how the interaction between CNVs and other genetic variations like genes can have association with a disease related traits.

Chapter 1

Introduction

1.1 Background

One of the many significant discoveries made after the Human Genome Project was the identification of single nucleotide polymorphisms (SNPs) as a significant source of genetic variation (Hood and Rowen 2013). This led to the hypothesis that single base changes cause the majority of phenotypic variability in human populations. Consequently, intensive efforts were made to develop high-throughput sequencing, genotyping platforms and SNP databases. Genetic research frequently employs a number of conventional SNP genotyping techniques based on electrophoresis systems, including CAPS (cleaved amplified polymorphic sequence), dCAPs (derived CAPS), and AS-PCR (allele-specific PCR) (Zhang et al. 2021). The development of some high-throughput SNP genotyping techniques, like the Gene Chip microarray or the KASP platform, has been made easier by the ongoing advancements in high-throughput sequencing (competitive allele-specific PCR) (Semagn et al. 2014). However, meeting the rapidly rising demand for SNP genotyping is challenging (Wang et al. 2015).

Until recently, the vast majority of gene-mapping studies had primarily focused on the role of SNPs in human diseases like (Visscher et al. 2021; Degtyareva et al. 2021; Rehman et al. 2022). In fact, dozens of SNP-based susceptibility variants for human diseases have been identified using population-based studies to identify genetic determinants of common disease.

However, research over the past ten years have led to a greater understanding of the importance of structural genetic variation in modulating gene expression

and disease phenotype. Copy number variation (CNV) is the most common form of structural genetic variation (McCarroll and Altshuler 2007). CNV is a general term for a molecular phenomenon in which stretches of DNA sequence are either deleted or amplified. The number of deletions and repeats varies among individuals of the same species and some of these differences may lead to disease related traits.

In addition, the impact of CNVs on other genetic variations may have a significant association with disease related traits. For example, when there is an intersection between a CNV and a gene, the CNV can effect on gene expression and has the potential to disrupt the gene structure and function. This suggests that the dual effect of the CNVs and other genomic activities like gene disruption, is very likely to have a significant association with disease related traits.

The first reported association of a CNV with a phenotype was in a non-human species, a reduced-eye mutant *Drosophila melanogaster*, with the bar eyes phenotype resulting from a single duplication of the *bar gene* (Pös et al. 2021). Reports of microscopically visible chromosomal aberrations in the human genome first released after the general human karyotype was established in the 1960s (Tijo and Levan 2004). Then, cytogeneticists discovered the genetic basis for many diseases, such as *Cri-du-Chat* syndrome linked to a partial deletion of chromosome 5's short arm, as early as 1963 (Cerruti Mainardi 2006).

As is reported by Pös et al. (2021), the biological roles of the CNVs range from seemingly no effect on common variability of physiological traits (Zhang et al. 2009), altered metabolic states (Elder et al. 2018), susceptibility to infectious diseases (Gonzalez et al. 2005; Harteveld and Higgs 2010), and host-microbiome interactions (Mohajeri et al. 2018; Poole et al. 2019; Greenblum et al. 2015), to a substantial contribution to common and rare genetic disorders or syndromes (Radvanszky et al. 2013).

In addition to their biological roles, CNV presence in our genomes may have a number of technical implications for bio-medicine. CNVs may act as bio-markers for specific pathological processes like cancer, as bio-markers of environmental exposures like radiation (Arlt et al. 2014), or even as potential confounding variables when analysing the results of specific genetic diagnostic tests (Kubiritova et al. 2019).

Considering the CNV biological roles and technical implications, they are

expected to have a significant impact on screening, diagnosis, prognosis, and monitoring of several disorders. However, despite of the development of better CNV genotyping platforms, we are still in the early stages of incorporating CNVs in genome-wide association studies. In contrast to the well-developed resources available for SNP-association studies, the methods for studying CNVs are still few and underdeveloped. On one hand, the existing methods for studying SNPs are not applicable to CNVs due to their multi-dimensional characteristics including chromosomal position, type, dosage and heterogeneity effect. On the other hand, a few existing methods that have been designed specifically to study CNVs, fall short in dealing with CNV characteristics. Furthermore, the association between the dual effect of CNVs and other genetic variations with disease related traits has not been studied yet based on our knowledge. Therefore, development of novel technical and statistical methods to optimally study CNVs will be necessary.

The research detailed in this thesis was aimed at developing statistical association tests to study the association between CNVs and disease related traits. The remainder of this chapter provides an introduction to the research questions considered (Section 1.2) and the resulting contributions to knowledge (Section 1.3). Finally, Section 1.4 outlines the structure of this thesis.

1.2 Research questions

As mentioned in the previous section, while there are lots of computational association studies that have investigated the association between SNPs and diseases or traits, methods for studying CNVs are underdeveloped due to the multi-dimensional characteristics of the CNVs. Currently, biologists have to do extensive investigation of affected individuals at whole genome level to identify chromosomal regions that CNVs in them are significantly associated with disease related traits. The examples are available in Table 2.1. There is also one more open question around the association between CNVs' sequential order and disease related traits. Specifically, SNPs do not usually function individually, rather, they work in coordination with other SNPs to manifest a disease or trait. Therefore, many sequence studies have been done to test the association between SNPs and disease or traits. However, the association between the sequential order of CNVs and disease-related traits has not been studied, to our knowledge, and it is still unclear if CNVs function individually or whether they work in coordination with

other CNVs.

Furthermore, another important field that is worth investigating and has not received much attention is the dual effects of CNVs and their intersected genes on disease development. CNVs overlap over 7000 genes, many of which are pivotal in biological pathways and can affect genes in different ways (De Smith et al. 2008). Therefore, investigating the association between CNVs which are intersected with the significant genes to a disease related traits can provide biologists with meaningful insights about the disease development.

These open challenges motivated the following research questions, which this project was designed to address:

RQ1: Can a kernel-based association test, deal with CNV multi-dimensional characteristics and heterogeneity effect to identify CNV chromosomal regions that are significantly associated with disease related traits?

RQ2: Can a sequential kernel-based association test, identify if there is any significant association between CNVs' sequential order and disease related traits?

RQ3: Can considering the effect of CNV-gene intersections in addition to the CNV characteristics, be helpful and informative in testing the association between CNVs and disease related traits?

1.3 Contributions to knowledge

The research presented in this thesis makes three corresponding contributions to knowledge, corresponding to the above research questions:

Contribution 1: A multi-dimensional kernel-based CNV association test that allows for the detection of CNV chromosomal regions significantly associated with disease related traits and improves on currently available methods for studying CNVs.

Contribution 2: A sequential multi-dimensional CNV kernel-based association test that allows investigating whether CNVs are randomly distributed across

the genome, or their order matters and have a significant association with disease related traits.

Contribution 3: The demonstration that considering the effect of CNV-gene intersections in addition to the CNV characteristics is informative and helpful in identifying the significant association between CNVs and disease related traits.

These contributions are briefly described below.

Contribution 1: A multi-dimensional CNV kernel-based association test that allows for the detection of CNV chromosomal regions significantly associated with disease related traits and improves on currently available methods for studying CNVs. The first contribution is a multi-dimensional CNV kernel-based association test for the detection of CNV chromosomal regions significantly associated with disease related traits. The multi-dimensional nature of the test provides an advantage over existing tests for considering all CNV characteristics in the association test. A multi-dimensional kernel framework is capable of measuring the similarity between CNV profiles utilizing CNV chromosomal region, type, dosage and heterogeneity effects. It contains two kernels. The first kernel, the single-pair CNV kernel, measures the similarity between a single CNV pair. The single-pair CNV kernel includes three sub-kernels. Each sub-kernel is responsible for measuring the similarity between two CNVs with respect to one of three CNV characteristics. The second sub-kernel, the whole chromosome kernel, aggregates the similarity between every possible CNV pair to measure the total similarity between the CNV profiles of the subjects. Finally, the association between CNVs across a chromosome and disease-related traits is tested by comparing the similarity in CNV profiles to that in the trait using an association test.

The development and testing of the multi-dimensional CNV kernel-based association test (MCKAT) are described in Chapter 3. The performance of the MCKAT is assessed using the simulated CNV data, frequent CNV data and rare CNV data. The results show that the multi-dimensional kernel-based model provides a CNV association test that is competitive with existing methods. First, MCKAT can identify chromosomal regions, at cytogenetic bands, that CNVs

on them are significantly associated with disease related traits using all CNV characteristics. Secondly, MCKAT is applicable to both CNV types: frequent and rare CNVs. Finally, MCKAT is capable of indicating stronger evidence, lower p-value, in detecting significant associations between CNVs and disease-related traits compared to existing methods, which is crucial in genome-wide association studies for multiple testing and avoiding false positive discoveries.

Contribution 2: A sequential multi-dimensional kernel-based CNV association test that allows investigating whether CNVs are randomly distributed across the genome, or their order matters and have a significant association with disease related traits. The second contribution is a sequential multi-dimensional CNV kernel-based association test to investigate whether CNVs are randomly distributed across the genome or there are any significant association between their orders and disease-related traits of interest. The sequential and multi dimensional nature of the test provides an advantage over existing methods for considering not only all CNV characteristics but CNV orders in the association test. Based on the literature, the association between CNV orders and disease-related traits has not been investigated in neither biological nor computational studies. Therefore, our proposed test is the first such method to test the association between the sequential order of CNVs and disease related traits.

The development and testing of the sequential multi-dimensional CNV kernel-based association test (SMCKAT) are described in Chapter 4. The performance of the SMCKAT is assessed using the simulated CNV data, frequent CNV data and rare CNV data. The results show that SMCKAT is applicable on both frequent and common CNV data and capable of identifying hot-spots on the genome where both CNV characteristics and the CNV sequential order are significantly associated with disease related trait of interest. SMCKAT is capable of handling both type I and type II errors. Using more strict rules in measuring similarity among CNV profiles, both CNV orders and characteristics, SMCKAT provides more specific significant CNV regions compared to existing methods.

Contribution 3: The demonstration that considering the effect of CNV-gene intersections in addition to the CNV characteristics is informative and helpful in testing the significant association between CNVs and

disease related traits. The third contribution is an investigation into the association of the dual effects of the CNVs and CNVs that are intersected with the genes that have been identified important in developing a specific disease. Both simulation and real data application results show that considering CNV-gene intersection data in designing the association test can help to measure the similarity between CNV profiles more precisely by using a CNV-gene intersection kernel. Using this approach can provide us with a statistical association test with higher power and better performance in handling both type I and II errors. This approach can be used to investigate the dual effects of CNVs and other genetic variations like SNPs to achieve more precise insights about the effect of different genetic variations together on the disease development.

1.4 Thesis structure

The remainder of this thesis is structured as follows:

Chapter 2 provides a review of the literature relevant to this project, including an overview of genetic variants and genetic association studies, with a focus on sequence variants and copy number variants, challenges of studying copy number variants, and a detailed review of the existing methods to test the association between copy number variants and disease related traits.

Chapter 3 details the development of a multi-dimensional CNV kernel-based association test and provides a comprehensive comparative assessment of the developed test against existing methods.

Chapter 4 details the development of a sequential multi-dimensional CNV kernel-based association test and provides a comprehensive comparative assessment of the developed test against existing methods.

Chapter 5 presents an investigation into the dual effects of CNV characteristics and the CNV-gene intersections on the disease development.

Chapter 6 concludes the thesis, discussing the implications of the results of this work, limitations and directions for possible future research.

Chapter 2

Literature review

The following sections provide a review of the literature relevant to this project. Section 2.1 reviews the biological domain of genetic variations, DNA sequence variations and copy number variants. Section 2.2 reviews the algorithms developed for testing the association between genetic variants, specifically collapsing methods. Section 2.3 explores the challenges of studying copy number variants. Section 2.4 discusses the existing methods for studying CNVs. Section 2.5 provides a summary of the gaps in the existing research and how they are addressed by the work presented in this thesis.

2.1 Genetic Variations

Genetically speaking, humans are 99.8-99.9 percent the same (Consortium et al. 2015). The remaining 0.1-0.2 percent is what makes them all unique and is called genetic variation. Genetic variation is a term used to describe the difference between human genomes. Different genetic variations contribute to human variability, which still needs to be fully characterized or properly understood. Genetic variation has two primary forms: sequence variations and structural alterations. DNA sequence variations and copy number variants are the most common form of sequence variations and structural alterations in the human genome, respectively (Frazer et al. 2009).

2.1.1 DNA sequence variations

DNA sequence variation or single nucleotide polymorphism (SNP) is the most common form of sequence variations. A SNP represents a difference in a

single DNA building block, called a nucleotide, anywhere in the genome between members of a species or paired chromosomes in an individual. In humans, each cell normally contains 23 pairs of chromosomes. Each chromosome is made up of DNA. A nucleotide is one of the building blocks of DNA, and it consists of one of four chemicals, including adenine, thymine, guanine, or cytosine. For example, one person may have the base pair thymine-adenine at a specific location in a specific chromosome, and another may have the base pair cytosine-guanine in the same location as is shown in Fig. 2.1. This means that there is a SNP in this particular position.

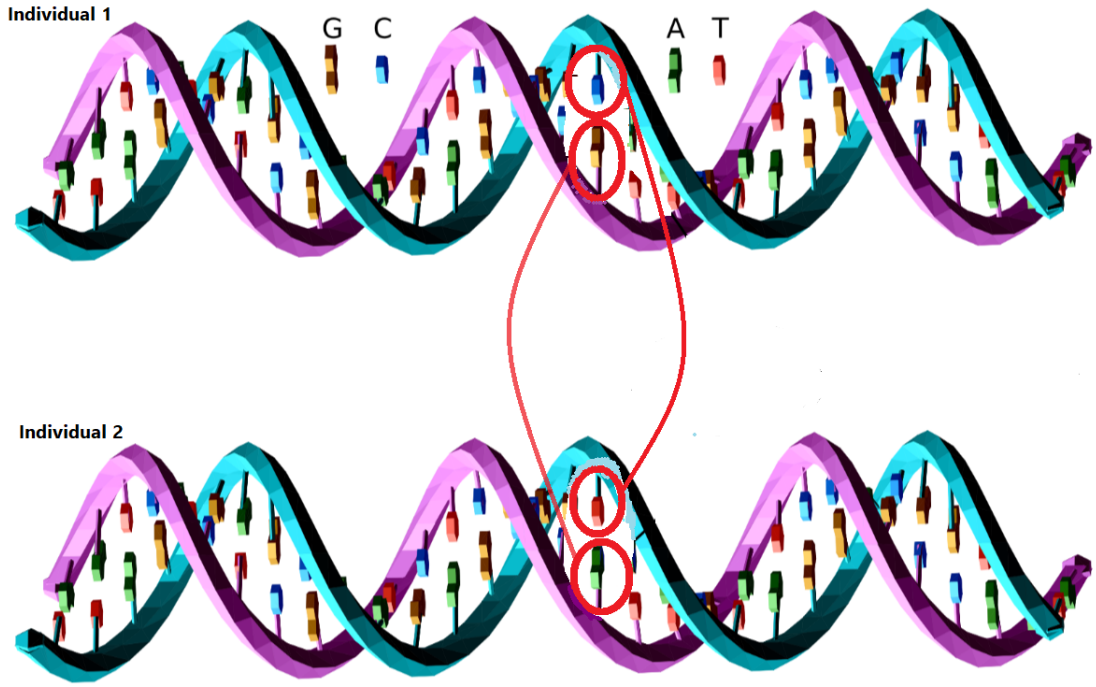

Figure 2.1: Single nucleotide polymorphism. A, T, G and C stand for adenine, thymine, guanine and cytosine respectively.

SNPs are classified into two major types based on the gene region they fall within: coding and non-coding regions. SNPs within a coding region do not always change the amino acid sequence of the protein produced due to the degeneracy of the genetic code. SNPs in the coding region have two types: synonymous and nonsynonymous SNPs. Synonymous SNPs change the amino acid sequence of a protein while nonsynonymous SNPs do not change the amino acid sequence of a

protein.

As reported in the USA National library of medicine (Consortium et al. 2015), SNPs occur almost once in every 1,000 nucleotides on average, which means roughly 4 to 5 million SNPs are in a person's genome. These variations may be unique or occur in many individuals. Scientists have found more than 100 million SNPs in populations around the world. These variations are commonly found in the DNA between genes. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting gene function. Therefore, they can act as biological markers, helping scientists locate genes that are associated with the disease.

2.1.2 Copy Number Variations

Copy number variation (CNV) is the most common form of structural alternation. CNVs are the gain or loss of DNA segments in the genome, ranging in size from one kilo-base to several mega-bases. The CNVs result in more or fewer copies of a DNA region with respect to the normal genome (Feuk et al. 2006). CNVs are described by three characteristics: type, chromosomal position, and dosage, as is shown in Fig. 2.2.

The type of the CNV is either amplification or deletion. The chromosomal position of the CNV is described by the start and end position of the CNV in the chromosome. Dosage represents the total number of copies of the CNV, with a value less than two for deletion and greater than two for amplification. Besides, CNVs have phenotypic heterogeneity effects. This means that different CNV types and dosages at the same position in the chromosome can have a different impact on the phenotype.

In general, biologists assign CNVs to one of two major groups, depending on the length of the affected chromosomal region and occurrence frequency (Schrider and Hahn 2010). The first group involves copy number polymorphisms (CNPs), widespread in the general population, with an average occurrence frequency greater than one percent. The second CNV group is rare variants that are much longer than CNPs, ranging from hundreds of thousands of base pairs to over 1 million base pairs.

In the clinical setting, each CNV can be assigned to one of three main categories of clinical significance (Nowakowska 2017):

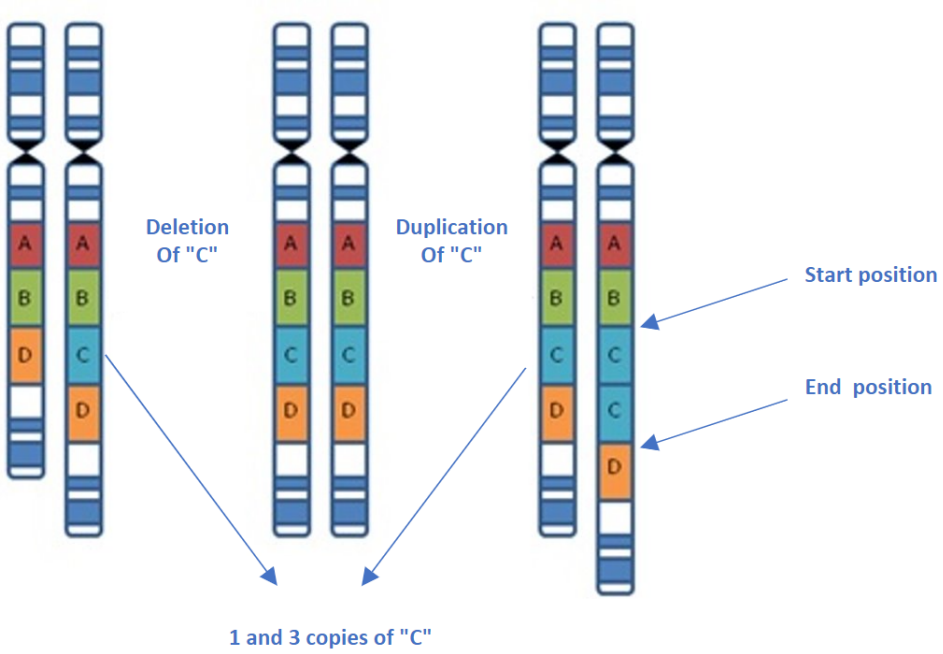

Figure 2.2: Characteristics of copy number variants: type, chromosomal position and dosage.

- Benign variants: the CNV is repeatedly observed in the normal population, and several peer-reviewed publications have reported there is no association between the CNV and specific disease related traits.
- Pathogenic variants: the CNV is observed in both normal and case population, and multiple peer-reviewed publications have reported there is a significant association between the CNV and specific disease related traits.
- Variants of uncertain significance (VOUS): the CNV is mostly not reported in the peer-reviewed publications and if reported there is no sufficient evidence for it's significant association with specific disease related traits. All CNVs that can not be classified as benign or pathogenic fall in this category.

There are several molecular mechanisms by which a CNV can convey a disease phenotype (Lupski and Stankiewicz 2005) shown in Fig. 2.3 including:

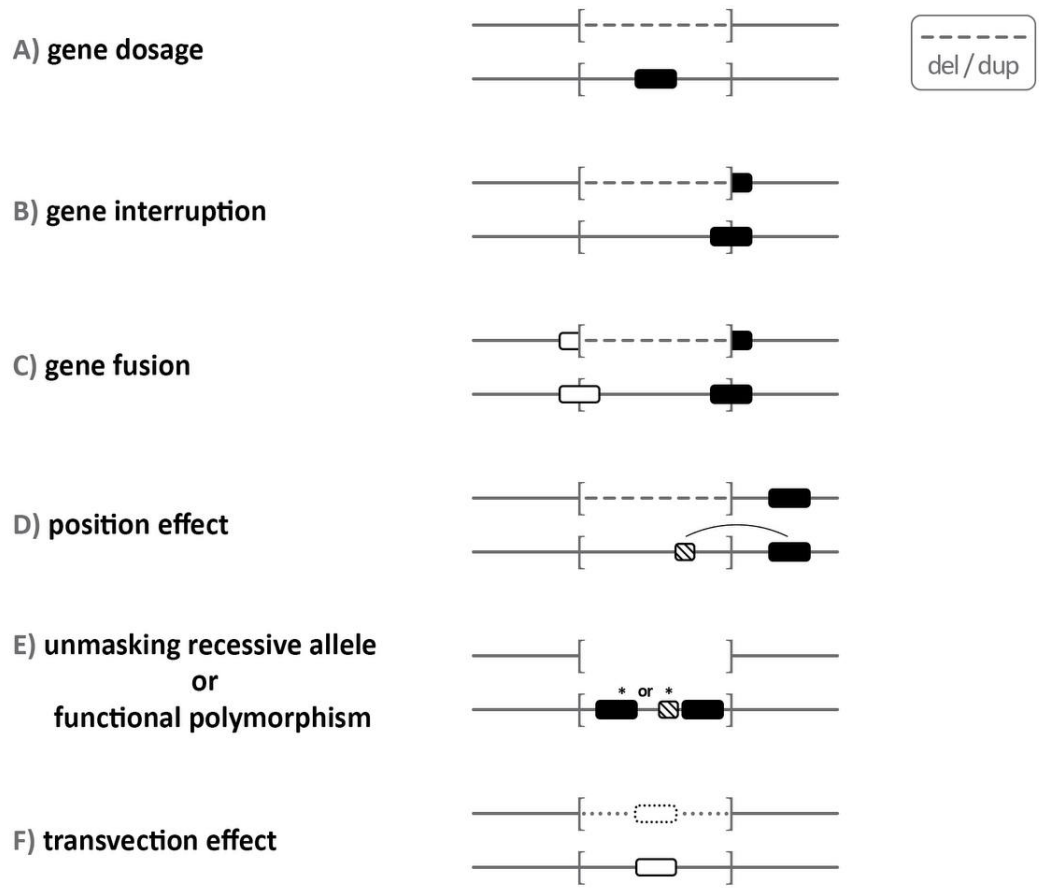

Figure 2.3: Molecular mechanisms of SNP phenotypes. The paired black lines represent chromosomal regions. Squared brackets ([ ]) represent the CNV region, both black and white squares show a gene, the dotted lines represent deletion or amplification.

- Gene dosage, where the CNV happens in a region that contains a dosage sensitive gene. The change in the copies of a dosage sensitive gene by deletion or amplification causes a phenotypic effect.
- Gene interruption, where the CNV interrupts a gene.
- Gene fusion, where the CNV forms a fusion gene from two independent genes existed in the CNV break-point.

- Position effect, where the CNV affects the expression or regulation of the gene that exists near the CNV break-point.
- Unmasking recessive allele, where the CNV caused to have only a single copy of a gene instead of two copies which alters the gene function.
- Interrupting effects of transvection, where the CNV affects the interaction between the alleles. Transvection is a phenomenon that occurs when an allele on one chromosome interacts with the homologous chromosome's corresponding allele. Transvection can lead to either gene activation or repression.

CNVs can cause a broad range of diseases through the above mentioned mechanisms, including Mendelian diseases known as single-gene diseases, complex diseases due to many genes, and cancer. Therefore, understanding the relationship between CNVs and diseases may provide important insights into genetic causes, leading to effective means in preventing and treating the diseases. As more CNVs are detected throughout the human genome, their potential role in developing diseases is being recognized based on the literature.

2.1.3 CNVs and Diseases

It has long been known that large chromosomal aberrations are identified being associated with human diseases. Down syndrome caused by human chromosome trisomy 21 is the best-known example which was revealed in 1959 by Lejeune (Zigman 2013). Conventional microscopy can detect such large chromosome abnormalities. It was subsequently found that copy number changes caused by submicroscopic genomic deletions were involved in human diseases and other traits, including thalassaemia (Higgs et al. 1979) and red-green blindness (Nathans et al. 1986). Besides deletions resulting in defects in gene function, duplication involving a dosage-sensitive gene can also cause disease. In 1991, the first disease-associated submicroscopic duplications were identified in the 17p12 locus. This duplication can lead to Charcot-Marie-Tooth disease (Lupski et al. 1991). Table 2.1 contains some studies that examine disorders caused by CNVs.

Table 2.1: Examples of disorders conveyed by CNVs

Disorder	Studies
Alzheimer's disease	Cuccaro et al. (2017) Song et al. (2015)
Autism	Yingjun et al. (2017) Vorstman et al. (2017) Stefano et al. (2019)
Crohn's disease	McCarroll et al. (2008) Fellermann et al. (2006)
HIV susceptibility	Yim et al. (2015) Gonzalez et al. (2005)
Parkinson's disease	La Cognata et al. (2017)
Schizophrenia	Marshall et al. (2017) Rees et al. (2016)

2.2 Genetic Association Studies

Genetic association studies are a major tool for studying the complex relationship between genetic variants and human health conditions. They test for a correlation between genetic variants and disease-related traits to identify candidate genome regions or genes that contribute to a specific disease-related trait. If an association is present, the genetic variant will be seen more often than expected by chance in an individual carrying the disease-related trait (Lewis and Knight 2012). Currently, SNPs are the most widely tested genetic variants in association studies but CNVs are also used.

In recent years, numerous new statistical methods have been proposed for detecting associations of genetic variants with diseases. Collapsing methods are a dominant mode of genetic association analysis, which study the association between a group of genetic variants and traits.

2.2.1 Collapsing Methods

There has been a surge in interest to find any associations between genetic variants and disease related traits with the advent of new sequencing technologies. Collapsing methods are the most common methods that are used to test the association between genetic variants, specifically SNPs, and disease related traits. Collapsing methods aggregate genetic variants in a specific chromosomal region

or a set of regions. Then, the association between the variants and disease related traits is tested by using statistical methods. So far, several association tests based on the collapsing method have been proposed and evaluated (Dering et al. 2014).

This section includes an overview of how collapsing methods work following the approach in Dering et al. (2011). First, the options in coding and analyzing genetic variants are discussed. Then, statistical tests for testing the genetic variants under the collapsing framework are explained.

Coding for Collapsing Methods

Generally, there are two fundamental ideas used for collapsing genetic variants: indicator coding and proportion coding. Let n_u and n_a be the number of unaffected and affected subjects respectively. n denotes the total of unaffected and affected subjects as $n = n_u + n_a$. A region of interest (ROI) is an arbitrary chromosomal region defined by base-pair positions on a chromosome, or a gene, a gene cluster, a combination of different genes (e.g., several genes from the same pathway).

The first approach, indicator coding, produces a binary variable x_{ij} as

$$x_{ij} = \begin{cases} 1 & \text{if variant present at position } i, \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

which indicates whether a variant is present or not at the chromosomal position i within the ROI in subject j. Then, x_j denotes whether subject j carries any rare variant in the ROI as

$$x_j = \begin{cases} 1 & \text{if } \sum_{i=1}^{K} x_{ij} > 0, \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

where K denotes the number of sites with variants in the ROI.

In the second approach, proportion coding, the number of variants of subject j is counted over all the K sites as

$$x'_j = \frac{1}{2K} \sum_{i=1}^{K} x'_{ij} \tag{2.3}$$

where x'_{ij} is the number of variants at position i of the subject j.

In both Equations 2.2 and 2.3, all variants are weighted equally. However, the frequency of a variant occurrence can affect on its effect size. Therefore, several

approaches like Madsen and Browning (2009) and Price et al. (2010) have been proposed to up or down variant weight in a ROI. Madsen and Browning (2009) proposed a weighting approach in which variants are weighted according to their occurrence frequency in unaffected subjects. More specifically, the position i is weighted inversely proportional to its variance as

$$\widehat{w}_i = 1/\sqrt{n_i \widehat{p}_i^{u'} \left(1 - \widehat{p}_i^{u'}\right)} \tag{2.4}$$

where

$$\widehat{p_i^{u'}} = \frac{\sum_{j=1}^{n_i^{u'}} x_{ij}^{u'} + 1}{2n_i^u + 2} \tag{2.5}$$

is an estimated frequency of a variant at position i in unaffected subjects and $x_{ij}^{u'}$ is the number of the variants in the unaffected subject j. Similarly, Price et al. (2010) proposed another weighting approach as

$$\widehat{w}_i = 1/\sqrt{\widehat{p_i^u}\left(1 - \widehat{p_i^u}\right)} \tag{2.6}$$

where $\widehat{p_i^u}$ is the estimated frequency of a variant at position i. However, the fact that a rare variant is present only in affected subjects is not considered in this coding approach.

Both the indicator and proportion coding approaches are flexible about the units that are considered for collapsing. The units can be genes, gene clusters and chromosomal regions as was explained previously. Furthermore, different variant frequency thresholds can be utilized for collapsing like 0.01 or 0.05 based on Price et al. (2010). In more details, only variants with occurrence frequency less than the threshold are considered to be collapsed. However, the analysis in the ROI can be done on common variants as well by considering greater thresholds. The threshold can be either variable or fixed.

There are different approaches for grouping variants including grouping based on the variant functionality or the variant predicted effect. In the grouping approach according to the variant functionality, only synonymous or non-synonymous variants can be grouped and counted. Conversely in the grouping approach based on the variant effect, variants with effects like beneficial, neutral or deleterious can be grouped separately. There are different standard packages like PolyPhen2 (Ramensky et al. 2002), SNAP (Bromberg et al. 2008) and SNPs3D (Yue et al. 2006) that can be used for grouping.

The different unit definitions, threshold types and variant groupings can be implemented by integrating indicator variables that flag the presence of the variant in Equations. 2.2 and 2.3.

Statistical Tests for Testing the Genetic Variants

There are several statistical approaches for genetic variants analysis under collapsing framework like Standard methods, collapsing and summation test, weighted-sum collapsing, and data-adaptive summation approaches. These approaches can be considered as the representatives for the similar existing approaches and are described briefly in the following.

Logistic regression or any other model from the family of generalized linear models are examples of the standard methods that can be used in genetic variant analysis as

$$E\left(y_j\right) = \beta_0 + \beta_1 x_j + \beta_2' z_j \tag{2.7}$$

where y_j is the phenotype of the subject j, and z_j is a vector of covariates like age and gender. The $\beta = (\beta_0, \beta_1, \beta_2)$ is the parameter vector to be estimated using likelihood ratio or score tests.

The collapsing and summation test (CAST) by Morgenthaler and Thilly (2007) is a straightforward approach for comparing case subjects with control subjects. In CAST, the number of control subjects with a variant is compared with the number of case subjects with a variant in a 2×2 frequency table. A limitation of the CAST is that it is restricted to rare variant analysis, and rare and frequent genetic variants can not be investigated jointly in this approach.

The weighted-sum (WS) collapsing approach proposed by Madsen and Browning (2009), can investigate rare and frequent genetic variants jointly. In WS, more weight is given to rare variants because they are expected to have stronger effects than the frequent ones.

The data adaptive summation approach proposed by Han and Pan (2010), is one the the first testing methods which can differentiate beneficial, neutral and deleterious effects of the genetic variant. They have introduced data-adaptive weights in which negative weights are allocated to the variant with beneficial effect and a positive weight is assigned to the variants with deleterious effect. Then, all variants are summed and used in the regression model.

Collapsing methods have been widely used in testing association between SNPs and disease related traits. However, due to the specific characteristics of CNVs applying collapsing methods on CNV data is not a straightforward task.

2.3 Challenges of Studying CNVs

As explained in previous sections, CNVs are described by three characteristics: type, chromosomal position and dosage. Each of these characteristics are multidimensional. The CNV type could be either amplification or deletion. The CNV chromosomal position is defined by start and end position. The CNV dosage can be any value less than two for deletion and greater than two for duplication. Therefore, SNPs collapsing methods can not be applied on CNV data straightforwardly because of the following reasons.

First, based on the Equation 2.1 collapsing methods assume SNPs as binary events: mutation versus no mutation. However, this binary approach is not applicable on CNVs considering their multidimensional characteristics.

Second, according to Equations 2.2 and 2.3 collapsing methods target only one feature of the SNPs like the total number of mutations. However, each of the three CNV characteristics can affect the CNV impact on the disease risk and are required to be considered in the association test together.

Third, in addition to the mentioned three multidimensional characteristics, CNVs often have both between loci and within locus etiological heterogeneity. Conversely, SNPs only exhibit between loci heterogeneity. Between loci etiological heterogeneity means CNVs in different loci, chromosomal positions, can have different effect on disease risk. On the other hand, within locus etiological heterogeneity means that even CNVs at the same locus but with different types or dosage can have different impact on disease risks. For example, Levinson et al. (2011) have identified a CNV of the deletion type at 22q11.2 chromosomal position as a risk factor for schizophrenia, whereas a CNV of the amplification type has been detected as a potentially protective factor by Rees et al. (2014).

Due to all these reasons, applying most SNP collapsing methods to CNV association analysis is not straightforward.

2.4 Methods for Studying CNVs

There are two ways to address the aforementioned difficulties in CNV association analysis: using collapsing random effect method or using kernel-based method. In the first strategy, multidimensional CNV information is broken into pieces: chromosomal location, type and dosage. Then, some collapsing methods are applied on a certain CNV information piece. In the second strategy, an appropriate kernel function is used to summarize the similarity between two CNVs. Then, this similarity is compared to the similarity in phenotypes to test whether there is any association between genotype and phenotype. A high correspondence between genotypic similarity and phenotypic similarity may suggest the existence of an association.

In the following subsections, we discuss three state of the art methods that have used these strategies to test the association between CNVs and disease related traits: CNV collapsing random effects test, CNV kernel association test and copy number profile curve-based association test.

2.4.1 CNV Collapsing Random Effects Test

CNV collapsing random effects test (CCRET), which is proposed by Tzeng et al. (2015), is an extension of collapsing methods applicable to CNV data. CCRET deals with CNV multidimensional characteristics by breaking CNV information into some pieces. Then, it applies the collapsing method on the chosen CNV characteristic and tests it's effect on disease related traits.

An overview of CCRET is shown in Fig. 2.4 and consists of four key steps to test the association between the CNV dosage and disease related traits. First, CCRET forms CNV regions (CNVRs) based on a predetermined amount of overlap among CNV chromosomal positions of the CNVs existing in subjects' CNV profiles as is depicted in the part (I) of the figure. In the step (II), CNV information of every subject in each CNVR is stored in a matrix. In more detail, CCRET breaks down the CNV information into three pieces: dosage, length and gene intersection. Then, each piece is stored in an input matrix. Z^{DS}, Z^{Len} and Z^{GI} matrices are used to store dosage, length and gene intersection information of the CNVs respectively. In the third step, CCRET chooses one CNV feature as the feature of interest and creates a similarity matrix by measuring the similarity

between CNV profiles based on the chosen feature. CCRET treats the other two features as a binary event which is explained previously and only the feature of interest can take different values. Finally, CCRET uses a score test to test the association between the CNV feature of interest, which is dosage as an example in Fig. 2.4 , and disease related traits.

Although CCRET is a more complex and modified version of the collapsing method, it still falls short in some aspects:

- CCRET does not use all multidimensional CNV characteristics at the same time. It is only capable of testing one of the CNV feature's effect on disease related traits at each time. Therefore, all CNV characteristics are not considering in the association test.

- CCRET forms CNV regions using an arbitrary overlapping threshold (a predetermined threshold). Since, there is not a unique overlapping pattern among all studies, the arbitrary choice of the overlapping threshold impacts the formation of CNV regions. Consequently, how the between and within locus heterogeneous effects of CNVs are interpreted is not clear.

- CCRET only considers CNV length not the CNV chromosomal positions which results in considering CNVs with the same length as similar to each other without taking into account their chromosomal start and end positions.

- CCRET can consider at most one CNV for each sample in each CNV region. Otherwise, it is not clear how to calculate input matrices.

Therefore, even an extension of collapsing methods is not capable of dealing with CNV multidimensional characteristics and utilizing them in the association test.

2.4.2 Kernel-based Association Tests

In statistics, a kernel refers to a function that measures the similarity or dissimilarity between data points (Shawe-Taylor et al. 2004). Kernels play a fundamental role in various statistical methods, particularly in kernel methods, where they are used to transform data into higher-dimensional feature spaces to capture complex patterns and relationships. Formally, a kernel function, denoted as $K(x, y)$, takes two data points, x and y, as input and outputs a real value that represents the degree of similarity between them. The kernel function satisfies

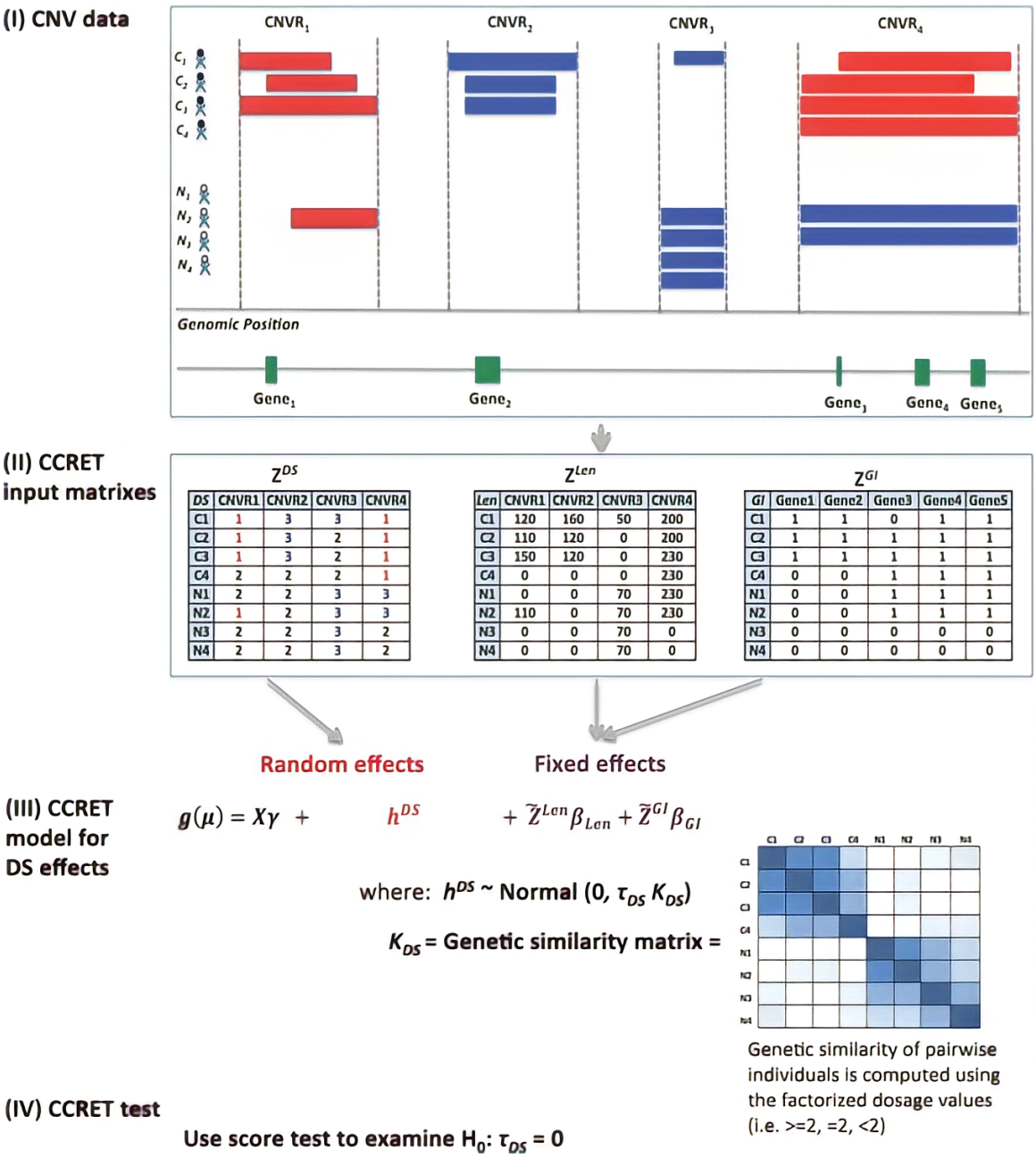

Figure 2.4: An overview of the CCRET with the dosage model as an example from (Tzeng et al. 2015). C_{1-4}: cases, N_{1-4}: controls, CNVR: copy number variation region, red rectangle: deletion, blue rectangle: duplication, green rectangle: gene. DS: dosage, Len: length, GI: gene intersection. In part (I), CNVRs are created, in part (II) CNV information fir each subject is stored in a matrix, and in part (III) the association between CNV characteristics and disease related traits is tested.

certain properties, such as symmetry ($K(x, y) = K(y, x)$) and positive definiteness, which ensures that it provides a valid measure of similarity. Kernels are commonly used in statistical techniques such as kernel density estimation, kernel regression, support vector machines (SVM), and kernel-based association tests. The choice of the kernel function depends on the characteristics of the data and the specific problem at hand.

Several popular kernel functions exist, each with its own properties and applicability to different types of data. Some commonly used kernels include: Gaussian kernel, Polynomial kernel and Laplacian kernel. These are just a few examples of kernels, and there are numerous other types available. The choice of the kernel depends on the specific problem, the underlying assumptions, and the desired properties of the transformed feature space.

Kernels are a powerful tool in statistics as they enable the analysis of complex data and the extraction of nonlinear patterns. They provide a flexible framework for statistical modeling and inference, allowing researchers to uncover hidden structures and relationships in diverse data sets.

In recent years, there has been a growing interest in leveraging kernel methods to analyze and identify associations between variables in complex data sets. One powerful tool that has emerged is the kernel-based statistical association test (Pfister et al. 2018). Association testing is a fundamental task in statistics, aiming to assess the presence of a relationship or dependency between variables. Traditional association tests, such as Pearson's correlation coefficient or chi-square test, are primarily designed for linear relationships. However, real-world data sets often exhibit complex nonlinear patterns that cannot be adequately captured by linear methods.

The kernel-based statistical association test extends traditional association tests to handle nonlinear relationships by incorporating kernel functions. The test operates by first transforming the data using a kernel function, which implicitly maps the variables into a higher-dimensional space. Then, the test statistic is computed based on the transformed data, capturing the strength and significance of the association between the variables.

Kernels have been widely used in genetic association studies as a similarity measure to conduct statistical tests like in Liu et al. (2007, 2008); Wu et al. (2010, 2011); Zhan et al. (2015a,b). A typical kernel association test includes the

following two steps. First, the similarities between two multidimensional genetic variants are summarized by an appropriate positive semi-definite kernel. Then, the calculated similarity between genetic variants is compared to the similarity between disease related traits. A high correlation between genotypic similarity and phenotypic similarity may indicate the presence of an association.

Two following methods are existing CNV kernel-based association tests that use kernels to measure similarity between CNV profiles:

CNV Kernel Association Test

CNV kernel association test (CKAT) is a kernel-based association test is proposed by Zhan et al. (2016) to test the association between CNVs and disease related traits. Motivated by kernel strategy, CKAT measures the similarity between CNV profiles using two kernels: the single-CNV kernel and the CNV region kernel. These two kernels are designed so that they are capable of dealing with both CNV multidimensional characteristics and heterogeneity effects.

The single-CNV kernel, is responsible for measuring the similarity between two CNVs. Lets assume that $X = (X^{(1)}, X^{(2)})$ denotes a CNV. $X^{(1)}$ is the length of the CNV which is calculated by subtracting CNV start position from CNV end position. $X^{(2)}$ is the type information of the CNV, taking values 1 and 3 for deletion and duplication respectively. The single-CNV kernel measures the similarity between two arbitrary CNVs X_1 and X_2 as

$$K\left(X_1, X_2\right) = \exp\left\{-\frac{\left(X_1^{(1)} - X_2^{(1)}\right)^2}{\rho}\right\} \times \left[\frac{I\left(X_1^{(2)} = X_2^{(2)}\right) + 1}{2}\right] \tag{2.8}$$

The first term is the contribution of the CNV length. A Gaussian kernel function and shape parameter ρ which is set to 1 in CKAT are used to measure the similarity between two CNVs from length aspect. The second term is the contribution of the CNV type. The $I(\cdot)$ is the identity function that takes the value 0 when the two CNVs are of the different type and 1 otherwise.

The second kernel, the CNV region kernel, measures the similarity between two CNV profiles with respect to all CNVs existing in a CNV region. In CKAT, each chromosome is treated as a CNV region that can contain any number of CNVs. Let R_i and R_j be the CNV profile of sample i and j. The CNV region

kernel, K_R, in a particular region between sample i and j is defined as

$$k_R(R_i, R_j) = \begin{cases} 0 & \text{if } p_i p_j = 0 \\ \max_{l=0,1,\ldots,p_i-p_j} \sum_{t=1}^{p_j} k\left(X_{t+l}^i, X_t^j\right) & \text{if } p_i \geq p_j > 0, \\ \max_{l=0,1,\ldots,p_j-p_i} \sum_{t=1}^{p_i} k\left(X_t^i, X_{t+l}^j\right) & \text{if } p_j > p_i > 0 \end{cases} \tag{2.9}$$

where p_i and p_j are the number of CNVs in the CNV profile of subjects i and j in the region respectively and $k(\cdot, \cdot)$ is the single-CNV kernel defined previously. The maximum operation in the definition of $k_R(\cdot, \cdot)$ searches for the best CNV-to-CNV correspondence in the CNV profiles of subjects i and j in the CNV region. The output of $k_R(\cdot, \cdot)$ is a similarity matrix K, where $k_{ij} = k_R(R_i, R_j)$ represents the similarity between CNV profile of subjects i and j in a specific region.

Finally, the association between CNVs and disease related traits is tested by a logistic regression kernel-based model. The following logistic regression model is used to relate disease to CNVs where Z are covariates such as age, gender, and $f(\cdot)$ is a function calculated by the CNV region kernel and β and β_0 are intercepts

$$\text{logit}\left[\Pr\left(y_1 = 1\right)\right] = \beta_0 + Z\beta + f\left(R_i\right) \tag{2.10}$$

Using a biologically relevant region like a chromosome to define CNV regions in CKAT bypasses the need for an arbitrarily defined locus in CCRET that there was a need to define an arbitrary overlap threshold. Also, there is no limit to the number of CNVs in each regions like CCRET. However, there are still some limitations in CKAT:

- CKAT uses a shift-by-one scanning algorithm to align pairs of CNVs based on their ordinal position rather than considering all possible pairs. This strategy results in not optimally capturing the similarity between each possible CNV pair especially when dealing with frequent CNVs.

- CKAT does not exploit the full information of a CNV when measuring the similarity between two CNVs. The kernel function measures CNV similarity based on type and length features between two CNV events and the contribution of dosage is not considered.

- CKAT, like CCRET, considers CNVs with the same length to be similar to each other without taking into account the exact chromosomal position of the CNVs.

Therefore, there is still a need to improve the kernel-based association tests so that they be able to deal with CNV characteristics from all aspects.

Copy Number Profile Curve-based Association Test

Brucker et al. (2020) proposed the copy number profile curve-based association test (CONCUR) to address aforementioned limitations in quantifying CNV similarity using kernel-based methods. An overview of CONCUR is shown in Fig. 2.5. CONCUR has two major components: copy number (CN) profile curve and common area under the curve (cAUC) kernel. The CN profile curve describes an individual's CNVs across the genome or a region of interest. The cAUC kernel measures the CNV similarity between two individuals. Then, the CNV effects on the phenotype are characterized.

Copy number profile curves are a visualization of CNV activity across the genomes. CNV dosage is shown on the y-axis with the baseline of 2 since there is 2 copy of each segment in a normal genome. The start and end positions of the CNV are shown on the x-axis. By superimposing two CN profile curves, the same type of overlapping regions of CNVs are identified. Then, the common area under the curve between two individuals is calculated as the sum of all areas of commonality in their duplication profile curves plus the sum of all areas in their deletion profile curves. Finally, CONCUR regresses the trait values of CNV effects captured by the cAUC kernel and evaluates the association between traits and CNV profiles via a score-based variance component test. CONCUR outperforms both CCRET and CKAT by being free from a definition of a CNV region and exploiting all CNV characteristics for calculating the similarity between CNV profiles of two individuals. However, it is still not able to capture the similarity in an optimal way and deal with the CNV heterogeneity effect.

CONCUR calculates the similarity between CNV profiles by summing the common area of the deletion and duplication curves separately. Therefore, the case when two individuals have a CNV of different types in the same locus, is not considered in the similarity measurement. This is because the common area under the curve is not meaningful in this case. Although it is not clear when two individuals have different types of CNV in a same chromosomal position may result in more or less similarity between them, ignoring the occurrence of the CNV in that position may lead to missing the CNV phenotypic heterogeneity effect.

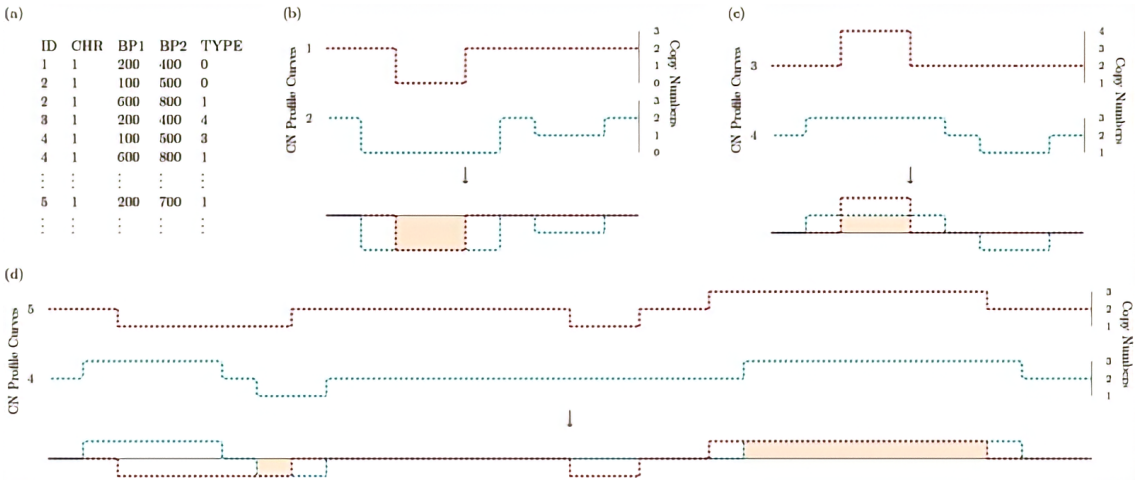

Figure 2.5: Diagram of copy number profile curves and common area under the curve by Brucker et al. (2020). (a) Example of CNV data describing individuals' CNV profile in chromosome 1. (b) Copy number (CN) profile curves of two individuals with overlapping deletions of dosage 0. (c) CN profile curves of two individuals with overlapping with overlapping duplications of dosage 3 and 4. (d) The cAUC between two individuals who have overlapping deletions of dosage 1 and overlapping duplications of dosage 3, so that the cAUC between the individuals is the sum of the two areas.

2.5 Research Gaps

Ongoing advances in high-throughput genomic technology have made possible the rapid identification of several genetic variants possible, including CNVs. Understanding the relationship between CNVs and diseases may provide important new insights into the underlying genetic causes and may also lead to effective methods of prevention and treatment. Therefore, there is increasing research interest in the role of CNV in the etiology of many complex diseases. However, as this literature review shows, although CNV data are available, methods to test the association between CNV and disease-related traits are few and have yet to be developed.

None of the methods in the literature, neither collapsing nor kernel-based, can deal with CNV multidimensional characteristics. CCRET, an extension of the collapsing method, is capable of testing the association between CNVs and diseases considering only one CNV characteristic at a time. CKAT, a kernel-based approach, not only ignores the CNV dosage in the association test but also it is

only applicable to rare CNVs, not frequent CNVs, based on the algorithm it uses to compare CNV profiles. Similarly, CONCUR, the other kernel-based approach, ignores the CNV heterogeneity effect.

In addition, to our knowledge, the CNV sequential order has not been studied yet. Therefore, there is an open question around CNV sequential order. More specifically, it is still unclear whether CNVs are randomly distributed across the genome, or their order is significant and associated with the disease, like SNPs.

The work presented in this thesis addresses the existing gaps in studying CNVs and answers open questions about the significance of the CNV sequential order in association with disease related traits. Testing the association between CNVs and disease-related traits based on all CNV characteristics is explored in Chapter 3 using a multi-dimensional kernel-based association test called MCKAT. In Chapter 4, a sequential multi-dimensional association test named SMCKAT is developed which tests whether if there is any significant association between CNV sequential order and disease related traits. In Chapter 5, our proposed association tests are used to test if considering the dual effects of CNVs and other genetic variation like genes is informative and helpful in testing the association between CNVs and disease related traits.

Chapter 3

MCKAT, a multi-dimensional copy number variant kernel association test

3.1 Introduction

As discussed in section 2.2, most work on testing the association between genetic variants and disease-related traits to date has focused on SNP data. However, these methods are not applicable to CNV data due to the multi-dimensional characteristics of CNVs. Therefore, there is a strong motivation to develop methods that specifically account for the CNV data. Of the methods that are designed for CNV data, CCRET (Tzeng et al. 2015) the random effect collapsing approach, is not capable of dealing with CNV multidimensional characteristics. Similarly, CKAT (Zhan et al. 2015b) and CONCUR (Brucker et al. 2020), the kernel based approaches fall short in dealing with CNV data as discussed in sections 2.4.2 and 2.4.3 respectively.

This chapter presents a multi-dimensional copy number variant kernel association test (MCKAT) which is not only capable of indicating stronger evidence in detecting significant associations between CNVs and disease-related traits, but is applicable to both rare and frequent CNV datasets. As with CONCUR and CKAT, the MCKAT is based on the kernel approach and so an appropriate group of kernels is used to summarize the similarity between CNV profiles. Then, this similarity is compared to the similarity between the presence of the disease-related trait in CNV profiles. A high correspondence between the CNV profiles' similar-

ity and the disease presence similarity may suggest existence of association. In contrast to CONCUR and CKAT, the multidimensional nature of the kernels presented in the MCKAT allows all CNV multidimensional characteristics plus the CNV heterogeneity effect be considered in the association test. Furthermore, the algorithm that measures the similarity between CNV profiles in the MCKAT is capable of capturing similarity between every possible CNV pair in CNV profiles which makes the association test applicable to both rare and frequent CNV data.

MCKAT is described in Section 3.2, including the specification of the kernels and the kernel based association test. The performance evaluation of the MCKAT using simulated data is described in Section 3.3. Results on real data are presented in Section 3.4. The results are discussed in Section 3.5 and Section 3.6 summarises and concludes the chapter.

The R code used to implement the MCKAT is given in Appendix A and is available at https://github.com/nesfehani/MCKAT. The work in this chapter addresses Contribution 1 listed in the Chapter 1, proposing an association test to deal with CNV multi-dimensional characteristics and the heterogeneity effect to identify CNV chromosomal regions that CNVs on them are significantly associated with disease-related traits. The contribution is addressed by developing a multi-dimensional kernel-based CNV association test which is included in a published article (Maus Esfahani et al. 2021b).

3.2 Model Development

We design a multi-dimensional kernel framework capable of measuring the similarity between CNV profiles utilizing all CNV characteristics. It contains two kernels. The first kernel, the single-pair CNV kernel, measures the similarity between a single CNV pair. It includes three sub-kernels. Each sub-kernel is responsible for measuring the similarity between two CNVs with respect to each of three CNV characteristics. The second sub-kernel, the whole chromosome kernel, aggregates the similarity between every possible CNV pair chromosome-wise to measure the total similarity between the CNV profiles of the subjects. Finally, the association between CNVs across a chromosome and disease-related traits is tested by comparing the similarity in CNV profiles to that in the trait using an association test.

3.2.1 Single-pair CNV Kernel

All CNV features including chromosomal position, type and dosage are used to measure the similarity between a single pair CNV. Let $X = (X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)})$ denote a CNV, where $X^{(1)}$ and $X^{(2)}$ are the start and end chromosomal positions of the CNV respectively, $X^{(3)}$ is the type information of the CNV taking the value 1 for a deletion and 3 for an amplification, and $X^{(4)}$ is the dosage information of the CNV taking the value of 0 or 1 for deletion, and > 2 for amplification. Considering two arbitrary CNVs X_1 and X_2, we define the kernel function between a CNV pair as

$$K_s(X_1, X_2) = \left[\frac{Intersection\left(\left(X_1^{(1)}, X_1^{(2)}\right), \left(X_2^{(1)}, X_2^{(2)}\right)\right)}{Union\left(\left(X_1^{(1)}, X_1^{(2)}\right), \left(X_2^{(1)}, X_2^{(2)}\right)\right)}\right] \times \left[\frac{\left(X_1^{(3)} == X_2^{(3)}\right) + 1}{2}\right] \times \left[\frac{1}{2^{\left|DR\left(X_1^{(4)}\right) - DR\left(X_2^{(4)}\right)\right|}}\right] \tag{3.1}$$

the first term is the CNV chromosomal position's contribution, which is described by measuring the mutual presence of a CNV with a specific start and end chromosomal position. It is defined as the size of the intersection of two CNVs divided by the size of their union. The maximum value for chromosomal position contribution is 1 when two CNVs have the same start and end position and 0 when two CNVs do not intersect.

The second term is the contribution from the CNV type. When two CNVs have the same type (both deletion or amplification), it takes the value of 1 and 0 when CNVs are of different types. The last term is the contribution of CNV dosage information. The similarity between two CNVs based on their dosage information is measured by a function called the Difference from the Reference (DR) as $DR(dosage) = |dosage - 2|$. We use 2 as a reference value. According to equation (3.1), the smaller difference between the DR value of two CNVs results in a greater similarity between them.

3.2.2 Whole Chromosome CNV Kernel

After measuring the similarity between two CNVs another kernel is needed to

compare the whole CNVs in a specific chromosome of one subject with another subject to calculate their similarity. To do this, we propose another kernel that is capable of measuring the similarity between all CNVs of two subjects in a chromosome.

Let $R_i = (X_1^i, ..., X_{p_i}^i)$ be the CNVs of subject i in a specific chromosome, where CNVs are ordered according to their chromosomal position and p_i is the number of CNVs of the sample i in the chromosome. Similarly, we have another CNV series $R_j = (X_1^j, ..., X_{q_j}^j)$ for subject j. Then, the whole chromosome CNV kernel between subject i and j in a particular chromosome is defined as

$$K_w(R_i, R_j) = \begin{cases} 0 & if \;\; p_i \times q_i = 0 \\ \\ \sum_{i=1}^{p_i} \sum_{j=1}^{q_j} K_s(X_i, X_j) & if \; p_i \times q_i \neq 0 \end{cases} \tag{3.2}$$

where $K_s(\cdot, \cdot)$ is the single pair CNV kernel from (3.1). The whole chromosome CNV kernel measures the similarity between every possible pair of CNVs in the CNV profiles of two subjects and aggregates these similarities to calculate the total similarity in a particular chromosome. To build a kernel-based association test described in the following section, we need to build a kernel similarity matrix K. K is a $n \times n$ matrix, where $K_{ij} = K_w(R_i, R_j)$. K_{ij} expresses the similarity between subject i and subject j measured by K_w.

3.2.3 Kernel-based Association Test

We use the following logistic regression model to test the association between CNVs and phenotype

$$logit[Pr(y_i = 1)] = \beta_0 + Z\beta + f(R_i) \tag{3.3}$$

Let $i = 1, 2, ..., n$ be the subjects and y_i the status of phenotype for subject i. $y_i = 1$ denotes the existence of that phenotype and $y_i = 0$ denotes its absence. Z is the covariant matrix which could include phenotype contributing factors such as certain inherited conditions, gender and age. $f(\,.\,)$ is a function of the CNV information, such as the CNV type and dosage, characterized by the whole chromosome CNV kernel $K_w(\,.\,,\,.\,)$.

According to equation (3.3), the association between the existence of a phenotype and CNVs can be examined by testing the hypothesis $H_0 : f(\cdot) = 0$. To do

this, we treat $f(\cdot)$ as a random effect vector with $N(0, \tau K)$ distribution. τ is a variance component parameter and k is the $n \times n$ similarity matrix generated by the whole chromosome CNV kernel K_w. Demonstrated by Liu et al. (2008), testing $H_0 : f(\ .\ ) = 0$ is equivalent to test $H_0 :\ \tau =\ 0$ under the logistic mixed effect model. Following Wu et al. (2010), Zhan et al. (2016) and Liu et al. (2008), we use a restricted maximum likelihood-based score test which is $Q = (y - \widehat{y})' K (y - \widehat{y})$.

The $\widehat{y}$ is the estimate of y in equation (3.3) under the null model $logit[Pr(y_i = 1)] = \beta_0 + Z\beta$. Then, we calculate the p-values of association between the status of the phenotype and CNVs by using Davies method Davies (1980) as implemented in the CKAT R package (Zhan et al. 2016).

3.3 Model Evaluation and Simulation Results

We conduct simulations to evaluate the performance of MCKAT and ensure that it can properly handle type I and II errors as well as having relatively high power in detecting existing associations. We focus on assessing MCKAT performance in detecting associations using chromosomal region × type × dosage effects in both rare and common CNV datasets. Apart from MCKAT, the CKAT and CONCUR approaches are also studied. We conduct our simulation studies under two main scenarios: rare CNVs and frequent CNVs. In the first scenario, we apply MCKAT, CKAT and CONCUR on a random chromosome to have limited number of CNVs for each subject to mimic a rare CNV dataset while in the second scenario, we apply them on the CNVs across whole genome to assess their performance in dealing with frequent CNV datasets. In the first scenario, each subject can have a maximum of five CNVs in their CNV profile to mimic rare CNV profile and in the second scenario there is no restriction on the number of CNVs to mimic frequent CNV profile. The dosage can take 0 or 1 for deletions and any value greater than two for amplifications in both scenarios. We simulated 10^5 datasets for each simulation scenario.

The CKAT evaluates the association between CNVs and disease-related traits through the following model (Zhan et al. 2015b):

$$logit(Pr(Y_i = 1)) = \beta_0 + \sum_{j=1}^{m_i} \left(\beta_j^{Del} I[X_{ij}^{(2)} = 1] + \beta_j^{Dup} I[X_{ij}^{(2)} = 3] \right) X_{ij}^{(1)} \quad (3.4)$$

where $X_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)})$ is the jth CNV of ith subject, β_0 is the prevalence rate

of the disease, and β_j^{Dup}, β_j^{Del} are the log of the odd ratio of CNV j for duplication and deletion respectively.

Similarly, CONCUR uses the following logistic model to generate disease-related trait status (Brucker et al. 2020):

$$logit(Pr(Y_i = 1)) = \gamma_0 + \beta_X X_i + \sum_{j=1}^{R} \beta_j^{\text{Dup}}\, Z_{ij}^{\text{Dup}} + \sum_{j=1}^{R} \beta_j^{\text{Del}}\, Z_{ij}^{\text{Del}} + \sum_{j=1}^{R} \beta_j^{\text{Len}}\, Z_{ij}^{\text{Len}} + \sum_{j=1}^{R} \beta_j^{\text{Dup*Len}}\, Z_{ij}^{\text{Dup}}\, Z_{ij}^{\text{Len}} + \sum_{j=1}^{R} \beta_j^{\text{Del*Len}}\, Z_{ij}^{\text{Del}}\, Z_{ij}^{\text{Len}} \tag{3.5}$$

where γ_0 and β_X are roughly set to -2 and $\log(1.1)$ respectively based on the baseline disease rate, $i = 1, ..., N$ indexes individuals, and $j = 1, ..., R$ indexes the CNV of regions. β_j^{Dup} and β_j^{Del} are the log odds ratio for the presence of a CNV versus its absence in segment j. Likewise, β_j^{Len} controls the effect of the CNV length and lets this effect differ by the CNV dosage value. Z^{Del} , Z^{Dup} and Z^{Len} are matrices which are generated based on CNV profiles. Z^{Dup} and Z^{Del} take value 1 for the CNV profiles that have a CNV in the CNV region j and 0 otherwise. Similarly, Z^{Len} codifies the length of the CNVs in the CNV regions and considers zero when a CNV profile is without CNVs in a specific region.

We use CNV datasets of 877 individuals with neurological deficits including dyslexia and intellectual disability, as well as 337 controls for our simulation studies. These datasets are publicly available in Girirajan et al. (2011). Briefly, the dyslexia dataset has 1,041 CNVs for 376 individuals and the intellectual disability dataset has 1,686 CNVs for 501 individuals. Similarly, the control dataset has 1,074 CNVs for 337 healthy subjects. The proportion of deletions to amplifications is almost 0.35 to 0.65 in all three datasets. The dosage value is 1 and 3 for all deletions and amplifications respectively in the datasets. Therefore, we randomly generate other values for the CNV dosage to conduct our simulation study and investigate the dosage effect in identifying existing associations. The simulated dosage value can take 0 or 1 for deletion types and 3, 4, ..., 7 for amplification types. We use equal probabilities when generating random dosage values for deletion and amplification, 0.5 (1/2) and 0.2 (1/5) respectively.

After preparing the CNV data, we propose the following logistic model to generate the case-control label Y_i using each CNV characteristics and their combined

effect:

$$
\begin{aligned}
logit(Pr(Y_i = 1)) &= \beta_0 + \sum_{j=1}^{m_i} \beta_j^{Len}(X_{ij}^{(2)} - X_{ij}^{(1)}) + \sum_{j=1}^{m_i}(\beta_j^{Del} I[X_{ij}^{(3)} = 1] \\
&+ \beta_j^{Amp} I[X_{ij}^{(3)} = 3]) + \sum_{j=1}^{m_i} \beta_j^{Dsg} |X_{ij}^{(4)} - 2| \\
&+ \sum_{j=1}^{m_i} \beta_j^{Len*Del*Dsg}(X_{ij}^{(2)} - X_{ij}^{(1)}) \times I[X_{ij}^{(3)} = 1] \times X_{ij}^{(4)} \\
&+ \sum_{j=1}^{m_i} \beta_j^{Len*Amp*Dsg}(X_{ij}^{(2)} - X_{ij}^{(1)}) \times I[X_{ij}^{(3)} = 3] \times X_{ij}^{(4)}
\end{aligned}
\tag{3.6}
$$

where $i = 1, ..., N$ indexes individuals, and $j = 1, ..., m_i$ indexes the CNVs of individual i. $X_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)}, X_{ij}^{(3)}, X_{ij}^{(4)})$ is the jth CNV of the ith individual as defined previously. β_0 corresponds to a baseline disease rate. β_j^{Len} controls the effect of chromosomal position, and β_j^{Del} and β_j^{Dup} are the log ratios of a CNV j for being related to a deletion versus an amplification and vice versa. Likewise, β_j^{Dsg} controls the effect of dosage in CNV j. $\beta_j^{Len*Amp*Dsg}$ and $\beta_j^{Len*Del*Dsg}$ allow the effect of the chromosomal position and CNV type to differ by dosage in CNV j.

The Q-Q (quantile-quantile) plots are very important in statistics to graphically analyze and compare two probability distributions by plotting their quantiles against each other. The Q-Q plot's points will flawlessly lie on the straight line y = x (45 degree line) if the two distributions we are comparing are exactly equal. When Q-Q plot's points are above or below the 45 degree line means there is the possibility to have type I error, rejecting the null hypothesis when it is true, and type II error, failure in rejecting null hypothesis when it is actually false, respectively. The QQ-plots of p-values of MCKAT versus CKAT and CONCUR under both simulation scenarios are presented in Figure 3.1.

As is shown in QQ-plot (a), MCKAT lies on the 45-degree line under different nominal significance levels even as low as 10^{-5} which means observed p-values calculated by MCKAT have the same value as the actual p-values. This indicates that MCKAT can have the correct type I and II error rates when testing an association between rare CNVs and disease-related traits. CKAT is more conservative when the significance level is small which means it is not capable of identifying significant association and handling type II error. Conversely, CONCUR is less

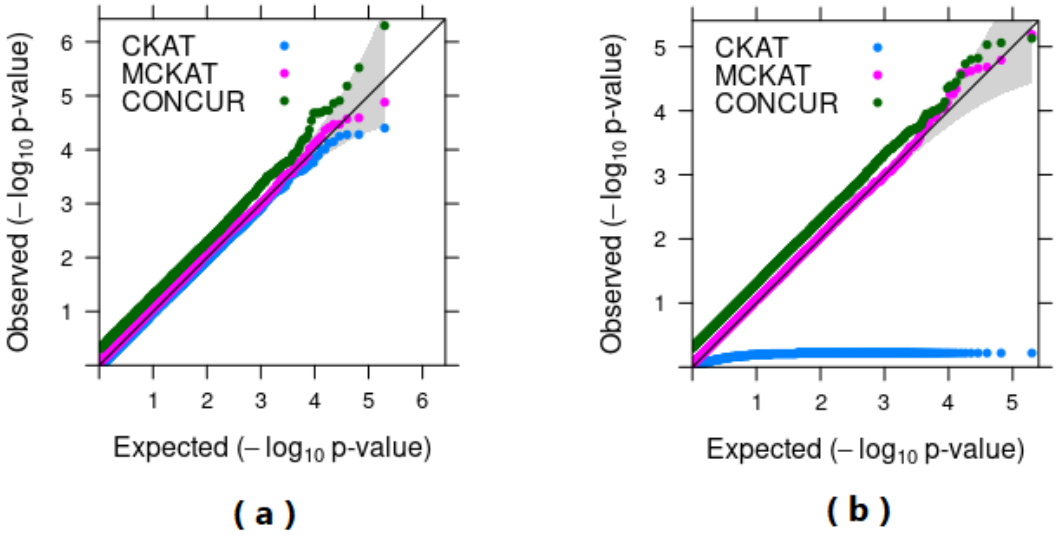

Figure 3.1: P-value based QQ-plots of MCKAT, CKAT and CONCUR under first (a) and second (b) simulation scenarios.

conservative when the significance level is small which means it identify significant associations when there are not any and can not handle type I error properly.

QQ-plot (b) presents the p-values of MCKAT, CKAT and CONCUR under the second simulation scenario. As shown, MCKAT can protect the correct type I and II error rates in the second scenario as well by being on the 45-degree line under different nominal significance levels. Similar with scenario one, CONCUR is less conservative when the significance level is small. However, CKAT can not identify any significant association in the frequent CNV data.

In statistics, the power of an statistical association test is the probability that the test correctly rejects the null hypothesis when an alternative hypothesis is true. Statistical power ranges from 0 to 1. As the power of a statistical test increases, the probability of making a type II error by wrongly failing to reject the null hypothesis decreases. The statistical test power is calculated using below equation:

$$\text{Power} = \Pr(\text{reject null hypothesis} \mid \text{alternative hypothesis is true}) \quad (3.7)$$

The empirical powers of MCKAT and CKAT under the first and second scenarios versus effect size are presented in Figures 3.2 and 3.3 respectively. The

effect size is a magnitude of an effect. As the effect size increases so does the likelihood of the detecting it which is essentially an increase in power. In more detail, the effect size is the estimation of the overlap between two distributions which is affected by both the distance between the population means and the standard deviations. The effect size is calculated using below equation:

$$\text{Effect Size} = \frac{\text{The estimated difference in the means}}{\text{Pooled estimated standard deviations}} \tag{3.8}$$

We observe that MCKAT has better power, higher performance in handling type II error, compared with CKAT under both scenarios. One reason might be that the MCKAT is designed to detect the dosage and the chromosomal position $\times$ type $\times$ dosage signals but CKAT struggles to pick up the signals due to its design. Another reason for CKAT's low power, especially under the second scenario, could be its scanning algorithm for aligning CNVs. CKAT's shift-by-one scanning algorithm may result in not capturing signals when dealing with greater numbers of CNVs in common CNV data.

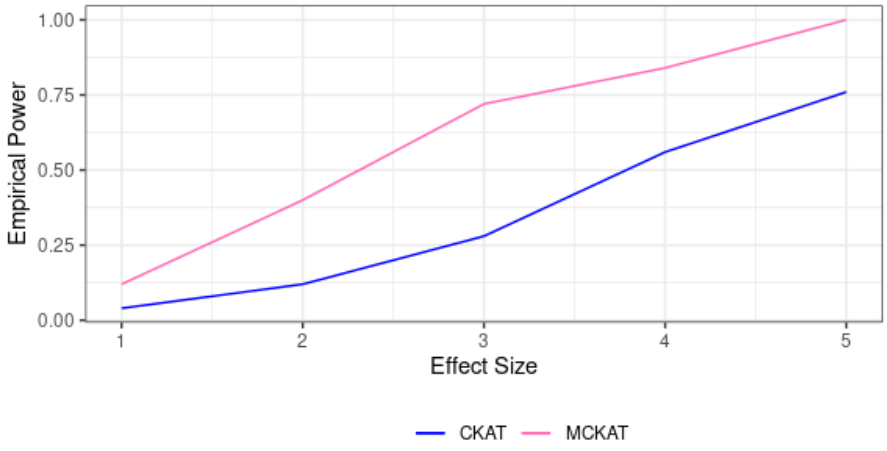

Figure 3.2: Empirical power of MCKAT and CKAT under first simulation scenario, rare CNV data.

Likewise, the empirical powers of MCKAT and CONCUR under the first and second scenarios are presented in Figures 3.4 and 3.5 respectively. We again observe that MCKAT has better power compared with CONCUR under both scenarios. One reason might be that CONCUR is not capturing the heterogeneity effect of CNVs. CONCUR, measures the similarity among CNV profiles by summing up

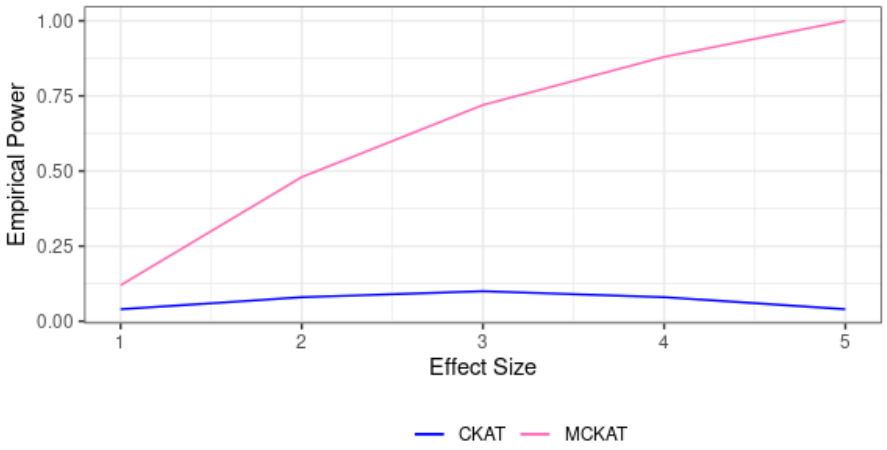

Figure 3.3: Empirical power of MCKAT and CKAT under second simulation scenario, frequent CNV data.

the similarities with respect to the CNV types, deletion and amplification. This approach may result in ignoring the within heterogeneity CNV effect specially when two CNVs with different types have a different impact on disease-related traits.

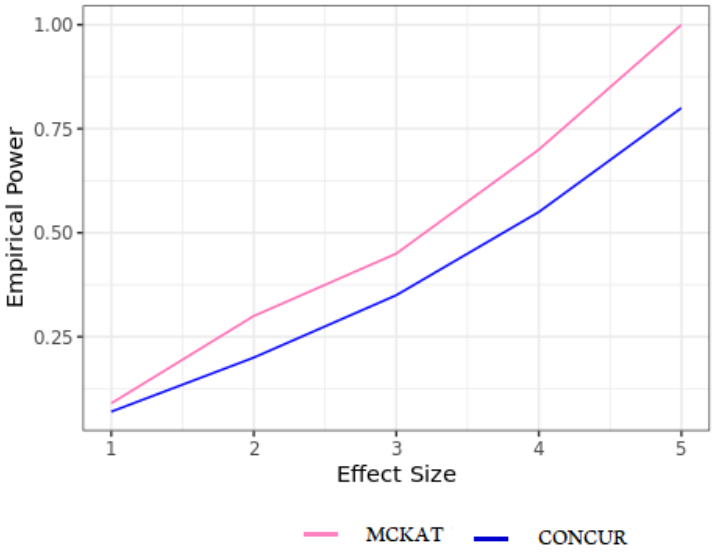

Figure 3.4: Empirical power of MCKAT and CONCUR under first simulation scenario, rare CNV data.

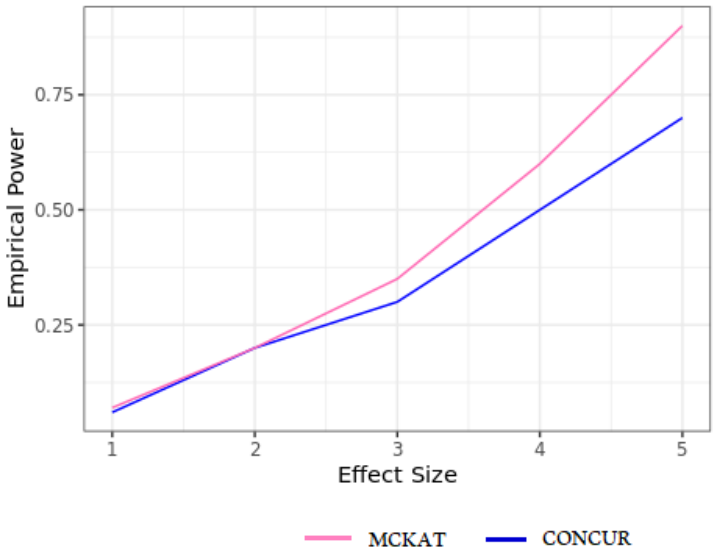

Figure 3.5: Empirical power of MCKAT and CONCUR under second simulation scenario, frequent CNV data.

3.4 Real Data Application Results

In real data applications we conduct the association test on autism and rhabdomyosarcoma datasets as representative examples of rare and red frequent datasets. First, we applied the MCKAT, CKAT and CONCUR on every chromosome to test if there is any association between CNVs and disease-related traits. Then, we partitioned the chromosomes into smaller regions called cytogenetic bands. We wanted to check if MCKAT can detect cytogenetic bands that CNVs on them are significantly associated with disease-related traits. The datasets and analysis results are described in the following.

3.4.1 Autism and Rhabdomyosarcoma Data

We apply MCKAT on both rare and frequent CNV public domain genome sequencing data sets to evaluate the performance. The two CNV datasets used in this study are from individuals with autism spectrum disorder (ASD) and rhabdomyosarcoma (RMS) cancer. The ASD data set contains a total of 2359 CNVs of 588 subjects publically available (Girirajan et al. 2011). Most of the

CNVs in the ASD data set are large and rare, while the RMS dataset contains common and small CNVs. The raw RMS dataset is publicly available through the National Institute of Health (NIH), database of Genotypes and Phenotypes (dbGaP). We use 59,131 processed whole-genome CNV data of 44 subjects (Shern et al. 2014). In both datasets, each CNV is presented by four characteristics: start and end position in the chromosome, type, and dosage. The type is either deletion or amplification, and the dosage is less than 2 for deletion and greater than 2 for amplification. MCKAT, CONCUR and CKAT are applied to the RMS and ASD CNV data.

3.4.2 Real Data Results

We conduct MCKAT analysis on each of the 23 chromosome pairs to test the association between CNVs in each chromosome and disease-related traits. The disease-related traits are cancer subtype and disease status in RMS and ASD CNV data sets, respectively. Then, we compare MCKAT results with those obtained from CKAT and CONCUR.

CNV Analysis on Rhabdomyosarcoma Data Set

First, we conduct the experiment on the RMS CNV data. The RMS occurs as two major histological subtypes: embryonal (ERMS) and alveolar (ARMS). The classification of the RMS subtype has a direct effect on the patients' treatment options. The RMS CNV data includes a total of 59,131 CNVs for 25 alveolar and 19 embryonal cancers. The p-values of MCKAT and CKAT are reported in Table 3.1. Bonferroni correction is used for adjusting the multiple testing to control the family-wise error rate (FWER) of $\alpha = 0.05$. Since 22 chromosomes and the sex chromosome are being tested, the p-value threshold for a whole-chromosome significance is calculated as $0.05/23 = 2.2 \times 10^{-3}$.

Based on the results reported in Table 3.1, MCKAT identifies CNVs in four chromosomes significantly associated with distinguishing RMS subtype at $FWER = 2.2 \times 10^{-3}$: chromosomes 2, 8, 11, and 13. These results are consistent with the existing biological knowledge, which shows the capability of the MCKAT in identifying chromosomes significantly associated with specific disease-related traits.

For example, El Demellawy et al. (2017) shows that RMS is associated with

Table 3.1: P-values of testing the association between RMS subtype and CNVs in each chromosome. (*) denotes significant association between RMS subtype and CNVs by MCKAT, CKAT and CONCUR, (#) denotes the total number of CNVs on that chromosome.

Chromosome	# CNVs	MCKAT	CKAT	CONCUR
chr1	4382	1.257×10^{-1}	4.427×10^{-1}	2.316×10^{-1}
chr2	5584	1.188×10^{-3} *	3.757×10^{-1}	1.512×10^{-2}
chr3	2925	1.424×10^{-1}	4.502×10^{-1}	2.275×10^{-1}
chr4	3068	4.606×10^{-1}	4.110×10^{-1}	4.319×10^{-1}
chr5	3237	7.607×10^{-2}	4.505×10^{-1}	6.512×10^{-2}
chr6	2777	5.054×10^{-1}	4.200×10^{-1}	3.749×10^{-1}
chr7	3549	4.421×10^{-1}	4.657×10^{-1}	4.021×10^{-1}
chr8	5365	4.308×10^{-7}*	4.064×10^{-1}	2.125×10^{-3}*
chr9	2474	5.666×10^{-2}	4.584×10^{-1}	4.925×10^{-2}
chr10	2378	9.667×10^{-2}	4.436×10^{-1}	5.041×10^{-2}
chr11	3449	1.107×10^{-3}*	3.655×10^{-1}	2.112×10^{-3}*
chr12	3773	3.638×10^{-1}	4.875×10^{-1}	3.825×10^{-1}
chr13	2462	1.241×10^{-3}*	3.916×10^{-1}	2.352×10^{-3}
chr14	1219	3.187×10^{-1}	4.613×10^{-1}	4.015×10^{-1}
chr15	1389	3.952×10^{-1}	4.659×10^{-1}	3.625×10^{-1}
chr16	1565	2.002×10^{-1}	4.960×10^{-1}	2.628×10^{-1}
chr17	1862	2.416×10^{-1}	4.658×10^{-1}	2.031×10^{-1}
chr18	1120	1.961×10^{-1}	4.717×10^{-1}	2.512×10^{-1}
chr19	1584	1.967×10^{-1}	4.948×10^{-1}	2.003×10^{-1}
chr20	1835	5.859×10^{-3}	4.237×10^{-1}	7.425×10^{-3}
chr21	648	3.531×10^{-2}	3.939×10^{-1}	4.107×10^{-2}
chr22	780	1.124×10^{-1}	4.327×10^{-1}	1.842×10^{-1}
chr X	1421	7.495×10^{-1}	4.917×10^{-1}	6.023×10^{-1}
chr Y	250	6.802×10^{-1}	4.755×10^{-1}	5.257×10^{-1}

specific chromosomal abnormalities that differentiate ARMS and ERMS. According to their study, approximately 80% of ARMS tumors show translocation between the *FOXO1* transcription factor gene located on chromosome 13 and the *PAX3* transcription factor gene on chromosome 2, and ERMS tumors demonstrate a higher frequency of specific genetic mutation on chromosome 11 compared with ARMS. The same has been revealed earlier in Sun et al. (2015). In addition to the association between chromosomal abnormalities on chromosomes 2, 11, and 13, Nishimura et al. (2013) found the ARMS subtype is significantly associated with amplifications on chromosome 8. Our findings show other mechanisms like

CNVs have the potential to play a significant role in causing any disease-related traits besides gene mutations and chromosomal translocations.

We apply CKAT and CONCUR on the RMS data set to compare their performance with MCKAT. As shown in Table 3.1, CKAT has low performance on the RMS data set, which includes common and small CNVs, and does not identify any chromosomes significantly associated with the RMS subtype. CKAT uses a parsimonious scanning algorithm to align pairs of CNVs based on their ordinal position. Using this strategy, each CNV is compared only with a limited number of adjacent CNVs resulting in sub optimal capture of the similarity between all possible CNV pairs. Furthermore, CKAT does not utilize CNV dosage and chromosomal position information in measuring the similarity between CNV profiles. CONCUR has better performance on the RMS data comparing with CKAT. It identifies CNVs in two chromosomes, 8 and 11, significantly associated with distinguishing RMS subtypes. These two chromosomes are identified by MCKAT as well. However, the MCKAT indicates stronger evidence and smaller p-value than CONCUR.

CNV Analysis on Autism Data Set

We apply MCKAT on the ASD data set to evaluate its performance on data sets that include large and rare CNVs. We aim to test if there is any association between CNVs and disease status. The ASD data set contains 1285 rare CNVs on 310 individuals with ASD and 1074 rare CNVs on 278 healthy individuals. Three factors characterize each CNV: the start and end chromosomal position and the type information.

As shown in Table 3.2, MCKAT, CKAT and CONCUR detect some chromosomes significantly associated with ASD status. The performance of MCKAT and CKAT are similar for the ASD dataset since this data set only contains rare and large CNVs. Therefore, the parsimonious scanning algorithm used in CKAT has a smaller adverse effect in measuring optimal similarity between CNV profiles. However, CKAT shows the similar low performance when the number of frequent CNVs are high in a chromosome like number 8. The performance of MCKAT and CONCUR are similar for the ASD dataset. However, MCKAT provides stronger evidence and a smaller p-value than CONCUR for the same chromosome. Among the detected chromosomes, all three methods, identify CNVs in chromosomes 3

and 22 as the most significant associated CNVs with ASD status. These results are consistent with previous biological studies, which identify chromosomes 3 and 22 being widely associated with the autism (Girirajan et al. 2011; Glessner et al. 2009; Freitag et al. 2010). A systematic review of the CNVs involved in ASD development which has the CNVs identifies by the MCKAT on other chromosomes 2, 11 and 16 has been done by Sener (2014).

Table 3.2: P-values of the testing association between ASD status and CNVs in each chromosome by MCKAT, CKAT and CONCUR. (*) denotes significant association between ASD and CNVs, (#) denotes the number of total CNVs on that chromosome.

Chromosome	# CNVs	MCKAT	CKAT	CONCUR
chr1	175	7.5×10^{-1}	8.2×10^{-2}	9.3×10^{-1}
chr2	45	2.3×10^{-5} *	1.7×10^{-4} *	2.1×10^{-3}*
chr3	49	0.0 *	0.0 *	1.5×10^{-3}*
chr4	112	7.5×10^{-1}	8.2×10^{-1}	8.3×10^{-1}
chr5	242	5.1×10^{-2}	2.3×10^{-2}	4.5×10^{-2}
chr6	17	2.9×10^{-3}	1.2×10^{-4} *	3.1×10^{-3}
chr7	25	1.0×10^{-1}	1.2×10^{-4} *	1.4×10^{-2}
chr8	3	2.6×10^{-1}	0.1×10^{-1}	2.0×10^{-1}
chr9	13	1.0×10^{-1}	7.7×10^{-1}	5.3×10^{-1}
chr10	130	4.3×10^{-1}	4.7×10^{-1}	4.9×10^{-1}
chr11	257	1.6×10^{-3}*	8.8×10^{-1}	2.1×10^{-3}*
chr12	3	3.8×10^{-1}	2.7×10^{-1}	4.2×10^{-1}
chr13	5	4.2×10^{-1}	7.4×10^{-1}	5.3×10^{-1}
chr14	2	4.0×10^{-1}	1.8×10^{-1}	3.3×10^{-1}
chr15	919	4.0×10^{-1}	5.4×10^{-1}	4.5×10^{-1}
chr16	140	1.7×10^{-3}*	3.7×10^{-1}	2.1×10^{-3}*
chr17	27	2.8×10^{-2}	2.3×10^{-3}	2.4×10^{-3}
chr18	6	4.2×10^{-1}	1.0	1.0
chr19	1584	1.9×10^{-1}	4.9×10^{-1}	2.3×10^{-1}
chr20	17	4.4×10^{-1}	1.3×10^{-1}	3.5×10^{-1}
chr21	0	1.0	1.0	1.0
chr22	166	0.0 *	0.0 *	1.2×10^{-4}*
chr X	2	3.2×10^{-1}	1.4×10^{-2}	4.5×10^{-1}
chr Y	1	2.9×10^{-1}	2.9×10^{-1}	3.1×10^{-1}

CNV Analysis on Cytogenetic Bands in RMS

We partitioned each chromosome into smaller regions based on the cytogenetic

bands. We applied MCKAT on each chromosome band to check if MCKAT is capable of detecting more specific regions rather than whole chromosomes. Figure 3.6 shows the significance level of all cytogenetic bands across each chromosome in the RMS dataset. We consider the p-value threshold for each chromosome as 2.1×10^{-3} calculated by dividing 0.05 by 23 chromosomes. CNVs within the bands with a calculated p-value above this threshold have a statistically significant association with the two main RMS subtypes. As is shown in figure 3.6 there are 22 cytogenetic bands across the genome, specifically across chromosomes 2, 8, 11, and 13, that CNVs in these bands are significantly associated with the RMS subtype.

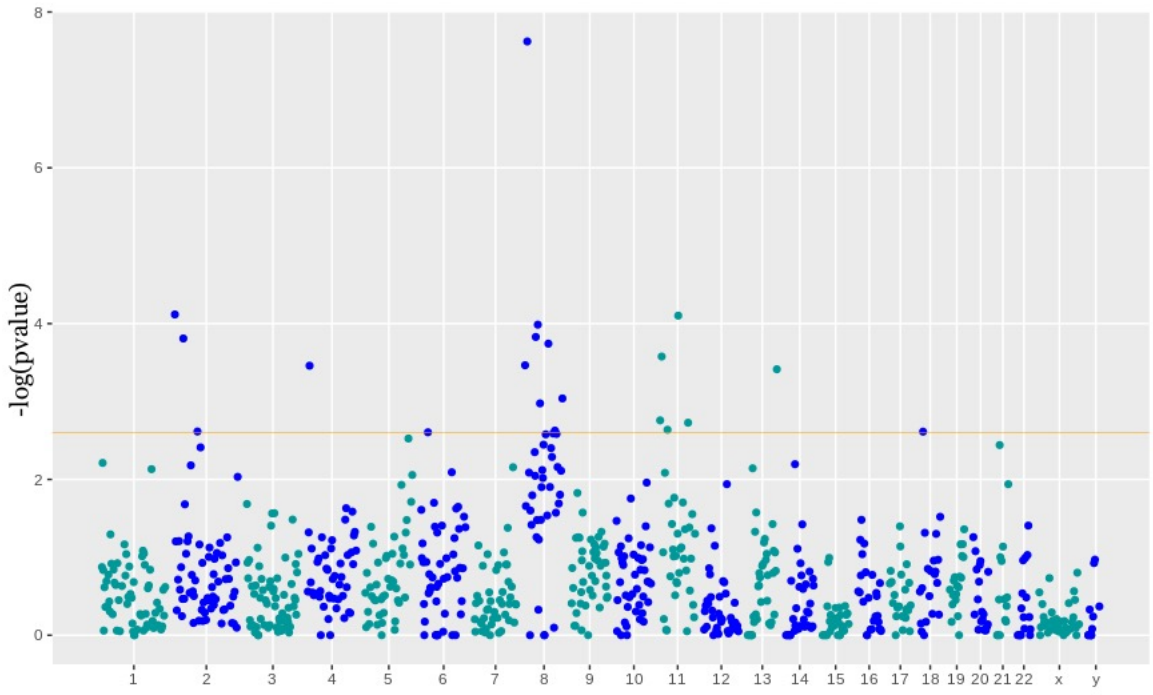

Figure 3.6: Manhattan plot showing the -log(pvalue) of testing association between CNVs on the chromosome cytogenetic bands and RMS sub types. Those with -log(pvalue) above the threshold line, are significantly associated with the RMS subtype

We use chromosomal ideograms to visualize the chromosomal position of these 22 cytogenetic bands identified as significantly associated with the RMS subtype. In Figure 3.7, we plot the calculated p-values against cytogenetic bands. It includes

the five identified significant chromosomes shown in Figure 3.6: 2, 4, 8, 11, and 13. The CNVs within the bands with a p-value that passes the threshold are significantly able to distinguish the RMS subtype.

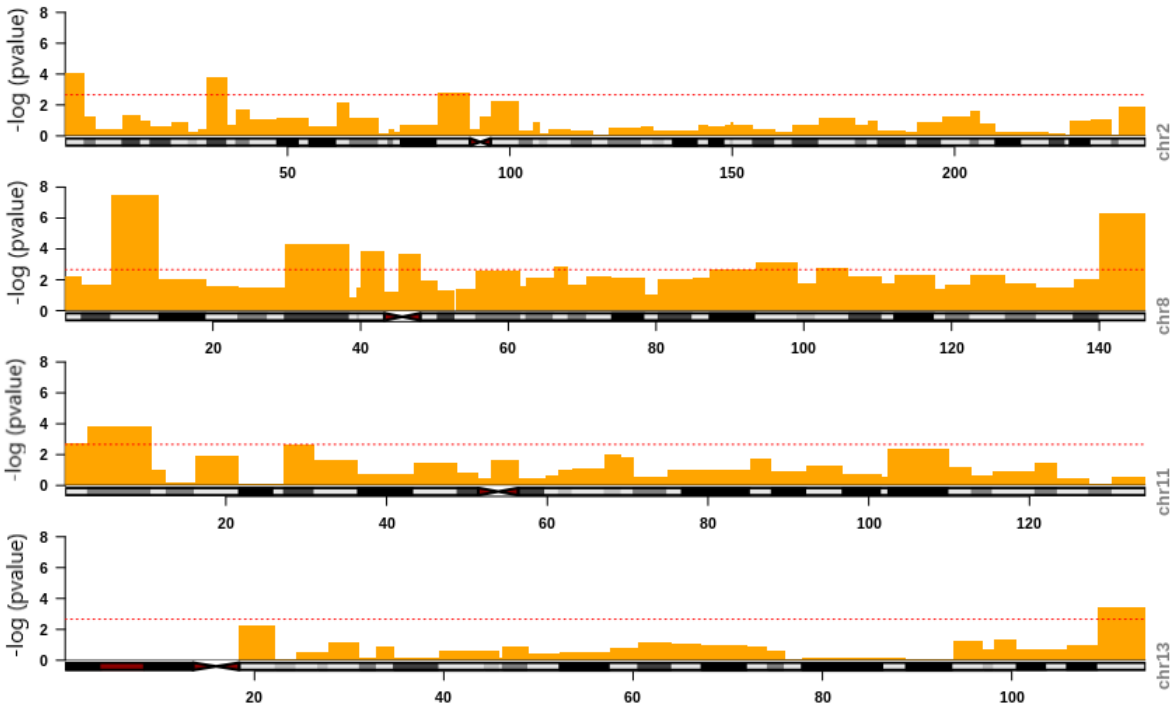

Figure 3.7: Chromosomal ideograms showing statistically significant cytogenetic bands that CNVs on them are associated with the RMS subtype for chromosomes 2, 8, 11 and 13.

More details regarding the p-values of these cyctogenetic bands are reported in Table 3.3 and Table 3.4. Table 3.3 contains the p-values of the association test between the RMS subtype and CNVs in each cytogenetic bands in chromosome 8. We chose Chromosome 8 since it was identified as the most significant one with the lowest p-value of 4.308×10^{-7} based on the results reported in Table 3.1.

Table 3.4 contains the p-values of all other bands across the genome that are identified as significantly associated with the RMS subtype.

Figures 3.8 to 3.13 shows the chromosomal ideograms for the other chromosomes where CNVs were not identified as statistically significantly associated with the RMS subtype by the MCKAT. As shown in these figures, none of the bands

Table 3.3: P-values of the testing association between RMS subtype and CNVs in each cytogenetic bands of chromosome 8 by MCKAT. (*) denotes significant association between RMS subtype and CNVs, (#) denotes the number of total CNVs on the band.

Arm	Band	Start	Stop	#CNVs	P-value
p	23.3	1	2,300,000	113	3.4×10^{-4} *
p	23.2	2,300,001	6,300,000	85	2.0×10^{-2}
p	23.1	6,300,001	12,800,000	304	4.7×10^{-8} *
p	22.0	12,800,001	19,200,000	101	8.2×10^{-3}
p	21.3	19,200,001	23,500,000	102	2.5×10^{-2}
p	21.2	23,500,001	27,500,000	82	3.6×10^{-2}
p	21.1	27,500,001	29,000,000	50	1.6×10^{-2}
p	12.0	29,000,001	36,700,000	190	3.7×10^{-5} *
p	11.23	36,700,001	38,500,000	48	3.7×10^{-3}
p	11.22	38,500,001	39,900,000	57	8.4×10^{-3}
p	11.21	39,900,001	43,200,000	147	1.0×10^{-3}
p	11.1	43,200,001	45,200,000	72	2.8×10^{-2}
q	11.1	45,200,001	47,200,000	41	2.1×10^{-2}
q	11.21	47,200,001	51,300,000	200	8.4×10^{-5} *
q	11.22	51,300,001	51,700,000	6	4.7×10^{-2}
q	11.23	51,700,001	54,600,000	61	6.1×10^{-2}
q	12.1	54,600,001	60,600,000	177	7.0×10^{-4} *
q	12.2	60,600,001	61,300,000	18	3.3×10^{-2}
q	12.3	61,300,001	65,100,000	134	1.1×10^{-2}
q	13.1	65,100,001	67,100,000	71	5.8×10^{-3}
q	13.2	67,100,001	69,600,000	54	4.3×10^{-3}
q	13.3	69,600,001	72,000,000	62	1.8×10^{-3}
q	21.11	72,000,001	74,600,000	144	8.4×10^{-3}
q	21.12	74,600,001	74,700,000	1	1.0
q	21.13	74,700,001	83,500,000	308	2.6×10^{-3}*
q	21.2	83,500,001	85,900,000	56	2.9×10^{-2}
q	21.3	85,900,001	92,300,000	185	1.0×10^{-4} *
q	22.1	92,300,001	97,900,000	182	1.0×10^{-2}
q	22.2	97,900,001	100,500,000	103	3.9×10^{-3}
q	22.3	100,500,001	105,100,000	162	4.6×10^{-3}
q	23.1	105,100,001	109,500,000	135	2.5×10^{-3}*
q	23.2	109,500,001	111,100,000	33	8.0×10^{-1}
q	23.3	111,100,001	116,700,000	185	2.3×10^{-3}*
q	24.11	116,700,001	118,300,000	53	2.6×10^{-2}
q	24.12	118,300,001	121,500,000	109	2.2×10^{-3}*
q	24.13	121,500,001	126,300,000	151	6.0×10^{-3}
q	24.21	126,300,001	130,400,000	208	1.9×10^{-2}
q	24.22	130,400,001	135,400,000	155	1.5×10^{-2}
q	24.23	135,400,001	138,900,000	162	7.7×10^{-3}
q	24.3	138,900,001	145,138,636	354	2.5×10^{-8} *

Table 3.4: The cytogenetic bands across the whole genome identified as significantly associated with the RMS subtype by MCKAT. (#) denotes the number of CNVs on the band.

Chr.	Arm	Band	Start	Stop	#CNVs	P-value
2	p	25.3	1	4,400,000	111	1.0×10^{-4}
2	p	22.3	31,800,000	36,300,000	117	1.0×10^{-4}
2	p	11.2	83,100,001	91,800,000	314	2.0×10^{-4}
11	p	15.5	1	2,800,000	304	4.7×10^{-8}
11	p	15.4	2,800,001	11,700,000	269	3.0×10^{-4}
11	q	14.1	27,200,001	31,000,000	100	2.0×10^{-4}
11	q	13.3	68,700,001	70,500,000	46	1.0×10^{-4}
11	q	22.3	103,000,001	110,600,000	145	1.9×10^{-3}
13	q	34.0	109,600,001	114,364,328	115	4.0×10^{-4}

in these chromosomes has a p-value that passes the threshold which means CNVs on them are not able to distinguish the RMS subtype statistically.

We form a new CNV profile for each subject for more investigation. These new CNV profiles include only CNVs in the 22 cytogenetic bands that have been identified significantly associated with RMS subtype shown in Tables 3.3 and 3.4. Then, we applied the MCKAT on these manually created CNV profiles. Based on the results, the combination of CNVs located in these bands has a statistically higher significant association with the RMS subtype of p-value equals to zero. This finding shows the combination of CNVs in cytogenetic bands that have been identified significantly associated with the RMS subtype has a high potential to be used in RMS subtype identification. We do not do cytogenetic band analysis on the ASD CNV data set, since it contains only rare CNVs and the frequency of the CNVs in chromosomes is too low which results in no existence of any CNVs on cytogetetic bands for most of the subjects.

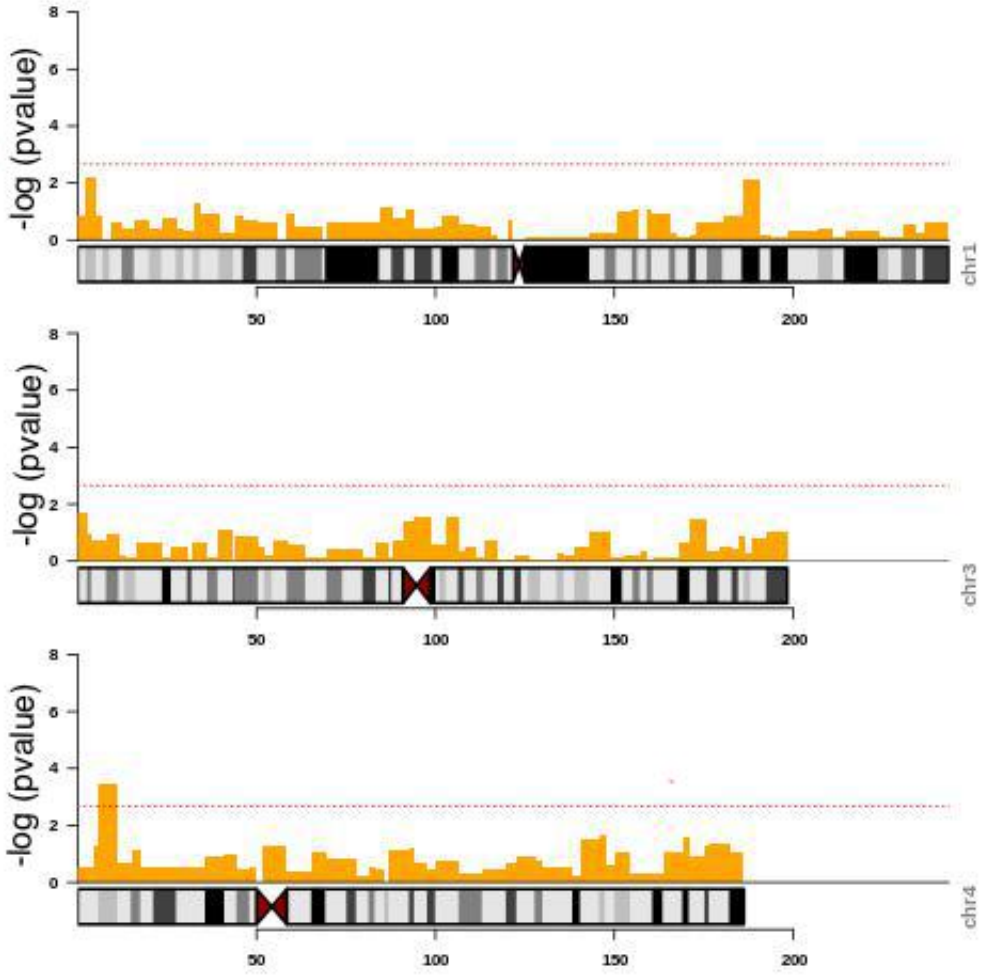

Figure 3.8: Chromosomal ideograms showing not statistically significant associated CNVs with the RMS subtype on cytogenetic bands for chromosomes 1, 3 and 4.

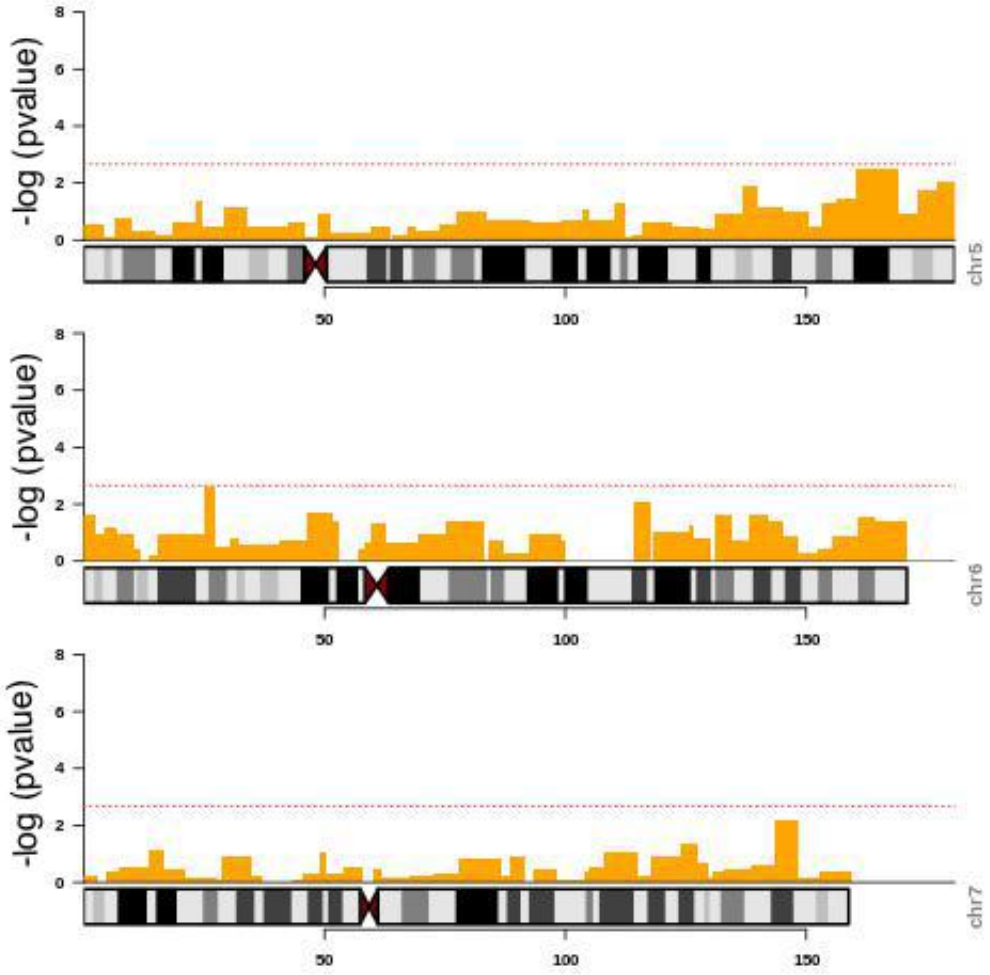

Figure 3.9: Chromosomal ideograms showing not statistically significant associated CNVs with the RMS subtype on cytogenetic bands for chromosomes 5, 6 and 7.

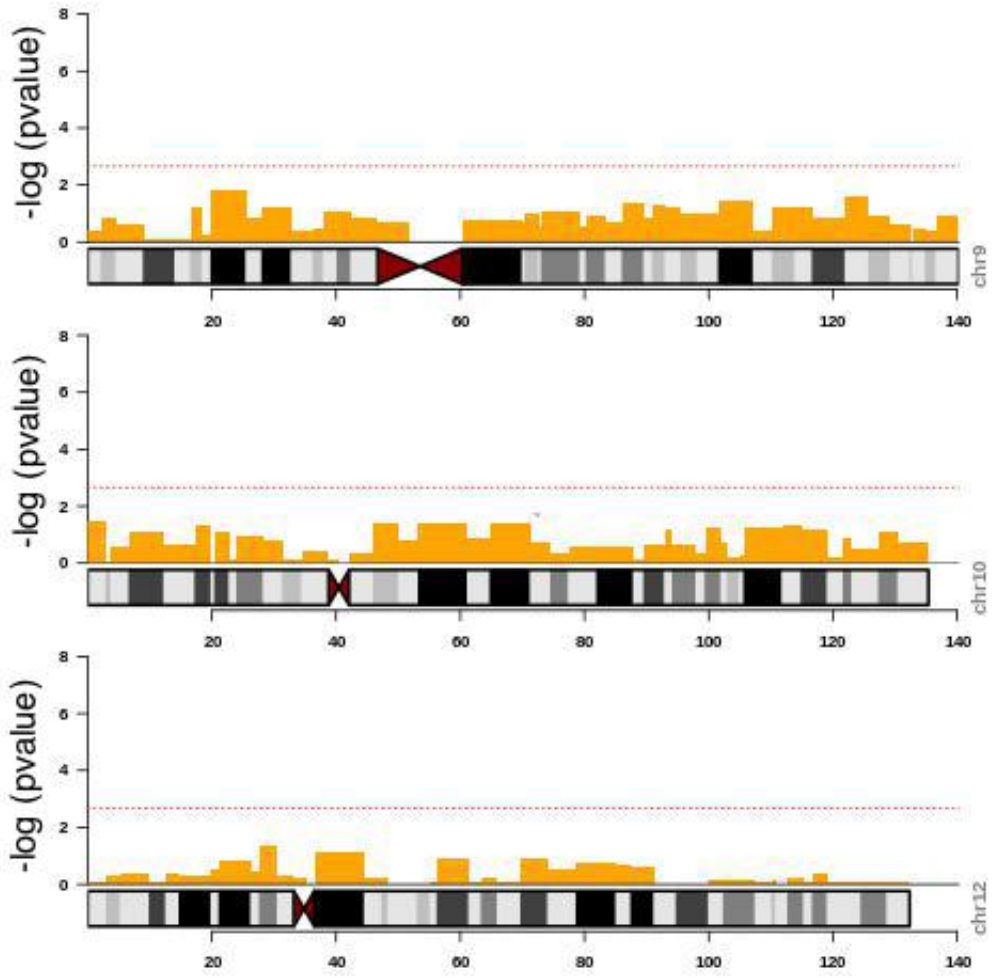

Figure 3.10: Chromosomal ideograms showing not statistically significant associated CNVs with the RMS subtype on cytogenetic bands for chromosomes 9, 10 and 12.

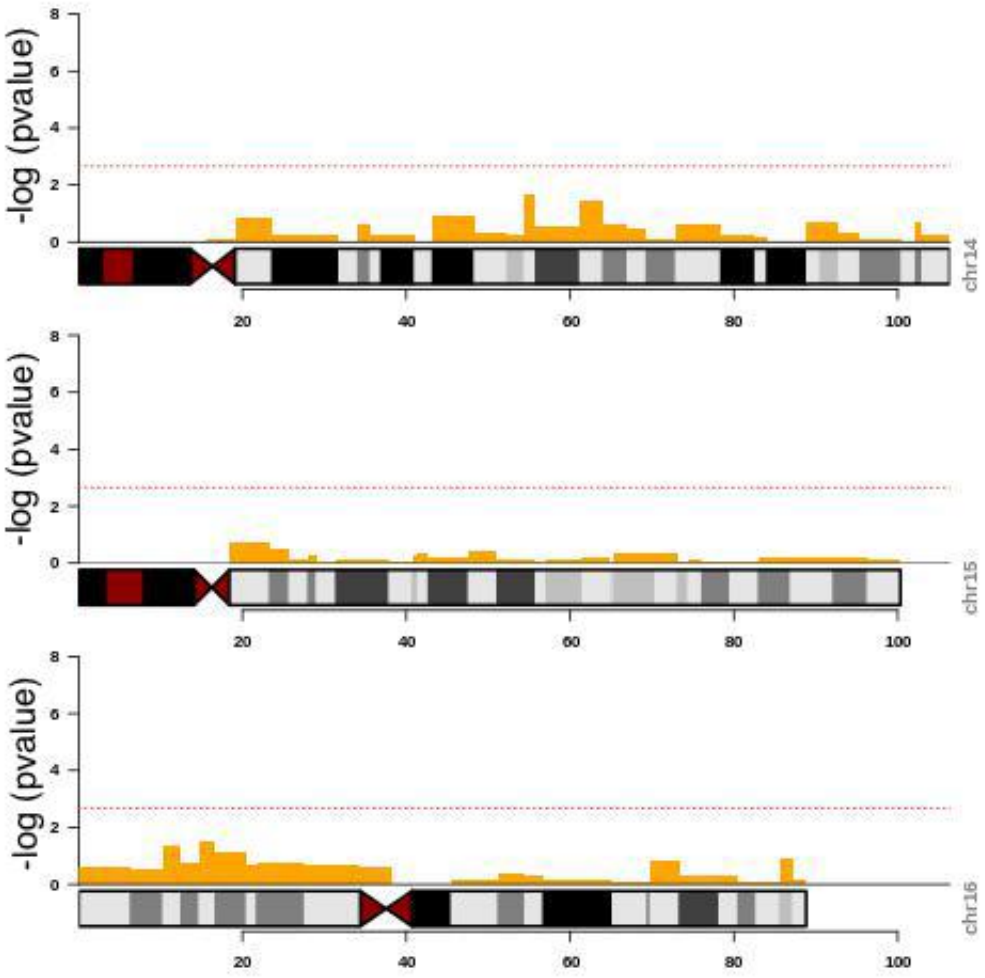

Figure 3.11: Chromosomal ideograms showing not statistically significant associated CNVs with the RMS subtype on cytogenetic bands for chromosomes 14, 15 and 16.

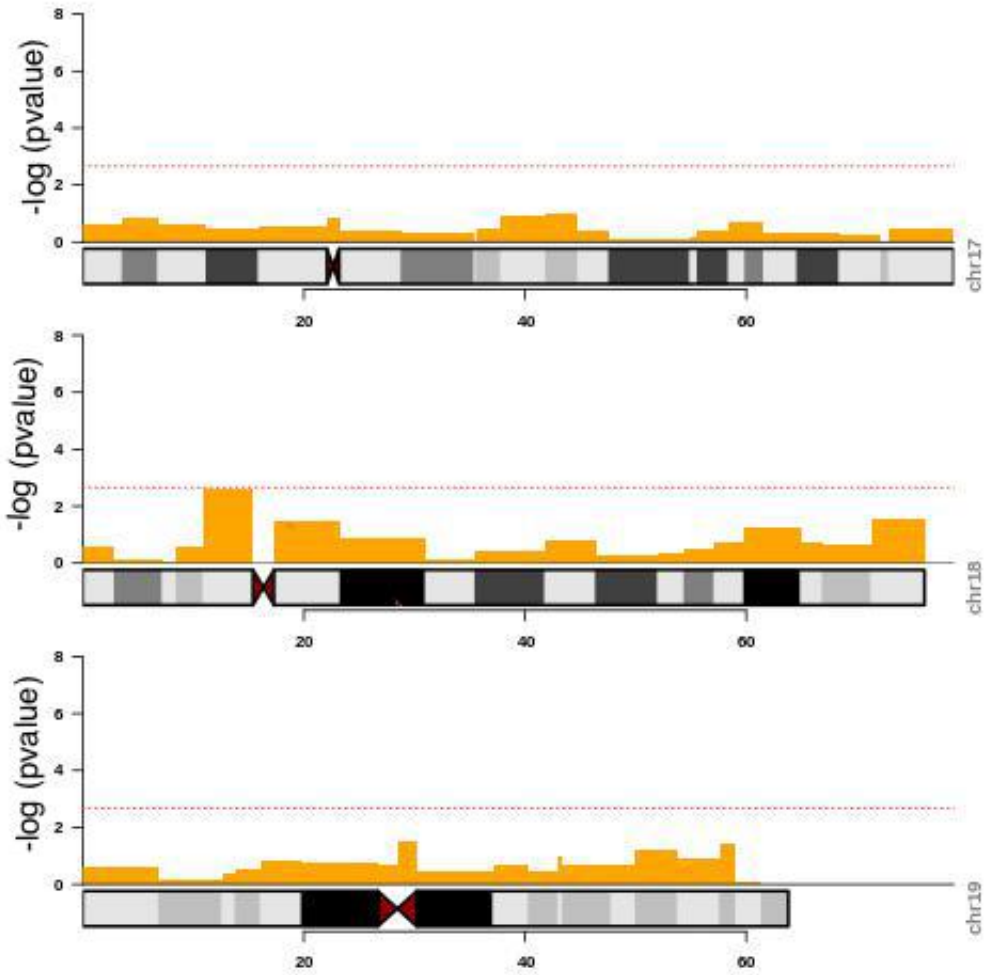

Figure 3.12: Chromosomal ideograms showing not statistically significant associated CNVs with the RMS subtype on cytogenetic bands for chromosomes 17, 18 and 19.

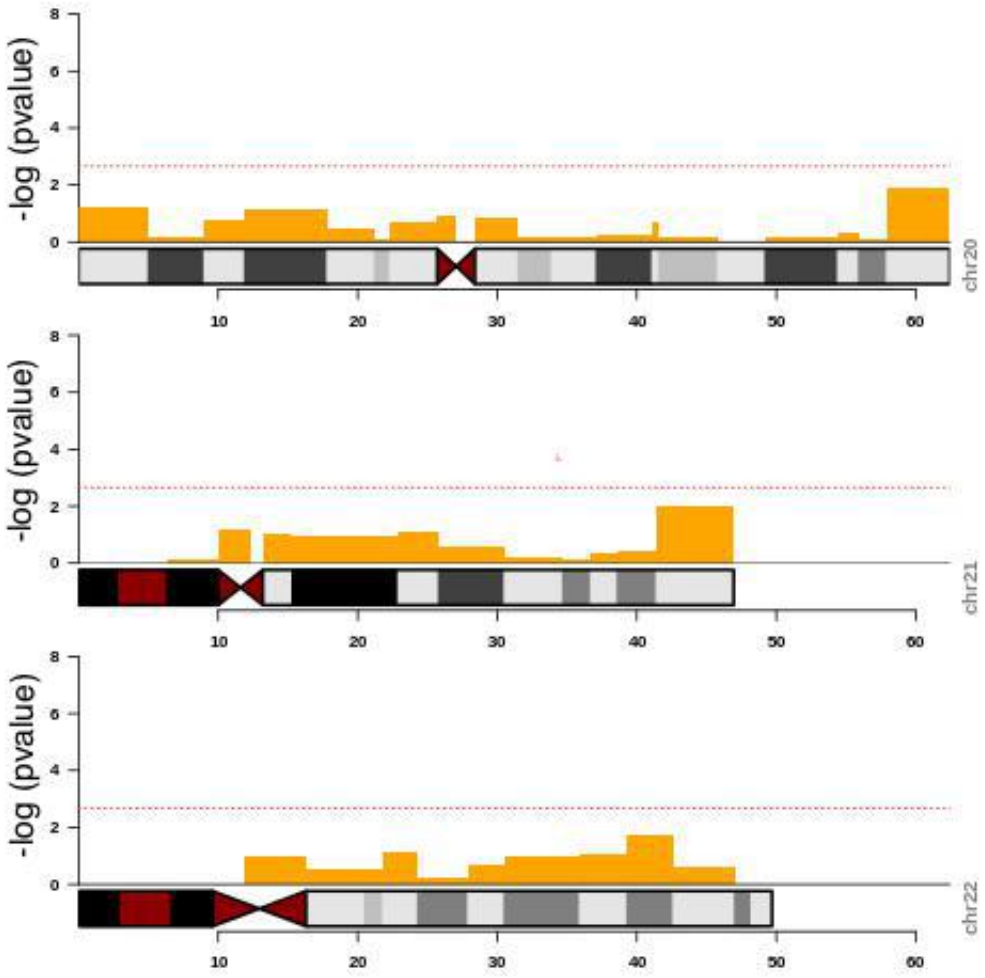

Figure 3.13: Chromosomal ideograms showing not statistically significant associated CNVs with the RMS subtype on cytogenetic bands for chromosomes 20, 21 and 22.

3.5 Discussion

MCKAT is an advanced approach to test the association between CNVs and disease-related traits. Our approach has several advantages over the existing methods. Firstly, as the CNVs have more complicated multi-dimensional features in comparison with other types of genetic variants like SNPs, this is the first time that all multi-dimensional features, including chromosomal position, type, dosage, and heterogeneity effect of the CNVs are utilized in testing the association between CNVs and disease-related traits.

Secondly, the previous kernel-based methods do not measure the similarity between CNV profiles in an optimal way due to deficiencies in the algorithm they used to pair CNVs. In our proposed approach, we measure the similarity between CNVs profiles in an optimal way by considering the similarity between all possible CNV pairs in two CNV profiles.

Thirdly, as the result of two aforementioned advantages, MCKAT provides smaller p-values compared to the other methods. As the p-value is a statistical measure used in hypothesis testing, it quantifies the strength of evidence against a null hypothesis. The smaller p-value can provide strong evidence to reject the null hypothesis and accept the alternative one. In almost all experiments that we have done to evaluate the performance of MCKAT, it provides us with the smaller p-values, stronger evidence, to check the association between CNVs and disease related traits. However, this improvement varies from one CNV data set to another one and can not be quantified as a constant value for all experiments.

Fourthly, the previous methods can only deal with a limited number of CNVs in chromosomal regions or rare CNV datasets. The results show that MCKAT is applicable to not only rare and large CNVs but also common and small CNVs.

Finally, MCKAT can help biologists detect significantly associated CNVs with any disease-related trait across a patient group instead of examining the CNVs case by case in each subject.

Although our experimental results are promising and based on these results MCKAT shows a better performance in most cases compared to the state-of-the-art CNV kernel approach, this study has limitations. There are not many publicly available CNV data sets. Besides, most available ones do not contain all CNV features together, in particular the dosage information. Consequently, our method is tested only on few publicly available datasets that include all

multi-dimensional CNV characteristics. However, for some of them, we had to simulates CNV features to be able to evaluate the performance of our proposed method. Applying MCKAT to more datasets containing all CNV features can help to determine its strengths and weakness.

Our study shows that CNVs in some chromosomal regions can have statistically significant association with disease-related traits, but it has the potential to reveal more new findings by conducting more comprehensive analysis.

3.6 Conclusion

The results presented in this chapter demonstrate that our method, MCKAT, provides improved outcomes for detecting significant associations between CNVs, both rare and frequent, and disease-related traits by indicating stronger evidence and smaller p-values than the existing methods. MCKAT can provide biologists with CNV hot spots on the genome at the cytogenetic band level for further investigation. Therefore, instead of examining all chromosomal regions across the genome which includes an enormous number of CNVs, they can narrow down their research to the identified regions by the MCKAT. This work forms Contribution 1 of this thesis which is proposing a multi-dimensional CNV kernel-based association test that allows for the detection of CNV chromosomal regions significantly associated with disease-related traits and improves on currently available methods for studying CNVs.

The method developed here allows for the more detailed exploration of chromosomal regions that CNVs on them have statistically significant association with disease-related traits. One such exploration is investigating the dual effect of the CNVs and their intersected genes on the disease development. This exploration is detailed in Chapter 5 of this thesis.

In the next chapter we will investigate if the CNV sequential order has any significant associations with disease-related traits by proposing a sequential multi-dimensional CNV kernel based association test.

Chapter 4

SMCKAT, a sequential multi-dimensional copy number variant kernel association test

4.1 Introduction

As discussed in section 2.5, there is an open question around the association of CNV sequential order and disease-related traits. More specifically, to our knowledge, it is still unclear whether CNVs are randomly distributed across the genome, or if their order is significant and associated with the disease, like SNPs. Consequently, we propose the first such method to test the association between the sequential order of CNVs and diseases.

Starting from MCKAT, we propose a sequential multi-dimensional CNV kernel-based association test (SMCKAT) for investigating the association between CNVs and disease or traits. SMCKAT is not only utilizing all multi-dimensional characteristics of CNVs but also the sequential order of CNVs in testing the association between CNVs and disease or traits. Based on the results in this chapter, SMCKAT is applicable on both rare and common datasets and is capable of identifying hot-spots on the genome where both CNV characteristics and the CNV sequential order are significantly associated with disease or traits.

The SMCKAT algorithm is described in Section 4.2, including the specification of the kernels and the kernel-based association test. The performance evaluation of the SMCKAT by using simulated data is described in Section 4.3. The results on real data are presented in Section 4.4. The results are discussed in Section 4.5

and Section 4.6 summarises and concludes the chapter.

The R code used to implement the SMCKAT is given in Appendix B and is available at https://github.com/nesfehani/SMCKAT. The work in this chapter addresses Contribution 2 listed in the Chapter 1, proposing an association test to investigate if CNVs' order matters and has a significant association with disease-related traits. The contribution is addressed by developing a sequential multi dimensional kernel based association test (SMCKAT) which is included in published article (Maus Esfahani et al. 2021a).

4.2 Model Development

We design a sequential multi-dimensional kernel framework capable of measuring the similarity between CNV profiles utilizing all CNV characteristics and the CNV sequential order. It contains two kernels. The first kernel, the pair group kernel, measures the similarity between two groups of CNVs at the same ordinal position of CNV profiles. It contains three sub-kernels. Each sub-kernel is responsible for measuring the similarity between two CNVs with respect to one of the three CNV characteristics. The second kernel, the whole genome group kernel, aggregates the similarity between every possible CNV pair group to measure the total similarity between the CNV profiles of the subjects. Finally, the association between CNV sequential order across a chromosome and disease-related traits is tested by comparing the similarity in CNV profiles to that in the trait using an association test. The SMCKAT workflow is summarized in Figure 4.1.

4.2.1 Pair CNV Group Kernel

AS in Chapter 3, let X denote a single CNV which is defined by four characteristics as $X = (X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)})$ where $X^{(1)}$ and $X^{(2)}$ are the CNV starting and ending positions on the chromosome, $X^{(3)}$ is the CNV type, and $X^{(4)}$ is the CNV dosage. First, we generate the CNV profile R for subject i with l CNVs as $R_i = (X_1^i, X_2^i, ..., X_{l_i}^i)$ where CNVs are sorted based on their chromosomal position. Secondly, we extract a CNV group of size n out of the CNV profile as $G_i = (X_m^i, X_{m+1}^i, ..., X_{m+n}^i)$ where m is the CNV of the first place in CNV group G and n is the group size that can take any value between 1 and l, the number of

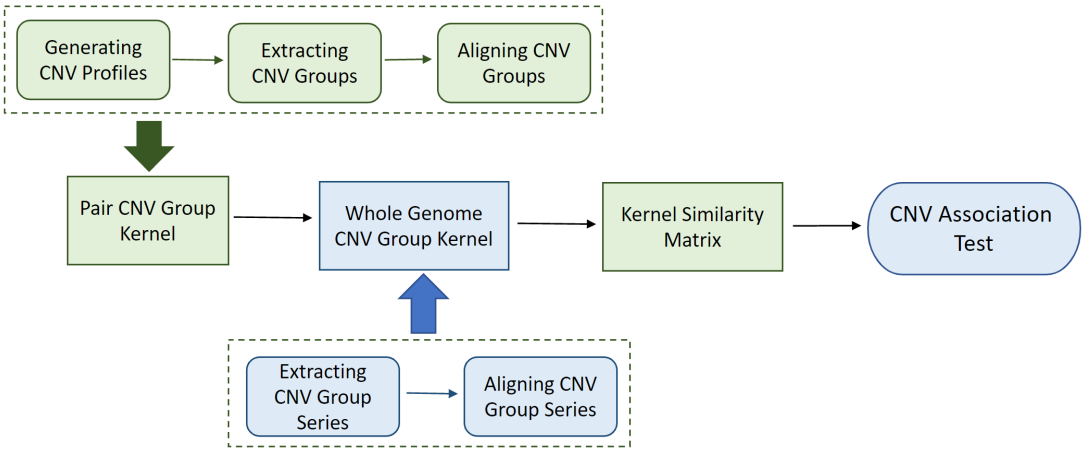

Figure 4.1: SMCKAT workflow diagram. Firstly, preparing CNV groups for each CNV profiles and aligning relevant CNV groups of each subject. Secondly, measuring the similarity between CNV groups by the pair CNV group kernel. Thirdly, extracting CNV group series for each subject and measuring the similarity between all CNV profiles by the whole genome CNV group kernel. Finally, testing the association between CNV characteristics and sequential order with disease-related traits.

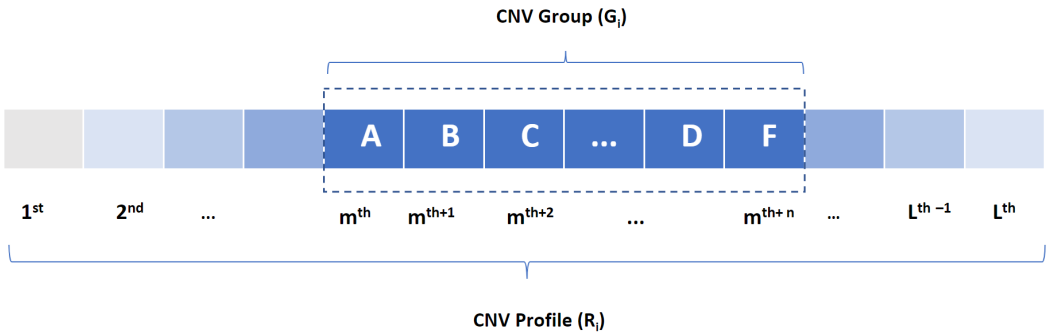

Figure 4.2: Generating CNV profile R_i where CNVs are sorted with respect to their chromosomal position. A, B,..., and F are arbitrary CNVs at m^{th}, m^{th+1}, ..., and m^{th+n} positions and G_i is a group of CNVs of size n.

existing CNVs in a CNV profile as is shown in Fig. 4.2.

We propose a pair CNV group kernel, K_{PG}, to measure the similarity between two CNV groups of size n, G_i and G_j, in two CNV profiles. First, K_{PG} aligns each CNV in the G_i with its relevant CNV in the G_j with respect to their position to generate n CNV pairs as is shown in Fig. 4.3.

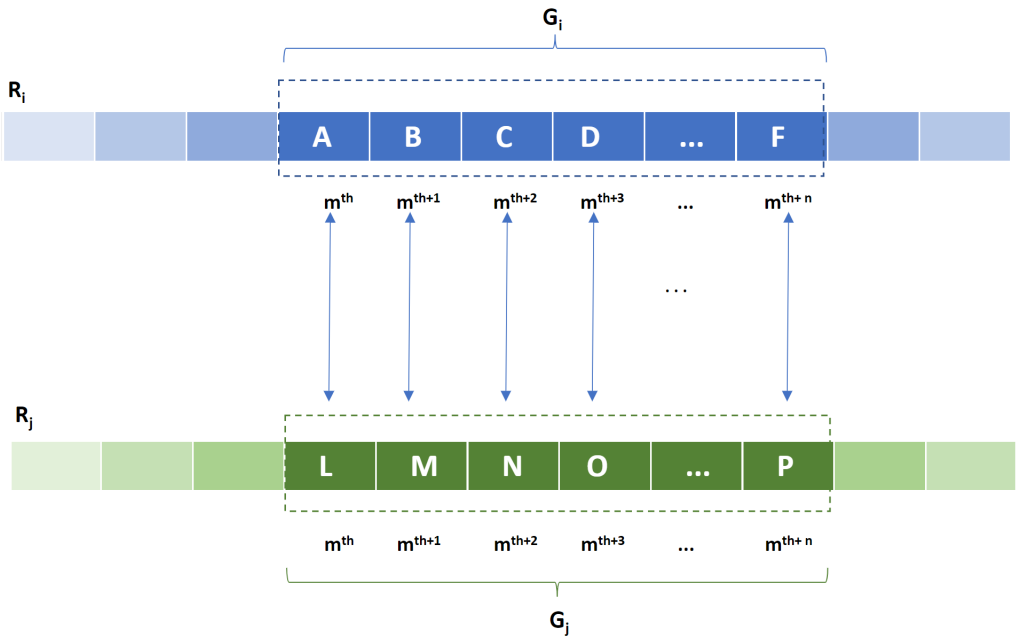

Figure 4.3: Aligning CNVs within two CNV groups of size n, G_i and G_j, to generate n CNV pairs.

Then, K_{PG} measures the similarity between each CNV pair using the single pair CNV kernel, K_S, we proposed in (Maus Esfahani et al. 2021b). K_S measures the similarity between a CNV pair by three sub-kernels considering all CNV features including chromosomal position, type and dosage. Finally, K_{PG} averages the similarities calculated by K_S between all generated CNV pairs to measure the similarity between two CNV groups, G_i and G_j, as

$$K_{PG}(G_i, G_j) = \sum_{m=1}^{n} \frac{K_s(X_m^i, X_m^j)}{n} \tag{4.1}$$

where K_s is defined in (Maus Esfahani et al. 2021b) as

$$K_s\left(X_m^i, X_m^j\right) = \left[\frac{Intersection\left(\left(X_m^{i(1)}, X_m^{i(2)}\right), \left(X_m^{j(1)}, X_m^{j(2)}\right)\right)}{Union\left(\left(X_m^{i(1)}, X_m^{i(2)}\right), \left(X_m^{j(1)}, X_m^{j(2)}\right)\right)}\right] \times \left[\frac{\left(X_m^{i(3)} == X_m^{j(3)}\right) + 1}{2}\right] \times \left[\frac{1}{2^{\left|DR\left(X_m^{i(4)}\right) - DR\left(X_m^{j(4)}\right)\right|}}\right] \tag{4.2}$$

the first term measures the mutual presence of a CNV with a specific start and end position by dividing the size of the intersection of two CNVs to their union size. The intersection function calculates the length of the chromosomal region that belongs to both CNVs. Similarly, the union function calculates the length of the chromosomal region that consists of both regions that belong to the first CNV and to the second CNV. The second term compares the CNV type of two CNVs to calculate the similarity between them. The third term measures the similarity between two CNVs with respect to their dosage. The DR is the difference from the reference function we proposed in Chapter 3 as $DR(dosage) = |dosage - 2|$. DR measures the difference between a CNV dosage and the reference dosage value 2.

4.2.2 Whole Genome CNV Group Kernel

Next, we create a window of size n which is the size of the CNV groups. We slide this window across the CNV profile R_i as is shown in Fig. 4.4 to extract all possible CNV groups of size n as $P_i = (G_1^i, ..., G_{pi}^i)$ where CNV groups are sorted based on their position and pi is the number of extracted CNV groups for the CNV profile R_i. Similarly, we have another CNV group series $P_j = (G_1^j, ..., G_{qj}^j)$ for CNV profile R_j.

Then, we propose the whole genome CNV group kernel, K_{WG}, to measure the similarity between two CNV group series P_i and P_j as

$$K_{WG}(P_i, P_j) = \begin{cases} 0 & \text{if } pi \times qi = 0 \\ \sum_{z=1}^{max(pi,qi)} max\left(K_{PG}(G_z^i, G_{z-1}^j), K_{PG}(G_z^i, G_z^j), K_{PG}(G_z^i, G_{z+1}^j)\right) & \text{if } pi \times qi \neq 0 \end{cases} \tag{4.3}$$

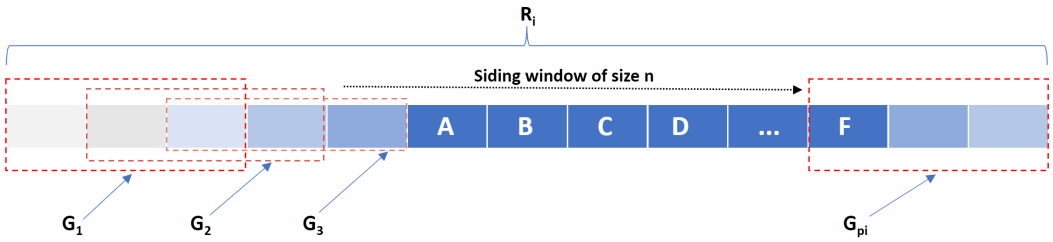

Figure 4.4: Sliding window of size n across CNV profile to extract CNV groups of size n.

where $K_{PG}(\cdot,\cdot)$ is the pair CNV group kernel from Equation 4.1. K_{WG} measures the similarity between the pair of CNV groups at the same position and aggregates these similarities to calculate the similarity in two CNV group series. The second maximum operation in the definition of K_{WG} searches for the the group pair among the existing the three pairs which has the highest similarity to align CNV groups in two CNV group series as shown in Fig. 4.5.

The kernel-based association test described in the following section, requires a kernel similarity matrix K. K is a $d \times d$ matrix, where $K_{ij} = K_{WG}(P_i, P_j)$ and d is the number of existing CNV profiles. K_{ij} expresses the similarity between CNV profile i and j measured by K_{WG}.

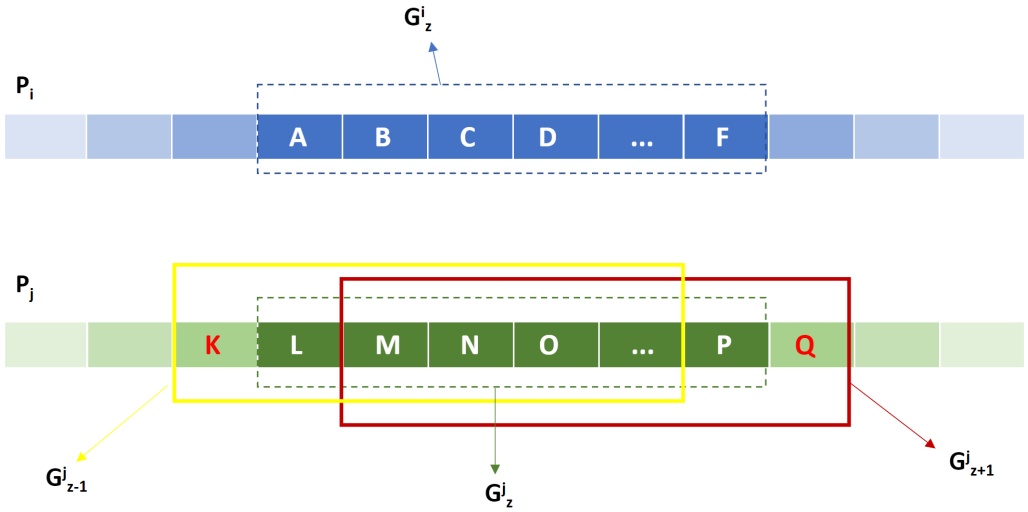

Figure 4.5: Aligning G_z^i to either of G_{z-1}^j, G_z^j or G_{z+1}^j of the highest similarity.

4.2.3 Kernel-based Association Test

We use the following logistic regression model to test the association between CNV sequential order and a disease-related trait

$$logit[Pr(y_i = 1)] = \beta_0 + Z\beta + f(P_i) \tag{4.4}$$

where y_i is the status of the disease-related trait with $y_i = 1$ denoting the existence of the trait and $y_i = 0$ denoting otherwise, and i is indexing the CNV profiles, and Z is the covariate matrix including information such as age and gender. P_i is the CNV group series of the profile R_i as explained previously. $f(\cdot)$ is a function spanned by the whole genome CNV group kernel $K_{WG}(\cdot, \cdot)$. According to equation (4.4), the hypothesis of no association between the CNV sequential order and the existence of a disease-related trait can be tested as $H_0 : f(\cdot) = 0$. To test this, one way is to treat the $f(\cdot)$ as a random effect vector which is distributed as $N(0, \tau K)$, where $\tau \geq 0$ and K is the $d \times d$ similarity matrix, treated as covariance matrix of the random effect, generated by K_{WG} as defined in Zhan et al. (2016). Liu et al. (2008) has shown that testing $H_0 : f(\cdot)$ is equivalent to testing $H_0 : \tau = 0$ in the logistic mixed effect model. Moreover, τ is a variance component parameter in the logistic mixed effect model, which can be tested using a restricted maximum likelihood-based score test Liu et al. (2008); Wu et al. (2010).

We use the following score test statistic where $\widehat{y}$ is estimated under the null model $logit[Pr(y_i = 1)] = \beta_0 + Z\beta$ and K is the similarity matrix explained in the previous section.

$$Q = (y - \widehat{y})' K (y - \widehat{y}) \tag{4.5}$$

Then, we used the Davies method (Davies 1980) as implemented in the CKAT R package (Zhan et al. 2016) to calculate the p-value of the proposed kernel based association test.

4.3 Model Evaluation using Simulated Data

As done in Chapter 3, we conduct simulations to evaluate the performance of SMCKAT and to ensure that it can properly handle type I and II errors as well as having relatively high power in detecting existing associations. Besides SMCKAT, MCKAT, CONCUR and CKAT are also studied. We conduct our simulation studies under two main scenarios. In the first scenario, we evaluate the

performance of the SMCKAT on the rare CNV data. In the second scenario, we evaluate the performance of the SMCKAT on the common CNV data.

We use the ASD dataset and the RMS dataset in the first and second simulation scenarios respectively. These datasets are studied in the real data analysis and further details regarding them are shared in the section 3.4.1. We simulated 10^5 datasets for each simulation scenario.

The ASD dataset has the same dosage value for all deletions and similarly the same dosage value for all amplifications. Therefore, other values for the CNV dosage are randomly generated to conduct our simulation studies and investigate the dosage effect in identifying existing associations. The simulated dosage value can take 0 or 1 for deletion types and 3, 4, ..., 7 for amplification types. We use uniform probabilities when generating random dosage values for deletion and amplification, 0.5 and 0.2 respectively, for the two and five respective values the dosage may take.

A case-control phenotype is generated for both SMCKAT and MCKAT from the following logistic model that we proposed in Chapter 3,

$$\begin{aligned} &logit(Pr(Y_i = 1)) = \beta_0 + \sum_{j=1}^{m_i} \beta_j^{Len}(X_{ij}^{(2)} - X_{ij}^{(1)}) + \\ &\sum_{j=1}^{m_i}(\beta_j^{Del} I[X_{ij}^{(3)} = 1] + \beta_j^{Amp} I[X_{ij}^{(3)} = 3]) + \sum_{j=1}^{m_i} \beta_j^{Dsg} |X_{ij}^{(4)} - 2| \\ &+ \sum_{j=1}^{m_i} \beta_j^{Len*Del*Dsg}(X_{ij}^{(2)} - X_{ij}^{(1)}) \times I[X_{ij}^{(3)} = 1] \times X_{ij}^{(4)} \\ &+ \sum_{j=1}^{m_i} \beta_j^{Len*Amp*Dsg}(X_{ij}^{(2)} - X_{ij}^{(1)}) \times I[X_{ij}^{(3)} = 3] \times X_{ij}^{(4)} \end{aligned} \tag{4.6}$$

where $X_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)}, X_{ij}^{(3)}, X_{ij}^{(4)})$ is the jth CNV of the ith individual as defined previously. β_0 corresponds to a baseline disease rate. β_j^{Len} controls the effect of chromosomal position, and β_j^{Del} and β_j^{Dup} are the log ratio of a CNV j for being deletion versus amplification and vice versa. β_j^{Del} and β_j^{Dup} share the same values but different signs. $\beta_j^{Len*Amp*Dsg}$ and $\beta_j^{Len*Del*Dsg}$ allow the effect of the chromosomal position and CNV type to differ by dosage in CNV j.

After generating phenotypes for SMCKAT and MCKAT, we use following logistic model that is proposed by Zhan et al. (2015a) to generate the phenotypes

under the CKAT method,

$$logit(\pi_i) = \beta_0 + \sum_{j=1}^{m_i} (\beta_j^{Del} I[X_{ij}^{(2)} = 1] + \beta_j^{Dup} I[X_{ij}^{(2)} = 3]) X_{ij}^{(1)} \tag{4.7}$$

where $X_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)})$ is the jth CNV of ith subject, $\pi_i = Pr(Y_i = 1)$, β_0 is the prevalence rate of the disease, and β_j^{Dup}, β_j^{Del} are the log of the odd ratio of CNV j for duplication and deletion respectively.

Similarly, we use following logistic model that is proposed by Brucker et al. (2020) to generate the phenotype under the CONCUR method,

$$\begin{aligned} logit(Pr(Y_i = 1)) = \gamma_0 + \beta_X X_i + \sum_{j=1}^{R} \beta_j^{\text{Dup}}\ Z_{ij}^{\text{Dup}}\ + \sum_{j=1}^{R} \beta_j^{\text{Del}}\ Z_{ij}^{\text{Del}} \\ + \sum_{j=1}^{R} \beta_j^{\text{Len}}\ Z_{ij}^{\text{Len}}\ + \sum_{j=1}^{R} \beta_j^{\text{Dup*Len}}\ Z_{ij}^{\text{Dup}}\ Z_{ij}^{\text{Len}}\ + \sum_{j=1}^{R} \beta_j^{\text{Del*Len}}\ Z_{ij}^{\text{Del}}\ Z_{ij}^{\text{Len}} \end{aligned} \tag{4.8}$$

where γ_0 and β_X are roughly set to -2 and $\log(1.1)$ respectively based on the baseline disease rate. $i = 1, ..., N$ indexes individuals, and $j = 1, ..., R$ indexes the CNV regions. β_j^{Dup} and β_j^{Del} are the log odds ratio for the presence of a CNV versus its absence in the segment j. Likewise, β_j^{Len} controls the effect of the CNV length and lets this effect differs by the CNV dosage value. Z^{Del} , Z^{Dup} and Z^{Len} are matrices which are generated based on CNV profiles. Z^{Dup} and Z^{Del} take value 1 for the CNV profiles that have a CNV in the CNV region j and 0 otherwise. Similarly, Z^{Len} codifies the length of the CNVs in the CNV regions and considers zero when a CNV profile is without CNVs in a specific region.

4.3.1 Simulation Results

The QQ-plots of comparing p-values of SMCKAT with MCKAT, CONCUR and CKAT under the first simulation scenario are presented in the following figures.

Based on the QQ-plots in Figures 4.6, 4.7 and 4.8, SMCKAT is on the 45 degree line under the first simulation scenario. This indicates that SMCKAT can properly handle the type I and II error rates under different nominal significance levels even as low as 10^{-4} when dealing with the rare CNV dataset. Similarly, MCKAT is capable of handling the type I and II error by being on the 45 degree line as is shown in Figure 4.6. However, SMCKAT is more conservative when

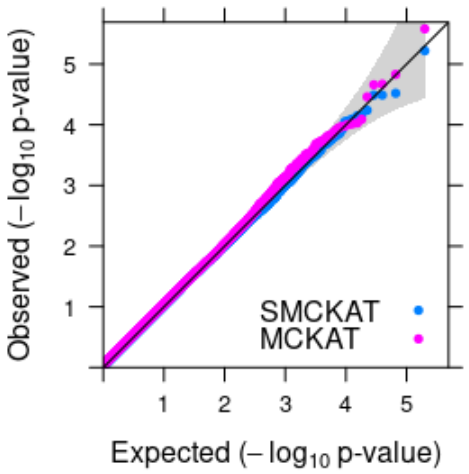

Figure 4.6: P-value based QQ-plots of SMCKAT and MCKAT under the first simulation scenario, the rare CNVs application.

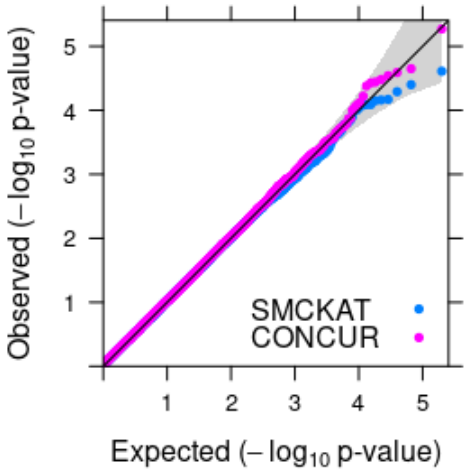

Figure 4.7: P-value based QQ-plots of SMCKAT and CONCUR under the first simulation scenario, the rare CNVs application.

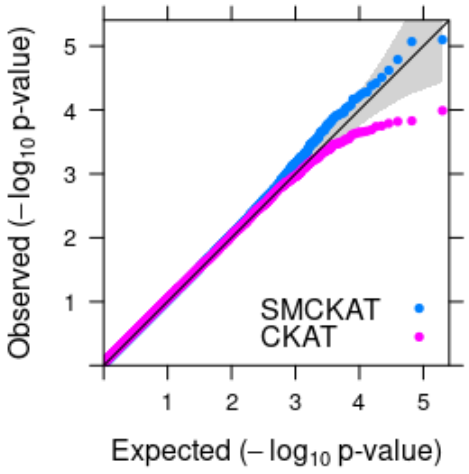

Figure 4.8: P-value based QQ-plots of SMCKAT and CKAT under first simulation scenario, the rare CNVs application.

the significance level is lower than 10^{-4} compared to MCKAT. This is due to more strict rules that SMCKAT takes into account to measure the similarity between CNV profiles. It uses not only the CNV characteristics but also the CNV sequential order in CNV profiles to test the association between CNVs and disease-related traits. We have the same observation when comparing SMCKAT with CONCUR. As shown in Figure 4.7, both SMCKAT and CONCUR can properly handle the type I and II error rates under different nominal significance levels even as low as 10^{-4} when dealing with the rare CNV dataset. SMCKAT is again more conservative when the significance level is lower than 10^{-4} while CONCUR is showing higher chance of committing type I error by being above the 45 degree line. However, as is shown in Figure 4.8, CKAT indicates a higher chance of committing type II errors by being below the 45 degree line for nominal significance levels as low as 10^{-3}.

The QQ-plots comparing p-values of SMCKAT with MCKAT, CONCUR and CKAT under the second simulation scenario are presented in the following. Based on the QQ-plots in Figures 4.9, 4.10 and 4.11, SMCKAT is on the 45 degree line under the second simulation scenario. This indicates that SMCKAT can properly handle the type I and II error rates under different nominal significance levels even as low as 10^{-4} when dealing with the frequent CNV dataset. Similarly, MCKAT is

capable of handling the type I and II error by being on the 45 degree line as shown in Figure 4.9. However, SMCKAT is more conservative when the significance level is lower than 10^{-4} compared to with MCKAT. It is the same behaviour observed in the first scenario from SMCKAT.

We have the same observation when comparing SMCKAT with CONCUR as is shown in Figure 4.10. Both SMCKAT and CONCUR can properly handle the type I and II error rates under different nominal significance levels even as low as 10^{-4} when dealing with the frequent CNV dataset. SMCKAT is again more conservative when the significance level is lower than 10^{-4} while CONCUR shows a higher chance of committing type I error by being above the 45 degree line. As is shown in Figure 4.11, unlike other three methods, CKAT has weak performance in handling type II errors when dealing with frequent CNVs. This means that CKAT is not capable of detecting any association between frequent CNVs and disease-related traits in this scenario.

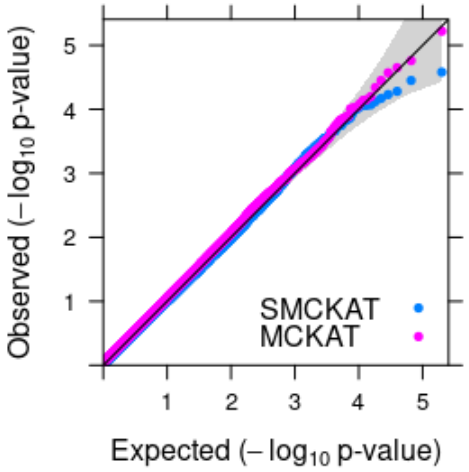

Figure 4.9: P-value based QQ-plots of SMCKAT and MCKAT under the second simulation scenario, the frequent CNVs application.

The empirical powers of SMCKAT, MCKAT, CONCUR and CKAT under the first and second scenarios are presented in Figures 4.12 and 4.13 respectively. The statistical association test power and how it is calculated are explained in details in Section 3.3. As shown in Figure 4.12, SMCKAT and MCKAT have almost similar powers when dealing with rare CNVs. However, CONCUR and

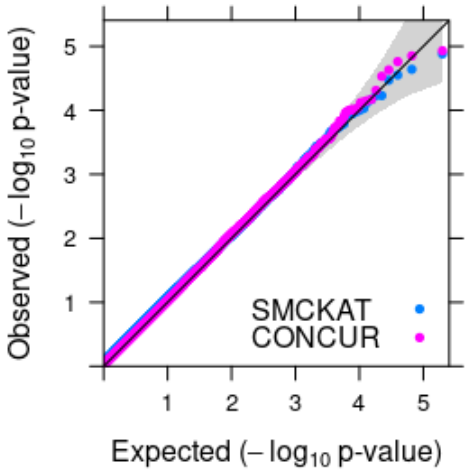

Figure 4.10: P-value based QQ-plots of SMCKAT and CONCUR under the second simulation scenario, the frequent CNVs application.

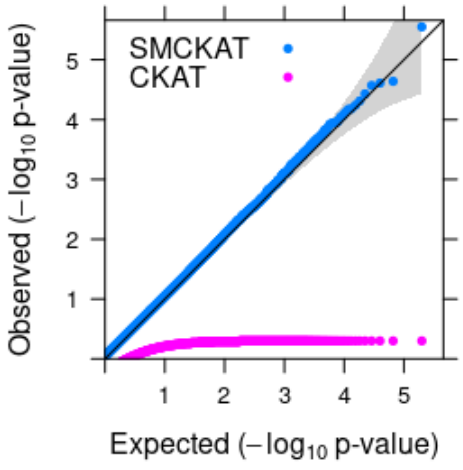

Figure 4.11: P-value based QQ-plots of SMCKAT and CKAT under the second simulation scenario, the frequent CNVs application.

CKAT show lower power compared with SMCKAT and MCKAT. The reason is that the CONCUR and CKAT are not considering all CNV characteristics when testing the association. CKAT considers neither CNV dosage nor heterogeneity effect when calculating similarity between CNV profiles. Furthermore, based on the CKAT design all possible CNVs in the CNVs profiles are not included in measuring the similarity between CNV profiles. Similarly, CONCUR ignores CNV heterogeneity effect by calculating similarities for deletion and amplification CNV types separately. These approaches may result in having lower power for both CKAT and CONCUR.

Similarly, in the second simulation scenario, SMCKAT and MCKAT have similar powers. CONCUR indicates lower performance compared with SMCKAT and MCKAT due to the same reason as in the first scenario. However, CKAT is showing too low power when dealing with frequent CNV data. As explained in the first scenario, CKAT shift-by-one scanning CNV algorithm, results in not measuring the similarity between the CNV profiles in an optimal way. This design type has bigger effect when we are dealing with more amount of CNVs that exist in the frequent CNV dataset.

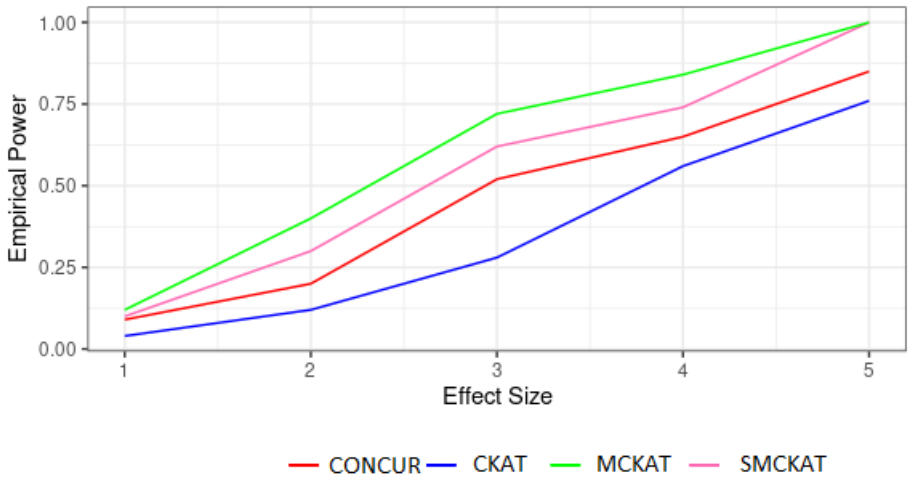

Figure 4.12: Empirical power of SMCKAT, MCKAT, CONCUR and CKAT under the first simulation scenario, rare CNV data.

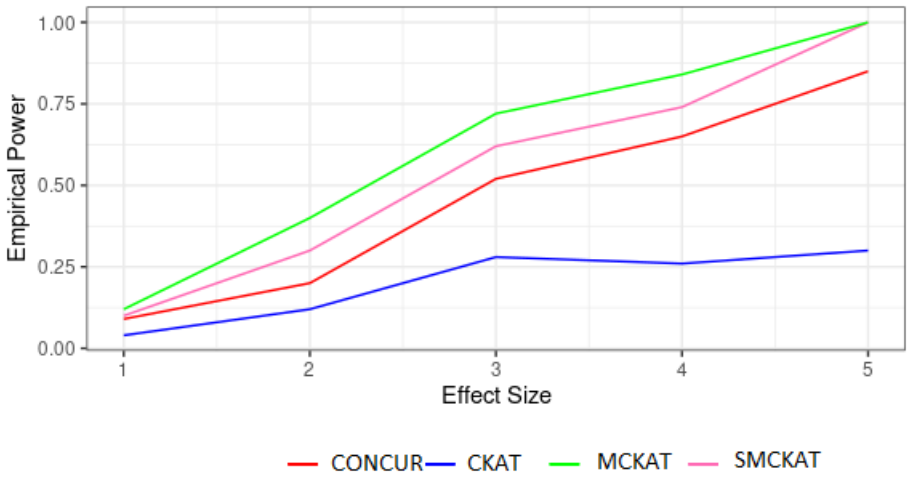

Figure 4.13: Empirical power of SMCKAT, MCKAT, CONCUR and CKAT under the second simulation scenario, frequent CNV data.

4.4 Real Data Application Results

We conduct SMCKAT analysis, for different CNV group sizes, on single chromosomes and the whole genome to test the association between CNV sequential order and disease-related traits. The disease-related traits studied in this thesis are cancer subtype for the RMS data set and disease status for the ASD data set. We compare SMCKAT results with those obtained from MCKAT and CKAT to evaluate SMCKAT performance on real CNV data.

4.4.1 CNV Analysis on Rhabdomyosarcoma Data Set

First, we conduct the experiment on the RMS data. The RMS occurs as two major histological subtypes, embryonal (ERMS) and alveolar (ARMS). The classification of the RMS subtype has a direct effect on the patient treatment options. The RMS data, which is explained in details in Section 3.4.2, includes a total of 59,131 CNVs for 25 alveolar and 19 embryonal cancers. We apply SMCKAT to each of 23 chromosome pairs, with different CNV group sizes, to test the association between CNV sequential order and RMS sub-type. Bonferroni correction is used for adjusting the multiple testing to control the family-wise error rate (FWER) of $\alpha = 0.05$. As in Chapter 3, since 22 chromosomes and sex chromosome are being tested, the p-value threshold for a whole-chromosome significance is calculated as

$0.05/23 = 2.2 \times 10^{-3}$. SMCKAT identifies four chromosomes out of existing 23 chromosomes where the CNV sequential order in these chromosomes is significantly associated with the RMS sub-type. The p-values of SMCKAT for these four chromosomes are reported in Table 4.1.

Table 4.1: P-values of the chromosomes that their CNV sequential orders are identified significantly associated with the RMS sub types for the different CNV group sizes.

Chr.	#CNV	n=1	n=2	n=3	n=4	n=5	n=6
2	5584	2.45×10^{-2}	5.10×10^{-2}	8.31×10^{-2}	3.49×10^{-3}	4.25×10^{-3}	3.21×10^{-2}
8	5365	2.61×10^{-5}	7.37×10^{-6}	1.13×10^{-6}	7.63×10^{-7}	4.99×10^{-8}	0
11	3449	2.03×10^{-2}	8.26×10^{-3}	2.93×10^{-3}	1.54×10^{-3}	5.82×10^{-4}	1.20×10^{-4}
13	2462	1.80×10^{-3}	3.56×10^{-3}	4.86×10^{-3}	6.06×10^{-3}	7.89×10^{-3}	6.23×10^{-2}

Based on the results, SMCKAT identifies CNV sequential order in chromosomes 2, 8, 11, and 13 significantly associated with distinguishing RMS subtype at $FWER = 2.2 \times 10^{-3}$.

Based on the literature, El Demellawy et al. (2017) shows that RMS is associated with specific chromosomal abnormalities that differentiate ARMS and ERMS. Based on their study, approximately 80% of ARMS tumors display a translocation between the *FOXO1* transcription factor gene located on chromosome 13 and the *PAX3* transcription factor gene on chromosome 2, and ERMS tumors show a higher frequency of specific genetic mutation on chromosome 11 than ARMS. Sun et al. (2015) has revealed the same earlier. Furthermore, Nishimura et al. (2013) has found the ARMS subtype is significantly associated with amplifications on chromosome 8.

Our findings show another mechanism like CNV sequential orders along with the CNV characteristics can play a significant role in causing any disease-related traits besides gene mutations, chromosomal translocations and any other genetic variations. However, since based on the literature, SMCKAT is the first study, both biological and computational, that is investigating the significance of CNV sequential order and it's association with disease-related traits, we are not capable of validating our results using the existing biological knowledge more precisely.

We test different CNV group sizes when applying SMCKAT to RMS data set. Based on the results reported in Table 4.4, SMCKAT shows the strongest evidence and smallest p-value, for chromosome 8 for all CNV group sizes. It means

subjects with the same RMS subtype may have the similar CNV sequential order on their chromosome 8.

We test SMCKAT on the RMS data set for group sizes greater than five. As is shown in for the gruop size 6 as an example, We observe an increasing trend in p-values, except for chromosome 8, by increasing the group size, which means there is a lower chance of having association between CNV sequential orders and disease-related traits for the larger group sizes.

4.4.2 CNV Analysis on Cytogenetic Bands in RMS

Based on the result reported in Table 4.4, there is strong statistical evidence, as supported by a p-value near to zero, that the CNV sequential order of chromosome 8 with group size of 6 is significantly associated with RMS sub-type. So, we choose chromosome 8 with CNV group size of 6 for further analysis. We partition chromosome 8 into smaller regions based on the cytogenetic bands. We apply SMCKAT on each cytogenetic band to check if SMCKAT is capable of detecting more specific regions rather than chromosomes. Then, we compare the results with MCKAT and CKAT. Table 4.2 contains the p-values of the association test in each cytogenetic bands in chromosome 8. Since 40 cytogenetic bands are being tested in chromosome 8, the p-value threshold for a band significance is calculated as $0.05/40 = 1.2 \times 10^{-3}$.

As is shown in Table 4.2, SMCKAT, MCKAT and CONCUR detect significantly associated cytogenetic bands with the RMS subtype while CKAT does not identify any significant regions. As we discussed in simulation study, this might be due to the CKAT scanning algorithm for aligning CNVs in the CNV profiles. The CKAT shift-by-one scanning algorithm can capture similarity only between a limited number of CNVs, which may result in low performance when dealing with frequent CNVs. MCKAT has identified 8 cytogenetic bands with CNVs that are significantly associated with RMS sub type. Four out of these eight cytogenetic bands, $8p23.1$, $8p12.0$, $8q11.21$ and $8q24.3$, are identified by CONCUR as well. Similarly, SMCKAT identifies two of these eight cyotogenetic bands, $8p23.1$ and $8p12.0$, significantly associated with the RMS sub type. Based on the results, SMCKAT is more conservative in identifying cytogenetic bands with CNVs that are significantly associated with disease-related traits. This might be due to considering not only CNV characteristics but the CNV order in testing

the association. Therefore, SMCKAT has the potential to provide us with more specific CNV regions when we are testing the association between CNVs and disease-related traits comparing with MCKAT, CONCUR and CKAT.

Table 4.2: P-values of the testing association between RMS subtype and CNVs in the chromosome 8 cytogenetic bands by SMCKAT, MCKAT and CKAT. (*) denotes significant association between RMS subtype and CNVs, (#) denotes the number of total CNVs on the band.

Arm	Band	Start	Stop	#CNVs	SMCKAT	MCKAT	CKAT	CONCUR
p	23.3	1	2,300,000	113	9.6×10^{-2}	3.4×10^{-4} *	4.9×10^{-1}	2.5×10^{-2}
p	23.2	2,300,001	6,300,000	85	3.0×10^{-2}	2.0×10^{-2}	3.9×10^{-1}	3.5×10^{-2}
p	23.1	6,300,001	12,800,000	304	1.8×10^{-4} *	4.7×10^{-8} *	4.7×10^{-1}	1.6×10^{-4} *
p	22.0	12,800,001	19,200,000	101	2.8×10^{-2}	8.2×10^{-3}	4.3×10^{-1}	3.4×10^{-2}
p	21.3	19,200,001	23,500,000	102	1.1×10^{-1}	2.5×10^{-2}	4.2×10^{-1}	4.9×10^{-2}
p	21.2	23,500,001	27,500,000	82	3.4×10^{-2}	3.6×10^{-2}	4.7×10^{-1}	2.6×10^{-2}
p	21.1	27,500,001	29,000,000	50	2.5×10^{-2}	1.6×10^{-2}	4.9×10^{-1}	1.3×10^{-2}
p	12.0	29,000,001	36,700,000	190	1.3×10^{-6} *	3.7×10^{-5} *	4.6×10^{-1}	9.5×10^{-4}*
p	11.23	36,700,001	38,500,000	48	1.0	3.7×10^{-3}	3.9×10^{-1}	1.5×10^{-1}
p	11.22	38,500,001	39,900,000	57	9.3×10^{-2}	8.4×10^{-3}	4.6×10^{-1}	1.6×10^{-3}
p	11.21	39,900,001	43,200,000	147	4.4×10^{-3}	1.0×10^{-4} *	3.6×10^{-1}	8.3×10^{-3}
p	11.1	43,200,001	45,200,000	72	8.8×10^{-2}	2.8×10^{-2}	4.5×10^{-1}	9.1×10^{-2}
q	11.1	45,200,001	47,200,000	41	1.0	2.1×10^{-2}	4.4×10^{-1}	3.4×10^{-1}
q	11.21	47,200,001	51,300,000	200	4.4×10^{-3}	8.4×10^{-5} *	4.0×10^{-1}	2.5×10^{-4}*
q	11.22	51,300,001	51,700,000	6	9.3×10^{-1}	4.7×10^{-2}	4.2×10^{-1}	8.3×10^{-2}
q	11.23	51,700,001	54,600,000	61	1.0	6.1×10^{-2}	4.6×10^{-1}	9.7×10^{-2}
q	12.1	54,600,001	60,600,000	177	9.1×10^{-3}	7.0×10^{-4} *	4.5×10^{-1}	5.4×10^{-3}
q	12.2	60,600,001	61,300,000	18	1.0	3.3×10^{-2}	4.5×10^{-1}	2.3×10^{-1}
q	12.3	61,300,001	65,100,000	134	4.9×10^{-2}	1.1×10^{-2}	4.1×10^{-1}	3.1×10^{-2}
q	13.1	65,100,001	67,100,000	71	4.4×10^{-2}	5.8×10^{-3}	4.4×10^{-1}	2.5×10^{-2}
q	13.2	67,100,001	69,600,000	54	5.8×10^{-2}	4.3×10^{-3}	4.6×10^{-1}	3.4×10^{-2}
q	13.3	69,600,001	72,000,000	62	1.4×10^{-2}	1.8×10^{-3}	3.7×10^{-1}	1.0×10^{-2}
q	21.11	72,000,001	74,600,000	144	4.8×10^{-1}	8.4×10^{-3}	3.3×10^{-1}	3.9×10^{-2}
q	21.12	74,600,001	74,700,000	1	1.0	1.0	1.0	1.0
q	21.13	74,700,001	83,500,000	308	1.0×10^{-2}	2.6×10^{-3}	4.9×10^{-1}	4.4×10^{-3}
q	21.2	83,500,001	85,900,000	56	4.8×10^{-2}	2.9×10^{-2}	4.1×10^{-1}	3.5×10^{-3}
q	21.3	85,900,001	92,300,000	185	4.7×10^{-3}	1.0×10^{-4} *	4.2×10^{-1}	2.8×10^{-3}
q	22.1	92,300,001	97,900,000	182	1.7×10^{-2}	1.0×10^{-2}	3.0×10^{-1}	1.3×10^{-2}
q	22.2	97,900,001	100,500,000	103	4.5×10^{-2}	3.9×10^{-3}	4.3×10^{-1}	2.6×10^{-2}
q	22.3	100,500,001	105,100,000	162	1.2×10^{-2}	4.6×10^{-3}	4.4×10^{-1}	0.9×10^{-3}
q	23.1	105,100,001	109,500,000	135	2.8×10^{-3}	2.5×10^{-3}	4.0×10^{-1}	1.4×10^{-3}
q	23.2	109,500,001	111,100,000	33	9.8×10^{-1}	8.0×10^{-1}	3.0×10^{-1}	4.3×10^{-1}
q	23.3	111,100,001	116,700,000	185	1.1×10^{-2}	2.3×10^{-3}	4.4×10^{-1}	0.4×10^{-2}
q	24.11	116,700,001	118,300,000	53	4.6×10^{-2}	2.6×10^{-2}	4.7×10^{-1}	2.3×10^{-2}
q	24.12	118,300,001	121,500,000	109	2.5×10^{-3}	2.2×10^{-3}	4.0×10^{-1}	2.0×10^{-3}
q	24.13	121,500,001	126,300,000	151	2.2×10^{-2}	6.0×10^{-3}	4.8×10^{-1}	5.4×10^{-3}
q	24.21	126,300,001	130,400,000	208	5.0×10^{-2}	1.9×10^{-2}	3.9×10^{-1}	3.2×10^{-2}
q	24.22	130,400,001	135,400,000	155	5.5×10^{-2}	1.5×10^{-2}	4.6×10^{-1}	4.6×10^{-2}
q	24.23	135,400,001	138,900,000	162	2.8×10^{-1}	7.7×10^{-3}	4.5×10^{-1}	8.3×10^{-2}
q	24.3	138,900,001	145,138,636	354	8.8×10^{-3}	2.5×10^{-8}*	4.2×10^{-1}	5.5×10^{-4}*

4.4.3 CNV Analysis on Autism Data Set

We apply SMCKAT on the ASD data set to evaluate its performance on the rare CNV type. We aim to test if there is any association between the sequential order of CNVs and ASD status. The ASD data set which is explained in details in section 3.4.2 contains 1285 rare CNVs on 310 individuals with ASD and 1074 rare CNVs on 278 healthy individuals. Since the ASD data set contains only rare and large CNVs, an arbitrary CNV profile may have no or few CNVs on some chromosomes. Therefore, instead of applying SMCKAT to all 23 chromosomes, we apply it to the whole genome. Then, we test if there is any association between the whole genome CNV sequential order and the ASD status. We consider 0.05 as the p-value threshold for the whole-chromosome significance. As in Table 4.3, there is strong statistical evidence, up to CNV group size of four, that subjects with the same disease status have similar CNV order in their CNV profiles. We test SMCKAT on the ASD data for the larger group sizes, five and six as well. We observe an increasing trend in p-values by increasing the group size as is shown in Table 4.3. which shows declining in the significance level of the CNV sequential order associated with the ASD status.

Table 4.3: P-values of testing the association between CNV sequential order and ASD status trying different CNV group sizes.

n	1	2	3	4	5	6
p-value	0	7.91×10^{-9}	3.09×10^{-6}	3.62×10^{-4}	4.89×10^{-3}	1.03×10^{-1}

4.5 Discussion

SMCKAT tests the association between the CNV sequential orders and disease-related traits. It checks if CNV sequential orders, have a significant association with disease-related traits. Our approach has several advantages over the existing methods. Firstly, it measures the similarity between CNV profiles by considering not only all CNV characteristics but also the CNV sequential order. It is the first approach to study the association between CNV sequential order and disease-related traits to our knowledge. Secondly, it is applicable to both rare and frequent

CNV data sets while previous methods like CKAT can only deal with common CNV data sets. Thirdly, SMCKAT is more stringent compared with the state-of-the-art approaches in detecting significant CNV regions due to more strict rules used in its design for measuring the similarity between the CNV profiles. Finally, SMCKAT has the potential to help biologists detect significantly associated CNV regions, more specific regions compared with the state-of-the-art kernel approach, with any disease-related trait across a patient group instead of examining the CNVs case by case in each subject.

Although our experimental results are promising and we were able to investigate if there is any association between the CNV sequential order and disease-related traits for the first time by achieving strong evidences, small p-values, this study has some limitations. There are not many publicly available CNV datasets. In more details, most of the available ones do not contain all CNV features: chromosomal regions, type and dosage. We test our method on available datasets and our simulated CNV data that includes all multi-dimensional CNV characteristics. For ASD data set that is used in this thesis, we had to simulate CNV dosage to be able to evaluate the performance of our proposed method. Applying SMCKAT to more data sets containing all CNV features can help to determine its strengths and weakness.

Our study shows that CNV sequential order has the potential to play a significant role in causing disease-related traits, but more new findings can be revealed by conducting more comprehensive analysis upon the availability of data.

4.6 Conclusion

The results presented in this chapter demonstrate that our method SMCKAT, provides the answer to the question of whether CNV sequential orders are significantly associated with disease-related traits. SMCKAT improves outcomes for detecting significant associations between CNVs, both rare and frequent, and disease-related traits. While MCKAT tests the association between CNVs and disease-related traits by using only CNV characteristics, SMCKAT utilizes both CNV characteristics and sequential order in the association test. SMCKAT has the potential to provide biologists with more specific CNV hot spots on the genome by being more stringent compared with existing methods. This work forms Contribution 2 of this thesis which is proposing a sequential multi-dimensional CNV

kernel based association test to investigate whether CNV sequential order has a significant association with disease-related traits.

In the next chapter we will do further explorations to investigate the dual effect of CNVs and other genetic variations like genes in disease development.

Chapter 5

CNV-gene intersection effect on testing the association between CNVs and disease-related traits

5.1 Introduction

Previous works, discussed in Chapters 3 and 4, have clearly demonstrated that there is an association between CNV characteristics and disease-related traits. The results presented in these chapters have shown that both MCKAT and SMCKAT can be used to identify chromosomal regions which CNVs on them have significant associations with disease-related traits. These chromosomal regions, as small as cytogenetic bands, can provide biologists with more specific CNV hot spots to identify CNVs that have the potential to cause any disease-related traits without the need to conduct an extensive investigation of the whole genome in affected individuals.

Another important field that is worth investigating and has not received much attention is the dual effects of CNVs and genes on disease development. It is believed that the distribution of CNVs in the genome is not random, and several hot spots of CNVs have been found. For example, as is reported by Cooper et al. (2007), 250 regions of 1 Mb DNA sequence have been found in which $> 50\%$ bases are within copy number variants. In particular, CNVs are more common near telomeres and centromeres, possibly because these genomic regions are repetitive (Nguyen et al. 2006). It has also been reported that there is a significant relationship between CNV regions and genes, with CNV-rich genomic

regions enriched with genes and vice versa (Cooper et al. 2007).

Numerous published studies have shown that the genome contains many CNV loci that have intersections with many genes. In this chapter, the association between the dual effects of CNVs and their intersected genes with disease-related traits is investigated. We test whether the CNVs that have intersections with a gene or gene set which are previously identified as significant to a disease-related trait in the biological literature, are significantly associated with that trait or not. The work in this chapter addresses Contribution 3 listed in the Chapter 1, which is the demonstration that considering the effect of CNV-gene intersections in addition to the CNV characteristics is informative and helpful in identifying the significant association between CNVs and disease-related traits.

5.2 Effects of CNVs on Gene Expressions

CNVs are responsible for at least 17.7% of the heritable variation in gene expression as is reported by Stranger et al. (2007). Genes can be impacted by copy number variation in various ways, which can be seen in how they affect phenotypes. First, variation in gene expression can happen simply through the dosage effects where whole genes vary in copy number. Losses or gains in copy number would decrease or increase expression levels, respectively. However, variation in expression due to CNVs may differ from one gene to another gene.

Gene expression variation is not only due to consequences of copy number variation on gene dosage. According to Stranger et al. (2007), approximately half of the effects of CNVs on gene expression levels are brought on by disruptions in the gene coding sequences, such as the deletion of exons, or by affecting regulatory elements and other functional regions. CNVs that affect a portion of a gene may also result in the formation of variant proteins by the creation of splice variants or exon shuffling (Korbel et al. 2007; Masson et al. 2008).

Gene expression may also be affected by amplifications or deletions lying outside coding sequences. The affect can be done by changing the efficiency or location of important regulatory elements through position effects. Furthermore, Stranger et al. (2007) found several significant CNV associations with genes that were more than 2 Mb away from them, providing evidence that CNVs can impact long-range gene regulation.

The impact of CNVs on gene expression and their potential to disrupt gene

structure and function, lead to the belief that they are very likely to play a significant role in developing human diseases. Based on the literature, a few investigations have been carried out so far to examine the dual effects of CNVs and genes on the disease development. Therefore, we conduct simulation studies to investigate the dual effects of CNVs and their intersected genes on the disease development using our proposed methods MCKAT and SMCKAT. We do not consider neither CONCUR nor CKAT in our simulation studies in this chapter, as they do not take the CNV-gene intersection information as input data and are not capable of using it in the association test.

5.3 Simulation studies

We conduct our simulation studies under four main scenarios. The simulation scenarios are based on the existence of CNV-gene intersections and the type of the CNV that intersects with a gene in a chromosomal region. In the first scenario we investigate the association between the CNVs and disease-related traits where there is no CNV-gene intersection. Whereas, in the other three scenarios we investigate the dual effects of CNVs and intersected genes on disease-related traits. In the first scenario, we have copy number variation regions with no intersection between CNVs and genes. Therefore, no genes are affected by any CNVs and the association between CNVs and disease-related traits are investigated on their own. In the second scenario, we have copy number variation regions with deletions only to investigate the association of CNVs with amplification type and their intersected gene with disease-related traits. In the third scenario we have copy number variation regions with amplifications only to investigate the association of CNVs with deletion type and their intersected gene with disease-related traits. Finally, in the fourth scenario, we have copy number variation regions in which both amplification and deletion CNVs intersect with genes. Under this scenario we investigate the effect of CNV types on their intersected genes and their association with disease-related traits. These four scenarios are depicted in Figures 5.1 to 5.4.

We use RMS CNV data set which is described in Section 3.4.2 as the base for simulating CNV data as it has frequent CNVs which increases the probability of the CNV-gene intersections. Briefly, the RMS CNV dataset includes 59,131 CNVs of 44 individuals, both frequent and rare CNVs, with different RMS cancer

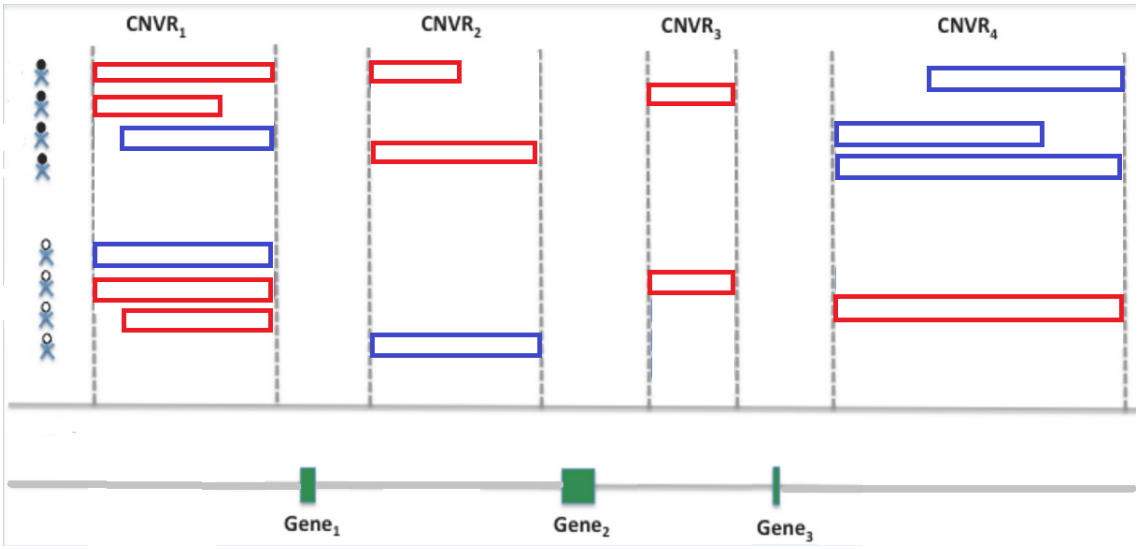

Figure 5.1: Scenario 1, no intersections between CNVs and genes. Each row is a CNV profile of a subject. CNVR: copy number variation region, blue rectangle: amplification and red rectangle: deletion.

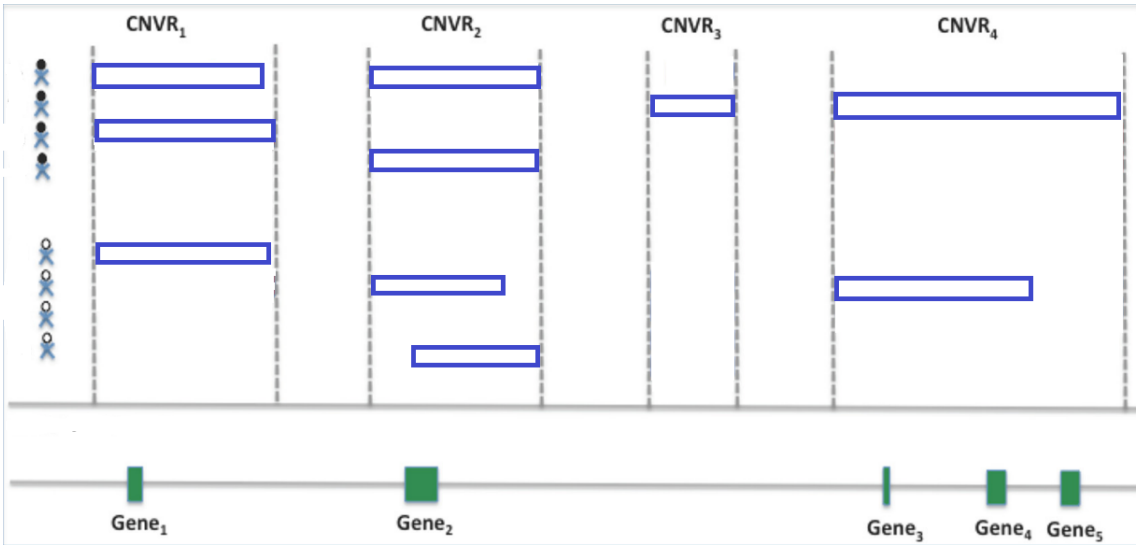

Figure 5.2: Scenario 2, genes have intersection only with CNVs of amplification type. Each row is a CNV profile of a subject. CNVR: copy number variation region, red rectangle: deletion.

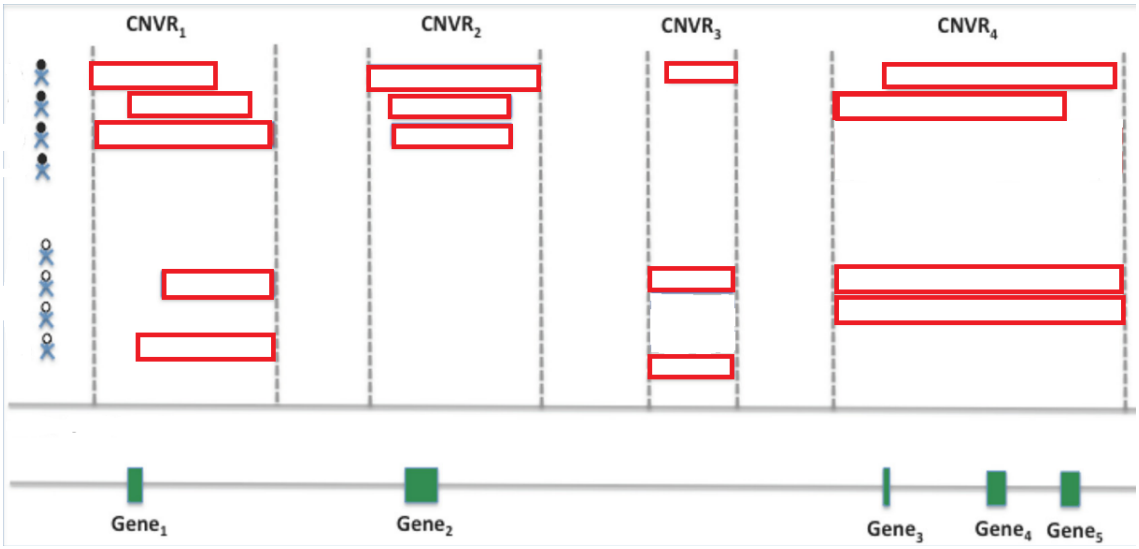

Figure 5.3: Scenario 3, genes have intersection only with CNVs of deletion type. Each row is a CNV profile of a subject. CNVR: copy number variation region, blue rectangle: amplification.

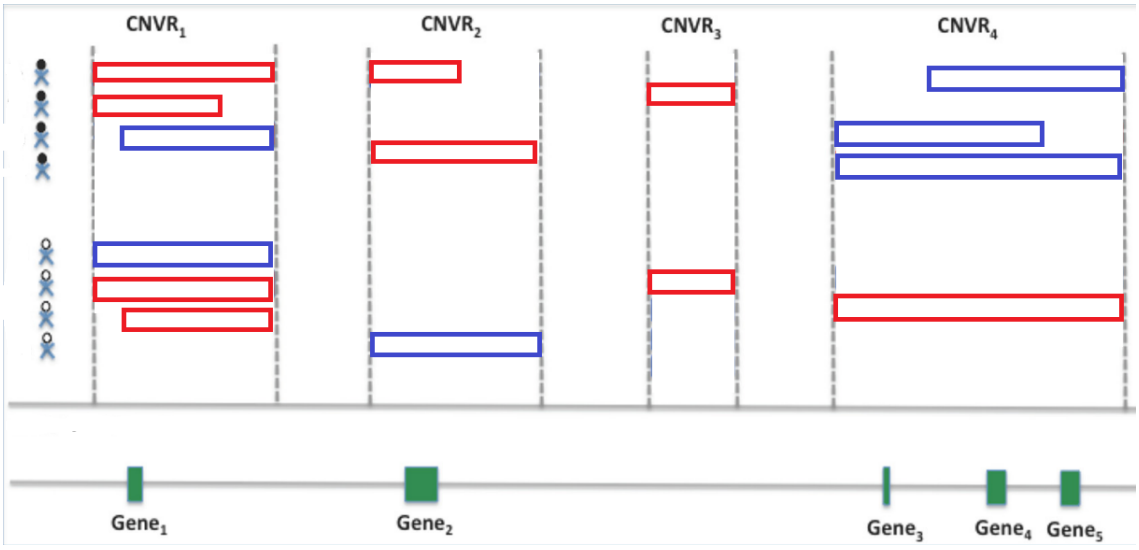

Figure 5.4: Scenario 4, genes have intersection with CNVs of both amplification and deletion types. Each row is a CNV profile of a subject. CNVR: copy number variation region, blue rectangle: amplification and red rectangle: deletion.

subtypes. The chromosomal position, dosage and type information are available for each CNV in RMS data. To simulate some chromosomal positions for genes across the genome we use the chromosomal positions of the genes that are reported by Shern et al. (2014) with significant frequency of somatic mutation in RMS patients. Table 5.3 contains these genes and their chromosomal position on the genome. Then, based on the intersection between chromosomal position of these genes and CNVs we split the RMS CNV data into three groups to perform four simulation scenarios explained previously. CNVs intersected with all 8 genes, where 3 genes are intersected by amplifications only, 2 genes are intersected by deletions only and 2 genes are intersected by both deletions and amplifications.

Gene	Chromosome	Cytogenetic band	Start position	End position
NRAS	1	p13.2	114,704,469	114,716,771
FGFR4	5	q35.2	177,086,915	177,098,144
PIK3CA	3	q26.32	179,148,357	179,240,093
FBXW7	4	q31.3	152,320,544	152,536,092
HRAS	11	p15.5	532,242	535,576
KRAS	12	p12.1	25,205,246	25,250,929
TP53	17	p13.1	7,668,421	7,687,490
NF1	17	q11.2	31,094,977	31,377,675

Table 5.1: Genes with significant frequency of somatic mutation across RMS patients

After dividing the CNV data based on the criteria we consider for each scenario, we simulated 10^5 datasets for each scenario. Then, we propose the following logistic model which includes CNV-gene intersection effects to generate the case-control

label Y_i:

$$
\begin{aligned}
logit(Pr(Y_i = 1)) &= \beta_0 + \sum_{j=1}^{m_i} \beta_j^{Len}(X_{ij}^{(2)} - X_{ij}^{(1)}) \\
&+ \sum_{j=1}^{m_i} (\beta_j^{Del} I[X_{ij}^{(3)} = 1] + \beta_j^{Amp} I[X_{ij}^{(3)} = 3]) \\
&+ \sum_{j=1}^{m_i} \beta_j^{Dsg} |X_{ij}^{(4)} - 2| \\
&+ \sum_{j=1}^{m_i} \beta_j^{Len*Del*Dsg} (X_{ij}^{(2)} - X_{ij}^{(1)}) \times I[X_{ij}^{(3)} = 1] \times X_{ij}^{(4)} \\
&+ \sum_{j=1}^{m_i} \beta_j^{Len*Amp*Dsg} (X_{ij}^{(2)} - X_{ij}^{(1)}) \times I[X_{ij}^{(3)} = 3] \times X_{ij}^{(4)} \\
&+ \sum_{j=1}^{m_i} \sum_{G=1}^{g} (\beta_g^{DelGI} I[X_{ij}^{(3)} = 1] \times Z_{jg}^{DelGI} + \beta_g^{AmpGI} I[X_{ij}^{(3)} = 3] \times Z_{jg}^{AmpGI})
\end{aligned}
\tag{5.1}
$$

where $i = 1, ..., N$ indexes individuals, and $j = 1, ..., m_i$ indexes the CNVs of individual i. $X_{ij} = (X_{ij}^{(1)}, X_{ij}^{(2)}, X_{ij}^{(3)}, X_{ij}^{(4)})$ is the jth CNV of the ith individual as defined previously. β_0 corresponds to a baseline disease rate. β_j^{Len} controls the effect of chromosomal position, and β_j^{Del} and β_j^{Dup} are the log ratio of a CNV j for being deletion versus amplification and vice versa defined based on thr CNV data including in each scenario. Likewise, β_j^{Dsg} controls the effect of dosage in CNV j. $\beta_j^{Len*Amp*Dsg}$ and $\beta_j^{Len*Del*Dsg}$ allow the effect of the chromosomal position and CNV type to differ by dosage in CNV j. $G = 1, ..., g$ indexes the genes included in our simulation study. $Z_{jg}^{DelGI} = 1$ if CNV j is intersected gene g by deletion and 0 when there is no intersection. Similarly, $Z_{jg}^{AmpGI} = 1$ if CNV j is intersected gene g by amplification and 0 when there is no intersection. β_g^{DelGI} and β_g^{AmpGI} are the log odd ratios of the gene g for a deletion intersection and a amplification intersection, respectively. β_g^{DelGI} and β_g^{AmpGI} share the same absolute values but are positive if we assume CNV-gene intersections have risk associated effects and negative if protective effects.

Finally, following the approach in Chapters 3 and 4, we apply MCKAT and SMCKAT on the CNV data that we prepare for each scenario. We check if they are capable of handling type I and II errors when we are considering CNV-gene intersection effect besides CNV characteristics in testing association between CNVs and disease-related traits. The QQ-plots of p-values of MCKAT and SMCKAT

under each simulation scenario are presented in Figures 5.5 to 5.8. The usage of QQ-plot and its meaning is explained in details in section 3.3.

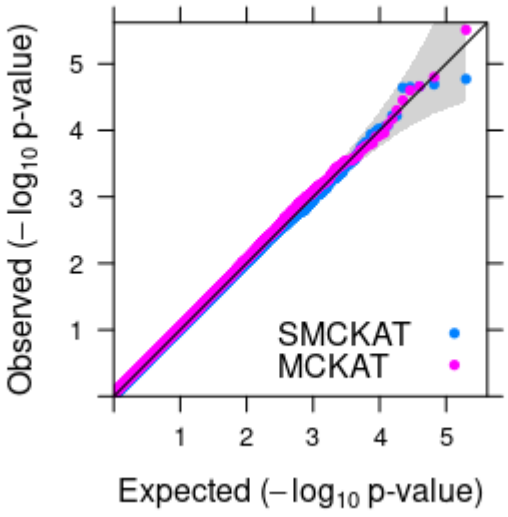

Figure 5.5: P-value based QQ-plots of MCKAT and SMCKAT under the first simulation scenario, no CNV-gene intersections.

As is shown in Figure 5.5, under the first scenario when there is no CNV-gene intersections, both MCKAT and SMCKAT p-values are on the 45-degree line under different nominal significance levels even as low as 10^{-5}. This means the observed p-values calculated by MCKAT and SMCKAT are the same as the expected p-values. This indicates that both methods are capable of handling type I and II errors in testing the association between CNVs and disease-related traits when there is no CNV-gene intersections. We had observed similar results for both MCKAT and SMCKAT under simulation studies and real data application studies in the Chapters 3 and 4.

In Figure 5.6, under the second scenario when there is only CNV-gene amplification intersections, both MCKAT and SMCKAT calculated p-values are on the 45-degree line under different nominal significance levels up to 10^{-4} are on the 45-degree line. This indicates that both methods are capable of handling the type I and II error in testing the association. We observe for significance levels lower than 10^{-4} both methods' p-values are slightly below or above the 45-degree

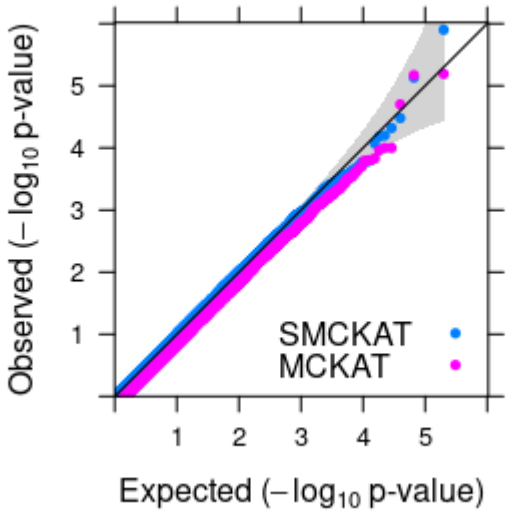

Figure 5.6: P-value based QQ-plots of MCKAT and SMCKAT under the second simulation scenario, only CNV-gene amplification intersections.

line. This means in some cases both MCKAT and SMCKAT may not identify any associations between CNV profiles and disease-related traits and vice versa when we need very low p-value significance level lower than 10^{-5}. However, a p-value less than 0.05 is typically considered to be statistically significant, in which case the null hypothesis should be rejected.

In Figure 5.7, under the third scenario when there is only intersections between CNV deletions and genes, we observe same results as of the second scenario. Both MCKAT and SMCKAT calculated p-values are on the 45-degree line under different nominal significance levels up to 10^{-4}. This indicates that both methods are capable of handling type I and II errors in testing the association in CNV-gene amplification intersection scenario as well. We observe for the significance level lower than 10^{-4} both methods' p-values are slightly below or above the 45-degree line. This means that in some cases both MCKAT and SMCKAT may not identify any associations between CNV profiles and disease-related traits and vice versa when we need very low p-value significance level lower than 10^{-5}. However, as mentioned previously, a p-value less than 0.05 is typically considered to be statistically significant, in which case the null hypothesis should be rejected.

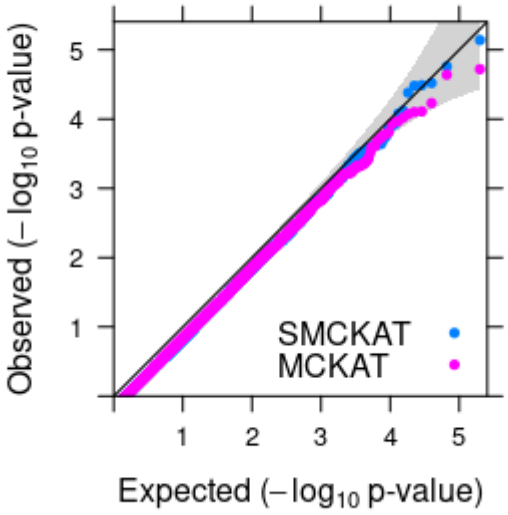

Figure 5.7: P-value based QQ-plots of MCKAT and SMCKAT under the third simulation scenario, only CNV-gene deletion intersections.

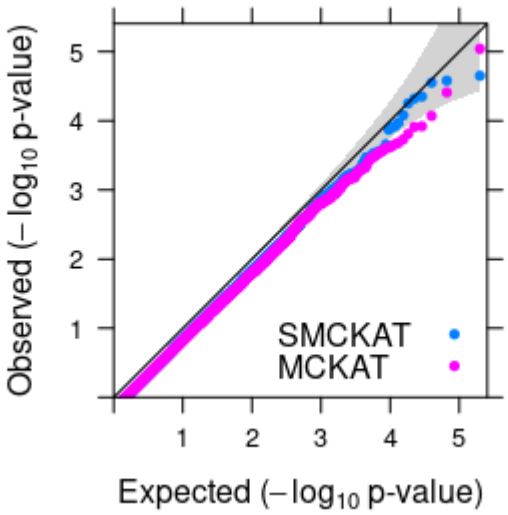

Figure 5.8: P-value based QQ-plots of MCKAT and SMCKAT under the fourth simulation scenario, CNV-gene both deletion and amplification intersections.

As shown in Figure 5.8, under the fourth scenario when there are CNV-gene intersections, no matter deletion or amplification, both MCKAT and SMCKAT calculated p-values are on the 45-degree line for the nominal significance as low as 10^{-3}. This means both MCKAT and SMCKAT are capable of handing type I and II errors up to this significance level. The slight drop in both MCKAT and SMCKAT in handling type I and type II errors may be result of CNV-gene intersection heterogeneity effect which is considered in this simulation scenario. The CNV-gene intersections heterogeneity effect means a CNV-gene amplification intersection can have a different effect on disease-related traits compared with a CNV-gene deletion intersection. One can have risk associated effects to a disease-related traits while other can have protective effects.

The empirical powers of MCKAT and SMCKAT, in the other words, the probability that they correctly reject the null hypothesis when the alternative hypothesises true, under two scenarios are presented in Figures 5.9 and 5.10 respectively. The null hypothesis is that there is no association between CNV characteristics and CNV-gene intersections with disease-related traits. In the first scenario there is no CNV-gene intersections in CNV profiles while in the second scenario there are CNV-gene intersections in CNV profiles. In both scenarios we observe that both MCKAT and SMCKAT ensure sufficient power, above 0.80, in detecting significant association between CNV profiles and the disease-related trait. We also observe that both MCKAT and SMCKAT show higher power when there are no gene intersections compared with the scenarios where CNV-gene intersections exist which could be due their designs. The built-in kernels in both MCKAT and SMCKAT are designed to measure the similarity between CNV profiles based on CNV chromosomal region, type and dosage, not CNV-gene intersection characteristics. Consequently, both methods struggle to pick up the CNV-gene intersection effect signals due to their kernels' design in testing the association. This situation can be improved by revising both MCKAT and SMCKAT kernels and adding another kernel to consider CNV-gene intersection effect in testing the association between CNV profiles and disease-related traits.

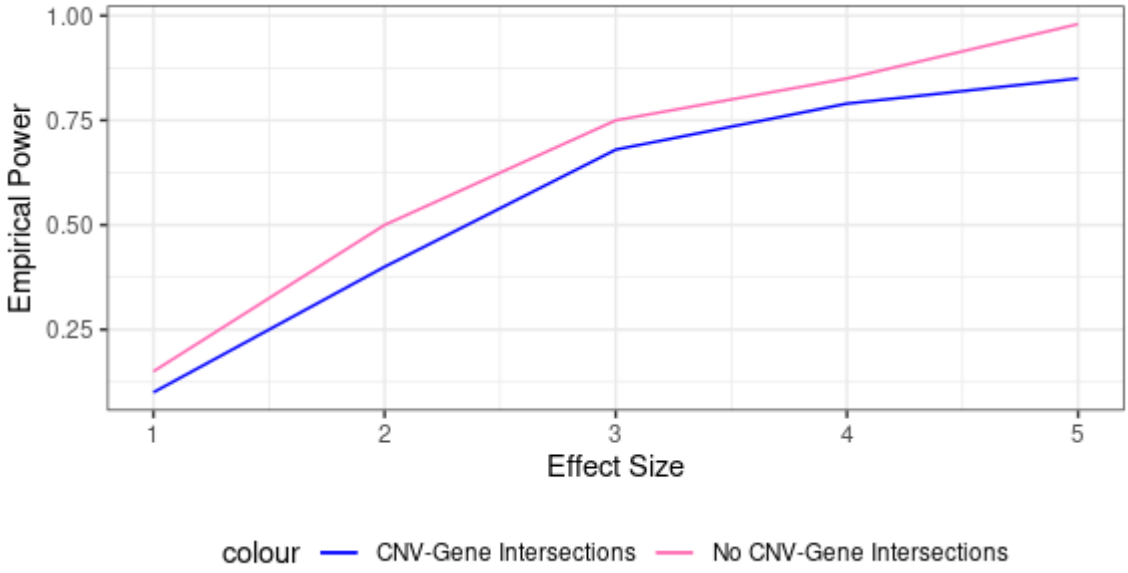

Figure 5.9: Empirical power of MCKAT under CNV-gene intersections and no CNV-gene intersections simulated scenarios.

Figure 5.10: Empirical power of SMCKAT under the CNV-gene intersections and the no CNV-gene intersections simulated scenarios.

5.4 Real data application results

We next conduct an investigation into the dual effect of CNV and genes intersections on the rhabdomyosarcoma CNV data. As explained previously, RMS occurs as two major histological subtypes: embryonal (ERMS) and alveolar (ARMS). The classification of the RMS subtype has a direct effect on the patients' treatment options. The RMS CNV data includes a total of 59,131 CNVs for 25 alveolar and 19 embryonal cancers. We investigate the association between RMS cancer sub type with CNVs, including all CNV characteristics and CNV-gene intersections. We use the list of genes that are identified by Shern et al. (2014) as the embryonal and alveolar classifier genes. These genes are reported in Table 5.4.

Embryonal classifier genes	**Alveolar classifier genes**
EFI4EBP1	TFAP2B
SAE1	CNR1
MFAPP2	NELL1
CPSF1	PIPOX
PGRMC1	CELA2A
PAFAH1B3	ALK
GPX7	NRCAM
FBN2	ASS1
ZIC1	DAPK1
KAZN	WSCD1
HMGA2	PGBD5
ASAP1	FAN1
TRPS1	TOX3
MAK16	TULP4
FZD7	NRN1
CAD	ABCG1
WDYHV1	JARID2
HOXC6	TAGLN3
ARHGEF40	MAGI1
GALNT2	PTBP2

Table 5.2: Genes reported by Shern et al. (2014) as embryonal and alveolar RMS cancer sub types classifier genes.

We divide the RMS CNV profiles into four groups based on whether they have CNV-gene intersections with the genes listed in Table 5.4 or not. This follows the four scenarios designed in the simulation studies. In the first CNV

profile group, there are no CNVs in the profiles that are intersected with the genes. In contrast, for the second and third CNV profile groups, profiles have CNVs that have CNV-gene only deletion or amplification intersections. The fourth CNV profile group, includes all CNV profiles that have CNVs, either deletion or amplification, intersected with the genes. Then, we apply MCKAT and SMCKAT on CNV profile groups to test the association between CNVs and the RMS cancer sub type considering both CNV characteristics and the CNV-gene intersection effect.

First, we conduct MCKAT analysis on each of 23 chromosomes. The p-values of testing the association between RMS subtype and CNVs in each chromosome and for each CNV profile group are reported in Table 5.3. Bonferroni correction is used for adjusting the multiple testing to control the family-wise error rate (FWER) of $\alpha = 0.05$. Since 22 chromosomes and a sex chromosome are being tested, the p-value threshold for a whole-chromosome significance is calculated as $0.05/23 = 2.2 \times 10^{-3}$.

Then, we conduct SMCKAT analysis on chromosomes which contain CNVs that are identified as being significantly associated with the RMS sub types that are, chromosomes 2, 11, 8 and 13, as reported in Table 5.3. The p-values of applying SMCKAT on the aforementioned chromosomes for each CNV profile group with group size of 5 are reported in Table 5.4.

Table 5.3: P-values of testing the association between RMS subtype and CNVs, both CNV characteristics and CNV-gene intersection, in each chromosome. (*) denotes significant association identified by MCKAT.

Chromosome	Scenario 1	Scenario 2	Scenario 3	Scenario 4
chr1	1.032×10^{-1}	1.637×10^{-1}	1.727×10^{-1}	4.316×10^{-1}
chr2	1.022×10^{-3} *	3.129×10^{-3} *	3.245×10^{-3}*	5.312×10^{-3} *
chr3	1.634×10^{-1}	1.829×10^{-1}	1.902×10^{-1}	2.175×10^{-1}
chr4	4.001×10^{-1}	6.023×10^{-1}	6.537×10^{-1}	8.329×10^{-1}
chr5	7.325×10^{-2}	4.021×10^{-1}	4.592×10^{-1}	6.212×10^{-1}
chr6	4.554×10^{-1}	5.031×10^{-1}	5.200×10^{-1}	7.324×10^{-1}
chr7	4.521×10^{-1}	6.635×10^{-1}	6.472×10^{-1}	7.098×10^{-1}
chr8	5.821×10^{-5}*	6.221×10^{-4}*	6.011×10^{-4}*	2.525×10^{-3}*
chr9	4.221×10^{-2}	5.887×10^{-1}	5.925×10^{-1}	8.211×10^{-1}
chr10	9.875×10^{-2}	8.041×10^{-1}	8.425×10^{-1}	9.025×10^{-1}
chr11	1.527×10^{-3}*	1.857×10^{-2}	1.655×10^{-2}	2.652×10^{-1}
chr12	4.524×10^{-1}	5.788×10^{-1}	5.882×10^{-1}	7.354×10^{-1}
chr13	2.462×10^{-3}*	1.241×10^{-3}*	3.916×10^{-1}	2.352×10^{-3}
chr14	1.219×10^{-1}	3.187×10^{-1}	4.613×10^{-1}	4.015×10^{-1}
chr15	4.002×10^{-1}	5.992×10^{-1}	5.679×10^{-1}	7.005×10^{-1}
chr16	1.565×10^{-1}	2.002×10^{-1}	4.960×10^{-1}	2.628×10^{-1}
chr17	3.613×10^{-1}	4.474×10^{-1}	4.788×10^{-1}	5.402×10^{-1}
chr18	2.021×10^{-1}	3.861×10^{-1}	3.723×10^{-1}	4.217×10^{-1}
chr19	1.995×10^{-1}	2.032×10^{-1}	2.045×10^{-1}	3.211×10^{-1}
chr20	6.032×10^{-3}	5.559×10^{-2}	5.487×10^{-2}	7.425×10^{-2}
chr21	4.231×10^{-2}	5.802×10^{-1}	5.922×10^{-1}	7.164×10^{-1}
chr22	2.332×10^{-1}	3.444×10^{-1}	3.827×10^{-1}	5.222×10^{-1}
chr X	8.035×10^{-1}	8.525×10^{-1}	8.917×10^{-1}	9.322×10^{-1}
chr Y	7.622×10^{-1}	8.002×10^{-1}	8.515×10^{-1}	1.114×10^{-1}

Table 5.4: P-values of testing the association between RMS subtype and CNVs, both CNV characteristics and CNV-gene intersections, in chromosomes 2, 11, 8 and 13 with group size of 5. (*) denotes significant association identified by SMCKAT.

Chr.	Scenario 1	Scenario 2	Scenario 3	Scenario 4
2	2.023×10^{-3}*	0.4533×10^{-2}	0.635×10^{-2}	2.237×10^{-2}
8	1.153×10^{-6}*	2.232×10^{-5}*	2.452×10^{-5}*	3.091×10^{-4}*
11	2.311×10^{-4}*	1.031×10^{-3}	1.852×10^{-3}	6.935×10^{-2}
13	1.095×10^{-3}*	1.211×10^{-2}	1.529×10^{-2}	4.863×10^{-1}

5.5 Discussion

Due to the unavailability of the public data that has both CNV characteristics and CNV-gene intersection information, we were not able to investigate the effect of CNV-gene intersection directly by adding another kernel to MCKAT and SMCKAT. However, we were able to investigate this with an indirect approach by conducting four different simulation scenarios.

Comparing the results reported in Tables 5.3 and 5.4 with the results reported in Tables 3.1 and 4.4, both MCKAT and SMCKAT identify the same chromosomes, chromosomes 2, 8, 11 and 13, that contain CNVs having a significant association with the RMS sub type under both scenario types: considering both CNV characteristics and CNV-gene intersection information in testing the association with the disease-related traits, and considering only CNV characteristics in testing the association which we have done in Chapters 3 and 4 as well.

The results are in line with what we observed in the simulation study. As we discussed in Section 5.3, under the scenario that there is no CNV-gene intersection, both MCKAT and SMCKAT, have the same performance as we observed in Chapters 3 and 4 by providing strong evidence, very low p-value significance level, to prove there a significant association between CNVs and disease-related traits.

In scenarios where CNV-gene intersections occur, both MCKAT and SMCKAT provide larger p-values compared to the scenario where no CNV-gene intersections are present. As discussed earlier, we hypothesize that this is because CNV-gene intersection data may be informative in testing the association between CNVs and disease-related traits. However, due to the kernel designs used in both MCKAT and SMCKAT, this information is not utilized.

Neither MCKAT nor SMCKAT, has a kernel to consider CNV-gene intersections data in measuring the similarity between CNV profiles for and use in an association test. We were not able to design a CNV-gene intersection kernel in both MCKAT and SMCAT design due to the unavailability of public data. In more details, we were not able to find CNV datasets that not only have all CNV multi-dimensional characteristics but the CNV interactions with other genetic variations like genes. Therefore, we designed both MCKAT and SMCKAT based on the availability of the the data. However, both MCKAT and SMCKAT can provide a powerful evidence, very small p-value, in testing the association between CNVs and disease-related traits considering a p-value less than 0.05 is typically

considered to be statistically significant, in which cases the null hypothesis should be rejected.

5.6 Conclusion

The results in this chapter provide a demonstration that considering CNV-gene intersection data in addition to the CNV characteristics including chromosomal regions, type and dosage is informative for testing the association between the CNVs and disease-related traits. This is due to the effects of CNVs on their intersected genes that eventually can lead to an association between CNV-gene intersections and the disease-related traits. This works forms Contribution 3 of this thesis which is the demonstration that considering the dual effects of CNV characteristics and CNV-gene intersection is informative in testing the association between CNVs and disease-related traits. This approach can be used to investigate the effects of the CNVs and other genetic variations to achieve more precise insights about the association of the CNVs and different genetic variations together with disease-related traits.

Chapter 6

Conclusion

The study of the copy number variants, their association with disease related traits and their interaction with other genomics events, are developing fields. The research presented in this thesis was aimed at addressing some different, but inter-related, questions in this area. In Chapter 3, the development and assessment of the multidimensional CNV kernel based association test to test the association between the CNVs and disease related traits is presented. In the Chapter 4, the development and assessments of the sequential multidimensional CNV kernel based association test for testing the association between the sequential order of CNVs in addition to their characteristics with disease related traits is presented. Chapter 5 presented an investigation into the association between the dual effects of CNVs and their intersected genes with disease related traits.

In addressing these questions, this research has made three corresponding contributions to knowledge:

Contribution 1: A multi-dimensional kernel-based CNV association test that allows for the detection of CNV chromosomal regions significantly associated with disease related traits and improves on currently available methods for studying CNVs.

Contribution 2: A sequential multi-dimensional CNV kernel-based association test that allows investigating whether CNVs are randomly distributed across the genome, or their order matters and have a significant association with disease related traits.

Contribution 3: The demonstration that considering the effect of CNV-gene intersections in addition to the CNV characteristics is informative and

helpful in identifying the significant association between CNVs and disease related traits.

The following sections discuss the interpretation, significance and limitations of these contributions, along with potential directions for further research, and last section provides a final conclusion to this thesis.

6.1 Identifying the Association between Copy number Variants and Disease related Traits

The multidimensional and sequential multidimensional CNV kernel based association test developed in Chapter 3 and 4 provide a way to identify CNV hot spots at the cytogenetic band level significantly associated with disease related traits. The first method, MCKAT, improves on existing methods by using kernels that are capable of dealing and utilizing the CNV multidimensional characteristics and providing stronger evidence to prove the existence of the associations. MCKAT can provide biologists with a list of CNV chromosomal region that contains CNVs which are significantly associated with a disease related traits, like disease status or cancer sub-type. Currently, biologists conduct an extensive investigation of the whole genome in affected and unaffected individuals to identify these CNV hot spots. MCKAT is applicable to both rare and frequent CNV data related to any diseases.

The second method, SMCKAT, tests the association between the CNV sequential orders and disease-related traits in addition to considering the CNV multi dimensional characteristics. It is the first approach to study the association between CNV sequential order and disease related traits to our knowledge. The motivation behind developing SMCKAT is that SNPs do not usually function individually. They work in coordination with other SNPs to manifest a disease or trait. Therefore, many sequence studies have been done to test the association between SNPs and disease or traits. However, the association between the sequential order of CNVs and disease-related traits had not been studied, to our knowledge, and it had been unclear that CNVs function individually or whether they work in coordination with other CNVs. Applying SMCKAT on CNV data, we observe the CNV sequential order has a significant association with disease related traits in some chromosomal regions. SMCKAT is more stringent compared

with the state-of-the-art approaches in detecting significant CNV regions due to stricter rules used in its design for measuring the similarity between the CNV profiles. Like MCKAT, SMCKAT is applicable to both rare and frequent CNV data related to any diseases.

While the main aim of this work is the development of association tests to test the association between CNVs and disease related traits, we conduct an investigation into the association between dual effects of CNVs and their intersected genes. The motivation of this investigation was the effect of CNVs on genes when they are intersected, which is explained in detail in Chapter 5. Based on the results, considering the dual effects of CNV characteristics and CNV-gene intersections is informative in testing the association between CNVs and disease related traits. This approach can be used to investigate the effects of the CNVs and other genetic variations to achieve more precise insights about the association of the CNVs and different genetic variations together with disease related traits.

6.2 Work Limitations and Future Works

There are some limitations to the work presented in Chapters 3, 4 and 5. Firstly, due to the existence of few CNV association tests, based on the literature, our proposed methods were compared to few methods. Our methods were applied on both simulated and real data. Their performances were evaluated and compared with all existing methods. However, testing our proposed methods on more available CNV data can help us to find out our proposed methods' limitations if there is any.

Secondly, there are few publicly available CNV datasets, specifically frequent CNV datasets. In addition, the available ones do not have all CNV characteristics including chromosomal regions, type and dosage. Therefore, for both MCKAT and SMCKAT, we had to simulate CNV data for our analysis in addition to using real datasets.

Last but not least, based on our knowledge, there is no data available including interactions between CNVs and other genetic variations like genes and SNPs. Having this type of data would be a good opportunity for further exploring of the association between the dual effect CNVs and other genomic events with disease related trait.

In our future work, we will expand both the MCKAT and SMCKAT frameworks

to be applicable to both qualitative and quantitative traits by using other methods other than logistic regression. Furthermore, we will revise MCKAT and SMCKAT designs by adding another kernel which is responsible for measuring the similarity between CNV profiles with respect to the interaction between CNVs and other genetic variations.

6.3 Conclusion

The research presented in this thesis represents advances in several different aspects of the analysis of the association between CNVs and disease related traits. The potential applications and extensions of this work are varied, with relevance to the through understanding of the biology of the CNVs, the continuing improvement of methods to detect precise CNVs, the availability of CNV data and their interactions with other genetic variations.

Chapter 7

Appendix

The R code for the MCKAT and SMCKAT models are given in the following pages. The main kernels are single-pair CNV kernel, whole chromosome CNV kernel, pair CNV group kernel and whole genome CNV group Kernel. The code for kernel-based association test is adapted from Zhan et al. (2016).

```
SinglePairCNVKernel <-
  function(x,y){            ##x,y are 4-d vectors
    if (length(interval_intersection(interval(x[1],x[2]),interval(y[1],y[2])))==0)
     { TotalSmilarity=0 }
    if (length(interval_intersection(interval(x[1],x[2]),interval(y[1],y[2])))!=0)
     {
      JacIndex= interval_measure(interval_intersection(interval(x[1],x[2]),interval(y[1],y[2])))/
      interval_measure(interval_union(interval(x[1],x[2]),interval(y[1],y[2])))  ##Contribution of length
      TypeCont= (((x[3]==y[[3]])+1)/2)     ##Contribution of type
      DosCont= 1/2^abs((abs(2-x[4])-abs(2-y[4])))     ##contribution of dosage
      TotalSmilarity= JacIndex*TypeCont*DosCont
    }
    return(TotalSmilarity)
  }
WholeChromosomeCNVKernel <-
  function(x,y){   ## x,y are matrices
    p=nrow(x)
    q=nrow(y)
    Pairsim=matrix(0,nrow =p,ncol =q)
    if(p*q==0){TotalSim=0}
    if(p*q!=0){
    for ( i in 1:p)
    {
    for (j in 1:q)
    {
    Pairsim[i,j]=KCNVPairSim(x[i,],y[j,])
    }
    }
      TotalSim=sum(rowSums(Pairsim))
    }
    return(TotalSim)
  }
MakeSimMatrix <-
  function(InputCNVData){
    SimMat=matrix(,nrow = length(InputCNVData),ncol=length(InputCNVData))
    for(i in 1:length(InputCNVData)) {
      for(j in 1:length(InputCNVData)){
        SimMat[i,j]=KCNVTotal(InputCNVData[[i]],InputCNVData[[j]])
      }
    }
    return(SimMat)
  }
```

```
PairCNVGroupKernel <- function(x,y,mersize){   ##x,y are matrices with 4 columns and kmersize rows
  Pairsim=list()
  for(m in 1:mersize){
    if (length(interval_intersection(interval(x[m,1],x[m,2]),interval(y[m,1],y[m,2])))==0)
    { Pairsim[m]=0 }
    if (length(interval_intersection(interval(x[m,1],x[m,2]),interval(y[m,1],y[m,2])))!=0)
    {
      JacIndex= interval_measure(interval_intersection(interval(x[m,1],x[m,2]),interval(y[m,1],y[m,2])))/
      interval_measure(interval_union(interval(x[m,1],x[m,2]),interval(y[m,1],y[m,2])))
       ##Contribution of length
      TypeCont= (((x[m,3]==y[[m,3]])+1)/2)     ##Contribution of type
      DosCont= 1/2^abs((abs(2-x[m,4])-abs(2-y[m,4])))     ##contribution of dosage
      Pairsim[m]= (JacIndex*TypeCont*DosCont)/mersize
    }
  }
  return((Reduce("+",Pairsim)))
}
WholeGenomeCNVGroupKernel <-
  function(x,y){   ## x,y are list of matrices
    p=length(x)
    q=length(y)
    if(p*q==0){TotalSim=0}
    if(p==q&q>0){
      Pairsim=rep(0,p)
      for ( i in 1:p)
      {
        Pairsim[i]=KmerPairSim(x[[i]],y[[i]],1)
      }
      TotalSim=sum(Pairsim)
    }
    if(p>q&q>0){
      SPairsim=rep(0,p-q+1)
      for ( l in 0:(p-q)){
        Pairsim=rep(0,q)
        for ( k in 1:q)
```

```
36          {
37            Pairsim[k]=KmerPairSim(x[[k+l]],y[[k]],1)
38
39          }
40          SPairsim[(l+1)]=sum(Pairsim)
41        }
42        TotalSim=max(SPairsim)
43      }
44      if(q>p&p>0){
45        SPairsim=rep(0,q-p+1)
46        for ( l in 0:(q-p)){
47          Pairsim=rep(0,p)
48          for ( k in 1:p)
49          {
50            Pairsim[k]=KmerPairSim(x[[k]],y[[k+l]],1)
51
52          }
53          SPairsim[(l+1)]=sum(Pairsim)
54        }
55        TotalSim=max(SPairsim)
56      }
57      return(TotalSim)
58    }
59    MakeSimMatrix <-
60    function(InputCNVData){
61      SimMat=matrix(,nrow = length(InputCNVData),ncol=length(InputCNVData))
62      for(i in 1:length(InputCNVData)) {
63        for(j in 1:length(InputCNVData)){
64          SimMat[i,j]=KCNVTotal(InputCNVData[[i]],InputCNVData[[j]])
65        }
66      }
67      return(SimMat)
68    }
```

```r
CNVAssociationTest <-
  function (y, K,  X=NULL) {
    n <- length(y)
    if (is.null(X)) {
      X1 <-  matrix(rep(1, length(y)), ncol=1)
    } else {
      X1 <- model.matrix(~. , as.data.frame(X))
    }

    glmfit <- glm(y ~ X1-1, family = binomial)

    betas <- glmfit$coef
    mu  <- glmfit$fitted.values
    eta <- glmfit$linear.predictors
    res.wk <- glmfit$residuals
    res   <- y - mu

    w   <- mu*(1-mu)
    sqrtw <- sqrt(w)

    adj <- sum((sqrtw * res.wk)^2)

    DX12 <- sqrtw * X1


    qrX <- qr(DX12, tol = 1e-7)
    Q <- qr.Q(qrX)
    Q <- Q[, 1:qrX$rank, drop=FALSE]

    P0 <- diag(length(y)) - Q %*% t(Q)

    DKD <- tcrossprod(sqrtw) * K
    tQK <- t(Q) %*% DKD
    QtQK <- Q %*% tQK
    PKP1 <- DKD - QtQK - t(QtQK) + Q %*% (tQK %*% Q) %*% t(Q)
    q1 <- as.numeric(res %*% K %*% res)
    q1 = q1 / adj
    ee1 = eigen(PKP1 - q1 * P0, symmetric = T, only.values=T)
    lambda1 = ee1$values[abs(ee1$values) >= 1e-10]
    p1 <- davies(0, lambda=lambda1, acc=1e-6)$Qq

    return(p1)
  }
```

Bibliography

Arlt, M. F., Rajendran, S., Birkeland, S. R., Wilson, T. E. & Glover, T. W., 2014, 'Copy number variants are produced in response to low-dose ionizing radiation in cultured cells', *Environmental and molecular mutagenesis*, vol. 55, no. 2, pp. 103–113.

Bromberg, Y., Yachdav, G. & Rost, B., 2008, 'Snap predicts effect of mutations on protein function', *Bioinformatics*, vol. 24, no. 20, pp. 2397–2398.

Brucker, A., Lu, W., Marceau West, R., Yu, Q.-Y., Hsiao, C. K., Hsiao, T.-H., Lin, C.-H., Magnusson, P. K., Sullivan, P. F., Szatkiewicz, J. P. et al., 2020, 'Association test using copy number profile curves (concur) enhances power in rare copy number variant analysis', *PLoS computational biology*, vol. 16, no. 5, p. e1007797.

Cerruti Mainardi, P., 2006, 'Cri du chat syndrome', *Orphanet journal of rare diseases*, vol. 1, no. 1, pp. 1–9.

Consortium, . G. P. et al., 2015, 'A global reference for human genetic variation', *Nature*, vol. 526, no. 7571, p. 68.

Cooper, G. M., Nickerson, D. A. & Eichler, E. E., 2007, 'Mutational and selective effects on copy-number variants in the human genome', *Nature genetics*, vol. 39, no. 7, pp. S22–S29.

Cuccaro, D., De Marco, E. V., Cittadella, R. & Cavallaro, S., 2017, 'Copy number variants in alzheimer's disease', *Journal of Alzheimer's Disease*, vol. 55, no. 1, pp. 37–52.

Davies, R. B., 1980, 'The distribution of a linear combination of $\chi 2$ random variables', *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 29, no. 3, pp. 323–333.

De Smith, A., Walters, R., Froguel, P. & Blakemore, A., 2008, 'Human genes involved in copy number variation: mechanisms of origin, functional effects and implications for disease', *Cytogenetic and genome research*, vol. 123, no. 1-4, pp. 17–26.

Degtyareva, A. O., Antontseva, E. V. & Merkulova, T. I., 2021, 'Regulatory snps: altered transcription factor binding sites implicated in complex traits and diseases', *International journal of molecular sciences*, vol. 22, no. 12, p. 6454.

Dering, C., Hemmelmann, C., Pugh, E. & Ziegler, A., 2011, 'Statistical analysis of rare sequence variants: an overview of collapsing methods', *Genetic epidemiology*, vol. 35, no. S1, pp. S12–S17.

Dering, C., König, I. R., Ramsey, L. B., Relling, M. V., Yang, W. & Ziegler, A., 2014, 'A comprehensive evaluation of collapsing methods using simulated and real data: excellent annotation of functionality and large sample sizes required', *Frontiers in genetics*, vol. 5, p. 323.

El Demellawy, D., McGowan-Jordan, J., De Nanassy, J., Chernetsova, E. & Nasr, A., 2017, 'Update on molecular findings in rhabdomyosarcoma', *Pathology*, vol. 49, no. 3, pp. 238–246.

Elder, P. J., Ramsden, D. B., Burnett, D., Weickert, M. O. & Barber, T. M., 2018, 'Human amylase gene copy number variation as a determinant of metabolic state', *Expert Review of Endocrinology & Metabolism*, vol. 13, no. 4, pp. 193–205.

Fellermann, K., Stange, D. E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C. L., Reinisch, W., Teml, A., Schwab, M., Lichter, P. et al., 2006, 'A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to crohn disease of the colon', *The American Journal of Human Genetics*, vol. 79, no. 3, pp. 439–448.

Feuk, L., Carson, A. R. & Scherer, S. W., 2006, 'Structural variation in the human genome', *Nature Reviews Genetics*, vol. 7, no. 2, p. 85.

Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J., 2009, 'Human genetic variation and its contribution to complex traits', *Nature Reviews Genetics*, vol. 10, no. 4, pp. 241–251.

Freitag, C. M., Agelopoulos, K., Huy, E., Rothermundt, M., Krakowitzky, P., Meyer, J., Deckert, J., Von Gontard, A. & Hohoff, C., 2010, 'Adenosine a 2a receptor gene (adora2a) variants may increase autistic symptoms and anxiety in autism spectrum disorder', *European Child & Adolescent Psychiatry*, vol. 19, no. 1, pp. 67–74.

Girirajan, S., Brkanac, Z., Coe, B. P., Baker, C., Vives, L., Vu, T. H., Shafer, N., Bernier, R., Ferrero, G. B., Silengo, M. et al., 2011, 'Relative burden of large CNVs on a range of neurodevelopmental phenotypes', *PLoS Genet*, vol. 7, no. 11, p. e1002334.

Glessner, J. T., Wang, K., Cai, G., Korvatska, O., Kim, C. E., Wood, S., Zhang, H., Estes, A., Brune, C. W., Bradfield, J. P. et al., 2009, 'Autism genome-wide copy number variation reveals ubiquitin and neuronal genes', *Nature*, vol. 459, no. 7246, pp. 569–573.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J. et al., 2005, 'The influence of ccl3l1 gene-containing segmental duplications on hiv-1/aids susceptibility', *Science*, vol. 307, no. 5714, pp. 1434–1440.

Greenblum, S., Carr, R. & Borenstein, E., 2015, 'Extensive strain-level copy-number variation across human gut microbiome species', *Cell*, vol. 160, no. 4, pp. 583–594.

Han, F. & Pan, W., 2010, 'A data-adaptive sum test for disease association with multiple common or rare variants', *Human heredity*, vol. 70, no. 1, pp. 42–54.

Harteveld, C. L. & Higgs, D. R., 2010, 'α-thalassaemia', *Orphanet journal of rare diseases*, vol. 5, no. 1, pp. 1–21.

Higgs, D. R., Old, J., Clegg, J. B., Pressley, L., Hunt, D., Weatherall, D. & Serjeant, G., 1979, 'Negro α-thalassaemia is caused by deletion of a single α-globin gene', *The Lancet*, vol. 314, no. 8137, pp. 272–276.

Hood, L. & Rowen, L., 2013, 'The human genome project: big science transforms biology and medicine', *Genome medicine*, vol. 5, pp. 1–8.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L. et al., 2007, 'Paired-end mapping reveals extensive structural variation in the human genome', *Science*, vol. 318, no. 5849, pp. 420–426.

Kubiritova, Z., Gyuraszova, M., Nagyova, E., Hyblova, M., Harsanyova, M., Budis, J., Hekel, R., Gazdarica, J., Duris, F., Kadasi, L. et al., 2019, 'On the critical evaluation and confirmation of germline sequence variants identified using massively parallel sequencing', *Journal of biotechnology*, vol. 298, pp. 64–75.

La Cognata, V., Morello, G., D'Agata, V. & Cavallaro, S., 2017, 'Copy number variability in parkinson's disease: assembling the puzzle through a systems biology approach', *Human genetics*, vol. 136, no. 1, pp. 13–37.

Levinson, D. F., Duan, J., Oh, S., Wang, K., Sanders, A. R., Shi, J., Zhang, N., Mowry, B. J., Olincy, A., Amin, F. et al., 2011, 'Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and vipr2 duplications', *American Journal of Psychiatry*, vol. 168, no. 3, pp. 302–316.

Lewis, C. M. & Knight, J., 2012, 'Introduction to genetic association studies', *Cold Spring Harbor Protocols*, vol. 2012, no. 3, pp. pdb–top068163.

Liu, D., Ghosh, D. & Lin, X., 2008, 'Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models', *BMC bioinformatics*, vol. 9, no. 1, p. 292.

Liu, D., Lin, X. & Ghosh, D., 2007, 'Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models', *Biometrics*, vol. 63, no. 4, pp. 1079–1088.

Lupski, J. R., de Oca-Luna, R. M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B. J., Saucedo-Cardenas, O., Barker, D. F., Killian, J. M., Garcia, C. A. et al., 1991, 'Dna duplication associated with charcot-marie-tooth disease type 1a', *Cell*, vol. 66, no. 2, pp. 219–232.

Lupski, J. R. & Stankiewicz, P., 2005, 'Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes', *PLoS genetics*, vol. 1, no. 6, p. e49.

Madsen, B. E. & Browning, S. R., 2009, 'A groupwise association test for rare mutations using a weighted sum statistic', *PLoS genetics*, vol. 5, no. 2, p. e1000384.

Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S., Antaki, D., Shetty, A., Holmans, P. A., Pinto, D. et al., 2017, 'Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects', *Nature genetics*, vol. 49, no. 1, p. 27.

Masson, E., Le Maréchal, C., Delcenserie, R., Chen, J.-M. & Férec, C., 2008, 'Hereditary pancreatitis caused by a double gain-of-function trypsinogen mutation', *Human genetics*, vol. 123, no. 5, pp. 521–529.

Maus Esfahani, N., Catchpoole, D. & Kennedy, P. J., 2021a, 'Smckat, a sequential multi-dimensional cnv kernel-based association test', *Life*, vol. 11, no. 12, p. 1302.

Maus Esfahani, N., Catchpoole, D., Khan, J. & Kennedy, P. J., 2021b, 'Mckat: a multi-dimensional copy number variant kernel association test', *BMC bioinformatics*, vol. 22, no. 1, pp. 1–16.

McCarroll, S. A. & Altshuler, D. M., 2007, 'Copy-number variation and association studies of human disease', *Nature genetics*, vol. 39, no. Suppl 7, pp. S37–S42.

McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P., Zody, M. C., Hall, J. L., Brant, S. R., Cho, J. H. et al., 2008, 'Deletion polymorphism upstream of irgm associated with altered irgm expression and crohn's disease', *Nature genetics*, vol. 40, no. 9, p. 1107.

Mohajeri, M. H., Brummer, R. J., Rastall, R. A., Weersma, R. K., Harmsen, H. J., Faas, M. & Eggersdorfer, M., 2018, 'The role of the microbiome for human health: from basic science to clinical applications', *European journal of nutrition*, vol. 57, no. 1, pp. 1–14.

Morgenthaler, S. & Thilly, W. G., 2007, 'A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast)', *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 615, no. 1-2, pp. 28–56.

Nathans, J., Thomas, D. & Hogness, D. S., 1986, 'Molecular genetics of human color vision: the genes encoding blue, green, and red pigments', *Science*, vol. 232, no. 4747, pp. 193–202.

Nguyen, D.-Q., Webber, C. & Ponting, C. P., 2006, 'Bias of selection on human copy-number variants', *PLoS genetics*, vol. 2, no. 2, p. e20.

Nishimura, R., Takita, J., Sato-Otsubo, A., Kato, M., Koh, K., Hanada, R., Tanaka, Y., Kato, K., Maeda, D., Fukayama, M. et al., 2013, 'Characterization of genetic lesions in Rhabdomyosarcoma using a high-density single nucleotide polymorphism array', *Cancer science*, vol. 104, no. 7, pp. 856–864.

Nowakowska, B., 2017, 'Clinical interpretation of copy number variants in the human genome', *Journal of applied genetics*, vol. 58, no. 4, pp. 449–457.

Pfister, N., Bühlmann, P., Schölkopf, B. & Peters, J., 2018, 'Kernel-based tests for joint independence', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 80, no. 1, pp. 5–31.

Poole, A. C., Goodrich, J. K., Youngblut, N. D., Luque, G. G., Ruaud, A., Sutter, J. L., Waters, J. L., Shi, Q., El-Hadidi, M., Johnson, L. M. et al., 2019, 'Human salivary amylase gene copy number impacts oral and gut microbiomes', *Cell host & microbe*, vol. 25, no. 4, pp. 553–564.

Pös, O., Radvanszky, J., Buglyó, G., Pös, Z., Rusnakova, D., Nagy, B. & Szemes, T., 2021, 'Dna copy number variation: Main characteristics, evolutionary significance, and pathological aspects', *biomedical journal*, vol. 44, no. 5, pp. 548–559.

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L.-J. & Sunyaev, S. R., 2010, 'Pooled association tests for rare variants in exon-resequencing studies', *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 832–838.

Radvanszky, J., Surovy, M., Polak, E. & Kadasi, L., 2013, 'Uninterrupted cctg tracts in the myotonic dystrophy type 2 associated locus', *Neuromuscular Disorders*, vol. 23, no. 7, pp. 591–598.

Ramensky, V., Bork, P. & Sunyaev, S., 2002, 'Human non-synonymous snps: server and survey', *Nucleic acids research*, vol. 30, no. 17, pp. 3894–3900.

Rees, E., Kendall, K., Pardiñas, A. F., Legge, S. E., Pocklington, A., Escott-Price, V., MacCabe, J. H., Collier, D. A., Holmans, P., O'Donovan, M. C. et al., 2016, 'Analysis of intellectual disability copy number variants for association with schizophrenia', *JAMA psychiatry*, vol. 73, no. 9, pp. 963–969.

Rees, E., Kirov, G., Sanders, A., Walters, J. T. R., Chambert, K., Shi, J., Szatkiewicz, J., O'dushlaine, C., Richards, A. L., Green, E. K. et al., 2014, 'Evidence that duplications of 22q11. 2 protect against schizophrenia', *Molecular Psychiatry*, vol. 19, no. 1, p. 37.

Rehman, M. Y. A., Briedé, J. J., van Herwijnen, M., Krauskopf, J., Jennen, D. G., Malik, R. N. & Kleinjans, J. C., 2022, 'Integrating snps-based genetic risk factor with blood epigenomic response of differentially arsenic-exposed rural subjects reveals disease-associated signaling pathways', *Environmental Pollution*, vol. 292, p. 118279.

Schrider, D. R. & Hahn, M. W., 2010, 'Gene copy-number polymorphism in nature', *Proceedings of the Royal Society B: Biological Sciences*, vol. 277, no. 1698, pp. 3213–3221.

Semagn, K., Babu, R., Hearne, S. & Olsen, M., 2014, 'Single nucleotide polymorphism genotyping using kompetitive allele specific pcr (kasp): overview of the technology and its application in crop improvement', *Molecular breeding*, vol. 33, no. 1, pp. 1–14.

Sener, E. F., 2014, 'Association of copy number variations in autism spectrum disorders: a systematic review', *Chinese Journal of Biology*, vol. 2014.

Shawe-Taylor, J., Cristianini, N. et al., 2004, *Kernel methods for pattern analysis*, Cambridge university press.

Shern, J. F., Chen, L., Chmielecki, J., Wei, J. S., Patidar, R., Rosenberg, M., Ambrogio, L., Auclair, D., Wang, J., Song, Y. K. et al., 2014, 'Comprehensive genomic analysis of Rhabdomyosarcoma reveals a landscape of alterations affecting a common genetic axis in fusion-positive and fusion-negative tumors', *Cancer Discovery*, vol. 4, no. 2, pp. 216–231.

Song, F., Han, G., Bai, Z., Peng, X., Wang, J. & Lei, H., 2015, 'Alzheimer's disease: genomics and beyond', *International review of neurobiology*, , vol. 121Elsevier, pp. 1–24.

Stefano, V., Viviana, C., Deny, M., Viola, A., Sara, L., Antonio, N. & Marco, T., 2019, 'Copy number variants in autism spectrum disorders', *Progress in Neuro-Psychopharmacology and Biological Psychiatry*.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., De Grassi, A., Lee, C. et al., 2007, 'Relative impact of nucleotide and copy number variation on gene expression phenotypes', *Science*, vol. 315, no. 5813, pp. 848–853.

Sun, X., Guo, W., Shen, J. K., Mankin, H. J., Hornicek, F. J. & Duan, Z., 2015, 'Rhabdomyosarcoma: advances in molecular and cellular biology', *Sarcoma*, vol. 2015.

Tijo, J. H. & Levan, A., 2004, 'The chromosome number of man', *Landmarks in Medical Genetics: Classic Papers with Commentaries*, vol. 42, no. 51, p. 68.

Tzeng, J.-Y., Magnusson, P. K., Sullivan, P. F., Szatkiewicz, J. P., Consortium, S. S. et al., 2015, 'A new method for detecting associations with rare copy-number variants', *PLoS genetics*, vol. 11, no. 10, p. e1005403.

Visscher, P. M., Yengo, L., Cox, N. J. & Wray, N. R., 2021, 'Discovery and implications of polygenicity of common diseases', *Science*, vol. 373, no. 6562, pp. 1468–1473.

Vorstman, J. A., Parr, J. R., Moreno-De-Luca, D., Anney, R. J., Nurnberger Jr, J. I. & Hallmayer, J. F., 2017, 'Autism genetics: opportunities and challenges for clinical translation', *Nature Reviews Genetics*, vol. 18, no. 6, p. 362.

Wang, C. K., Xu, M. S., Ross, C. J., Lo, R., Procyshyn, R. M., Vila-Rodriguez, F., White, R. F., Honer, W. G. & Barr, A. M., 2015, 'Development of a cost-efficient novel method for rapid, concurrent genotyping of five common single nucleotide polymorphisms of the brain derived neurotrophic factor (bdnf) gene by tetra-primer amplification refractory mutation system', *International journal of methods in psychiatric research*, vol. 24, no. 3, pp. 235–244.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J. & Lin, X., 2010, 'Powerful snp-set analysis for case-control genome-wide association studies', *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 929–942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X., 2011, 'Rare-variant association testing for sequencing data with the sequence kernel association test', *The American Journal of Human Genetics*, vol. 89, no. 1, pp. 82–93.

Yim, S.-H., Jung, S.-H., Chung, B. & Chung, Y.-J., 2015, 'Clinical implications of copy number variations in autoimmune disorders', *The Korean journal of internal medicine*, vol. 30, no. 3, p. 294.

Yingjun, X., Haiming, Y., Mingbang, W., Liangying, Z., Jiaxiu, Z., Bing, S., Qibin, Y. & Xiaofang, S., 2017, 'Copy number variations independently induce autism spectrum disorder', *Bioscience reports*, vol. 37, no. 4.

Yue, P., Melamud, E. & Moult, J., 2006, 'Snps3d: candidate gene and snp selection for association studies', *BMC bioinformatics*, vol. 7, no. 1, pp. 1–15.

Zhan, X., Epstein, M. P. & Ghosh, D., 2015a, 'An adaptive genetic association test using double kernel machines', *Statistics in biosciences*, vol. 7, no. 2, pp. 262–281.

Zhan, X., Girirajan, S., Zhao, N., Wu, M. C. & Ghosh, D., 2016, 'A novel copy number variants kernel association test with application to autism spectrum disorders studies', *Bioinformatics*, vol. 32, no. 23, pp. 3603–3610.

Zhan, X., Patterson, A. D. & Ghosh, D., 2015b, 'Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data', *Bmc Bioinformatics*, vol. 16, no. 1, p. 77.

Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R., 2009, 'Copy number variation in human health, disease, and evolution', *Annual review of genomics and human genetics*, vol. 10, p. 451.

Zhang, J., Yang, J., Zhang, L., Luo, J., Zhao, H., Zhang, J. & Wen, C., 2021, 'A new snp genotyping technology target snp-seq and its application in genetic analysis of cucumber varieties (vol 10, 5623, 2020)', *SCIENTIFIC REPORTS*, vol. 11, no. 1.

Zigman, W. B., 2013, 'Atypical aging in down syndrome', *Developmental disabilities research reviews*, vol. 18, no. 1, pp. 51–67.