ABSTRACT

BRUCKER, AMANDA RUTH. Association Testing and Detection of Copy Number Variants on the Global and Local Scale. (Under the direction of Jung-Ying Tzeng.)

Copy number variants are a type of structural genetic variant that comprise a large proportion of variation in the human genome, and our work here focuses on detecting and estimating their association with disease traits and phenotypes. Copy number variants (CNVs) are the relative gain or loss of DNA segments at least 1000 base pairs long (1kb), and have been found to carry relatively strong risks for neurodevelopmental and neuropsychiatric disorders. CNVs may drive or mitigate risk by interrupting genes or regulatory regions. An excess burden of rare CNVs (< 1% in the population) in particular has been found to correspond to higher risk of schizophrenia, bipolar disorder, autism spectrum disorder, and other complex polygenic traits. Researchers study these rare variants effects with collapsing methods that detect signals by aggregating the effects of multiple variants across loci and smoothing methods that borrow information across variants to estimate their effects. Here we propose two such methods which are suitable for both rare and common CNV analysis.

Following a literature and methods review in Chapter 1, in Chapter 2 we develop a kernel-based association test for rare CNV analysis. The semiparametric kernel machine regression framework is a powerful tool in modeling the pooled effects of rare variants across a test region, which individually may be heterogeneous in size and direction and may include interaction and nonlinear effects. To perform a kernel association test, a CNV locus or region needs to be defined so that locus-specific effects can be retained during aggregation. However, there is no universal definition of CNV loci since their exact locations and lengths will vary within and between studies. Different locus definitions can lead to different performance depending on the underlying effect patterns. In this work, we introduce a new kernel-based test called CONCUR (i.e., copy number profile curve-based association test) that is free from a definition of locus and evaluates CNV-phenotype associations by comparing individuals’ copy number profiles across the genomic regions. CONCUR is built on the proposed concepts of “copy number profile curves” to describe the CNV profile of an individual, and the “common area under the curve (cAUC) kernel” to model the multi-feature CNV effects. The proposed method captures the effects of CNV dosage and length, accounts for the numerical nature of copy numbers, and accommodates between- and within-locus heterogeneous effects without the need to define artificial CNV loci as required in current kernel methods. In a variety of simulation settings, CONCUR shows comparable or improved power over existing approaches. Real data analyses suggest that CONCUR is well powered to detect CNV effects in the Swedish Schizophrenia Study and the Taiwan Biobank.

Whereas the CONCUR kernel-based test is designed for genome-wide analysis, in Chapter 3 we propose a novel lasso-based framework for CNV analysis on the local scale. Following association testing, few methods exist for post-hoc CNV analysis of a region containing a significant signal. Visualization tools may highlight hotspots of CNV activity in the test region, which may be individually
regressed against the phenotype if the CNV frequency is large enough. However this approach will have no power to characterize rare variant effects, will again require the construction of CNV loci to test within, and will be laborious or intractable given a large number of hotspots. Our proposed method offers a formal modeling framework for the post-hoc analysis setting which estimates effects over the test region using a weighted fused lasso. The approach smooths rare variant effects towards their neighbors while detecting heterogeneous signals in high activity regions if they are present. In our penalized regression framework, a lasso penalty encourages a sparse solution and the fusion penalty imposes locally constant estimates, an assumption made in several CNV calling algorithms and which we believe is biologically plausible. In simulations and real data analysis we show the performance of the proposed method and other benchmark lasso approaches.
Association Testing and Detection of Copy Number Variants on the Global and Local Scale

by
Amanda Ruth Brucker

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina
2020

APPROVED BY:

________________________________________
Wenbin Lu

________________________________________
Arnab Maity

________________________________________
David Reif

________________________________________
Jung-Ying Tzeng
Chair of Advisory Committee
DEDICATION

To my family by birth and the family I found.
Amanda Ruth Brucker was born in Kansas City, KS in 1993 and grew up in the Midwest. She attended Cornell College studying mathematics and physics. After participating in an applied math REU in Lincoln, NE and a SIBS summer program in Pittsburgh, PA, she was inspired to pursue a graduate degree in statistics. In a gap year she worked in Washington, DC at the National Cancer Institute as a research fellow which opened the door to joining the NC State statistics program.

While progressing through her PhD, Amanda worked at the Duke Clinical Research Institute in Durham, NC, and found lifelong friends and mentors. After graduating, she will begin work as a biostatistician in the Duke Biostatistics and Bioinformatics Department.
ACKNOWLEDGEMENTS

First of all, I thank Jung-Ying Tzeng for her mentorship over the past 4 years and for always helping me grow as a statistical thinker, writer, and communicator. I am equally grateful to Wenbin Lu, whose relentless new ideas and keen insight got me out of many dead ends. I thank my committee members Denis Fourches for his incisive feedback on my preliminary exam, Arnab Maity for his good ideas at my research group presentations, and David Reif and Eric Chi for graciously standing in my committee last minute. I am also grateful to my collaborators Jin Szatkiewicz and Tzu-Pin Lu who were instrumental in our first paper, and to Rachel Marceau West for her edits and helping me find my footing in my first project. Thanks to the many great professors in the NC State statistics department, and thanks to Lanakila and Alison for keeping me on track. In addition, many thanks to my DCRI mentors Karen Chiswell and Ben Goldstein for helping me grow as an applied statistician.

Thanks to the 2014 team first year students for the hangouts, study groups, and homework sessions. Special thanks to Munir, Stephanie, Pat, Matt, and Isaac. Thanks to amazing TAs Suhyun, Shuhan, Debraj, and basic bootcamp leaders Ajay and Marshall. Thanks to my research group members Meng, Yueyang, Som, and David. Thanks to Mike Carter for his detailed feedback on my early thesis drafts. I also owe so much to the Dissertation Completion group, and I'm so grateful to have met all of you. A special thanks to Josephine for keeping up the regular writing sessions that I still love attending. Thanks to Tara, Julia, and Christine for being my support network. Many thanks to David and Joyce for helping me get through my final year and cross the finish line.

Thanks to my family for always being there. Tom, thanks for your patience, support, and love.
# TABLE OF CONTENTS

## LIST OF TABLES

<table>
<thead>
<tr>
<th>Chapter 1</th>
<th>Introduction</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.1 Kernel Machine Regression for Genetic Data Analysis</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1.1.1 Basic kernel machine</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1.1.2 Kernel machine regression for Gaussian response</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1.1.3 Connection to Linear Mixed Models</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>1.1.4 Kernel machine regression for exponential family responses</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>1.1.5 Kernel-based association tests</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>1.2 Lasso Methods for Genetic Data Analysis</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>1.2.1 Standard lasso for linear models</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>1.2.2 Fused lasso and generalized lasso for linear models</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>1.2.3 Constrained lasso for linear models</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>1.2.4 Constrained lasso for logistic regression</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>1.2.5 Additional topics</td>
<td>18</td>
</tr>
</tbody>
</table>

## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Chapter 2</th>
<th>Association Test Using Copy Number Profile Curves (CONCUR) Enhances Power in Rare Copy Number Variant Analysis</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.1 Introduction</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>2.2 Results</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>2.2.1 Overview of CONCUR</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>2.2.2 Simulation Studies</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>2.2.3 Real data application</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>2.3 Discussion</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>2.4 Materials and methods</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>2.4.1 Ethics statement</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>2.4.2 CONCUR method</td>
<td>45</td>
</tr>
</tbody>
</table>

## Chapter 3 | Weighted fused lasso for local analysis of copy number variants | 49 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.1 Introduction</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>3.2 Methods</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>3.2.1 Weighted sparse fused lasso</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>3.2.2 Generalized lasso formulation</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>3.2.3 Choice of weights</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>3.2.4 Model selection with $\lambda$</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>3.2.5 Connection to CNV profile curves</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>3.3 Simulations</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>3.3.1 Simulation model</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>3.3.2 CNV effects $\beta$</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>3.3.3 Implementation</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>3.3.4 Performance metrics</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>3.4 Results</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>3.4.1 Simulations</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>3.5 Real Data Analysis</td>
<td>66</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 2.1  **Type I error rates.** Type I error rates of three CNV tests evaluated based on 5000 replications. ................................................................. 31

Table 2.2  **Association test results for the effects of CNVs with > 100kb in length on schizophrenia risk in the Swedish Schizophrenia Study.** Raw p-values are reported for CONCUR, CCRET, and CKAT. Asterisks indicate p-values that were significant after a Holm multiple-testing adjustment. Pathways are ordered alphabetically. ................................................................. 39

Table 2.3  **Association test results for the effects of CNVs on triglyceride levels in the Taiwan Biobank.** Raw p-values are reported for CONCUR, CCRET, CKAT, and a negative control test in which CONCUR is applied to a randomly permuted response vector. Asterisks indicate p-values that were significant after a Holm multiple-testing adjustment. Pathways are ordered alphabetically. ................................. 41

Table 2.4  **Association test results for the effects of CNVs on triglyceride levels by chromosome in the Taiwan Biobank.** Results from the CONCUR, CCRET, and CKAT association tests are shown. The results of the negative control analysis reflect the p-value from CONCUR applied to a randomly permuted response vector. Asterisks indicate p-values that were significant after a Holm multiple-testing adjustment. Pathways are ordered according to chromosome. For interpretation of the by-chromosome association tests, the number of genes from the 15 lipid metabolism pathways that are intersected by CNVs is given (# Genes Interrupted). ................................. 43

Table A.1  **Table A.1 Descriptive statistics for hsa01040 pathway.** TG values are classified as Low (<the 25th percentile [<68 mg/dL]; n=2,931), Medium (the middle 50% [68 - 140 mg/dL]; n=5,844), and High (>the 75th percentile [>140 mg/dL]; n=2,889). The percent of individuals with CNVs is with respect to the total number of individuals in each TG category. The mean number of CNVs per individual and mean total length of CNVs (bp) per individual are reported, as well as the mean lengths (bp) and mean dosage per CNV. “Promising” associations with TG are marked with ** to indicate p-value< 0.01 and with * to indicate p-value< 0.05. ................................................................. 89
Figure 1.1  **CKAT CNVR-level kernel scanning alignment.** The CKAT CNVR kernel evaluates the 3 possible alignments of CNVs between two individuals $i$ and $j$ with $p_i = 4$ and $p_j = 2$. For the $R_i = [X_i^1, X_i^2, X_i^3, X_i^4]$ and $R_j = [X_j^1, X_j^2]$ shown here, the CNVR-level kernel function $k_R(R_i, R_j)$ is the maximum value of the 3 CNV-level kernel sums to the right of the illustrations. . . . . . . . . . . . . . . . . 13

Figure 2.1  **Diagram of copy number profile curves and common area under the curve.** (a) Example of CNV data in standard PLINK format describing profiles of individuals in a small region of chromosome 1. (b) Copy number (CN) profile curves illustrating the cAUC between individuals with overlapping deletions of dosage 0. (c) CN profile curves illustrating the cAUC between individuals with overlapping duplications of dosage 3 and 4. (d) CN profile curves which show 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. . . . . . . . . . . . . . . . . 27

Figure 2.2  **Power comparison between CONCUR, CCRET, and CKAT in the TGP-Chr1(a) simulations (causal dosage×length effects in chromosome 1 of TGP data).** The top panel shows power under combined duplication and deletion effects, the middle panel shows power under effects from duplications only, and the bottom panel shows power under effects from deletions only. Different proportions of deleterious vs. protective effects are considered as indicated by $(D_{Dup}, P_{Dup}, D_{Del}, P_{Del})$ with $D_{Dup}$ and $P_{Dup}$ reflecting the proportions of deleterious and protective segments among causal duplication segments, and with $D_{Del}$ and $P_{Del}$ defined similarly for causal deletion segments. . . . . . 33

Figure 2.3  **Power comparison between CONCUR, CCRET, and CKAT in TGP-Chr1(b) and TGP-Chr1(c) (causal dosage×length effects from both duplications and deletions in chromosome 1 of TGP data).** The top panel shows power under TGP-Chr1(b), in which the kernels are built on CNV data with error added to the CNV endpoints, mimicking the scenario of inaccuracy end points of called CNVs. The bottom panel shows power under TGP-Chr1(c), in which the disease base rate is more rare (5%). Different proportions of deleterious vs. protective effects are considered as indicated by $(D_{Dup}, P_{Dup}, D_{Del}, P_{Del})$ with $D_{Dup}$ and $P_{Dup}$ reflecting the proportions of deleterious and protective segments among causal duplication segments, and with $D_{Del}$ and $P_{Del}$ defined similarly for causal deletion segments. . . . . . 34

Figure 2.4  **Power comparison between CONCUR, CCRET, and CKAT in TWB-Chr1(a) (causal continuous dosage×length effects in chromosome 1 of TWB data).** The top panel shows power under combined duplication and deletion effects, the middle panel shows power under effects from duplications only, and the bottom panel shows power under effects from deletions only. Different proportions of deleterious vs. protective effects are considered as indicated by $(D_{Dup}, P_{Dup}, D_{Del}, P_{Del})$ with $D_{Dup}$ and $P_{Dup}$ reflecting the proportions of deleterious and protective segments among causal duplication segments, and with $D_{Del}$ and $P_{Del}$ defined similarly for causal deletion segments. . . . . . 36
Figure 2.5  
**Power comparison between CONCUR, CCRET, and CKAT for TWB-Chr1 (b)**  
(causal categorical dosage\times length effects in chromosome 1 of TWB data). The panels show power under combined duplication and deletion effects. Different proportions of deleterious vs. protective effects are considered as indicated by (D_{Dup}, P_{Dup}, D_{Del}, P_{Del}) with D_{Dup} and P_{Dup} reflecting the proportions of deleterious and protective segments among causal duplication segments, and with D_{Del} and P_{Del} defined similarly for causal deletion segments.

Figure 3.1  
**Example of CNV segments for 3 individuals.** The horizontal lines represent a section of the genome, with genomic location increasing along x-axis and thick horizontal lines indicating CNVs. Vertical lines mark the CNV segment boundaries for CNV activity among the 3 individuals in this region.

Figure 3.2  
**Performance metrics for model selection using BIC with thresholding across 100 replications for 6 lasso methods.** Metrics are reported at $\gamma = (0.5, 1, 2, 5, 10)$, where $\gamma$ is the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.

Figure 3.3  
**Performance metrics for model selection using AIC with thresholding across 100 replications for 6 lasso methods.** Metrics are reported at $\gamma = (0.5, 1, 2, 5, 10)$, where $\gamma$ is the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.

Figure 3.4  
$D_{IF}(k, 2)$ **model fits at $\lambda_{BIC}$ across 100 replications with $\gamma = 1$.** The mean $\hat{\beta}(\lambda_{BIC})$ is plotted in red on the y-axis and the true $\beta$ is shown in black. The coefficients are ordered on the x-axis according to the genomic locations of the segments the variables correspond to.

Figure 3.5  
**Performance metrics for model selection using BIC across 100 replications for 6 lasso methods.** Metrics are reported at $\gamma = (0.5, 1, 2, 5, 10)$, where $\gamma$ is the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.

Figure 3.6  
**Performance metrics for model selection using AIC with thresholding across 100 replications for 6 lasso methods.** Metrics are reported at $\gamma = (0.5, 1, 2, 5, 10)$, where $\gamma$ is the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.

Figure 3.7  
**Performance metrics for model selection using FSR control across 100 replications for 6 lasso methods, at $\gamma = 1$.** Metrics are reported at FSR control levels of 0.1, 0.15, 0.2, and 0.25. $\gamma$ is the value of the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.

Figure 3.8  
**Model fits at $\lambda_{FSR}$ across 100 replications for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.25.** The mean $\hat{\beta}(\lambda_{FSR})$ is plotted in red on the y-axis and the true $\beta$ is shown in black. The coefficients are ordered according as shown on the x-axis.

Figure 3.9  
**Model fits at $\lambda_{FSR}$ for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.25**  
The estimates $\hat{\beta}(\lambda_{FSR})$ are plotted in red on the y-axis. The coefficients are ordered according as shown on the x-axis.

Figure 3.10  
**Model fits for chr13q22-32 duplications analysis at $\lambda_{FSR}$ for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.25.** The estimates $\hat{\beta}(\lambda_{FSR})$ are plotted in red on the y-axis. The coefficients are ordered according as shown on the x-axis.
Figure A.1 Visualization of CNV activity in pathway hsa01040 by level of triglycerides (TG). CNV activity in genes in hsa01040 is shown by level of TG (Low, Medium, and High), with duplications in red and deletions in dark olive. Columns represent individuals, and genes shown here are the 23 genes in the pathway that contain CNVs, ordered by the number of CNVs contained therein.

Figure B.1 Performance metrics for model selection using FSR control across 100 replications for 6 lasso methods, at $\gamma = 0.5$. Metrics are reported at FSR control levels of 0.1, 0.15, 0.2, and 0.25. $\gamma$ is the value of the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.

Figure B.2 Performance metrics for model selection using FSR control across 100 replications for 6 lasso methods, at $\gamma = 2$. Metrics are reported at FSR control levels of 0.1, 0.15, 0.2, and 0.25. $\gamma$ is the value of the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.

Figure B.3 Performance metrics for model selection using FSR control across 100 replications for 6 lasso methods, at $\gamma = 5$. Metrics are reported at FSR control levels of 0.1, 0.15, 0.2, and 0.25. $\gamma$ is the value of the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.

Figure B.4 Performance metrics for model selection using FSR control across 100 replications for 6 lasso methods, at $\gamma = 10$. Metrics are reported at FSR control levels of 0.1, 0.15, 0.2, and 0.25. $\gamma$ is the value of the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.

Figure B.5 Model fits at $\lambda_{BIC}$ across 100 replications for 6 lasso methods, at $\gamma = 1$. The mean $\hat{\beta}(\lambda_{BIC})$ is plotted in red on the y-axis and the true $\beta$ is shown in black. The coefficients are ordered according as shown on the x-axis.

Figure B.6 Model fits at $\lambda_{AIC}$ across 100 replications for 6 lasso methods, at $\gamma = 1$. The mean $\hat{\beta}(\lambda_{AIC})$ is plotted in red on the y-axis and the true $\beta$ is shown in black. The coefficients are ordered according as shown on the x-axis.

Figure B.7 Model fits for chr13q22-32 deletions analysis at $\lambda_{BIC}$ with thresholding for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.2. The estimates $\hat{\beta}(\lambda_{BIC})$ are plotted in red on the y-axis, and the black line is the threshold. The coefficients are ordered according as shown on the x-axis.

Figure B.8 Model fits for chr13q22-32 deletions analysis at $\lambda_{AIC}$ with thresholding for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.2. The estimates $\hat{\beta}(\lambda_{AIC})$ are plotted in red on the y-axis, and the black line is the threshold. The coefficients are ordered according as shown on the x-axis.

Figure B.9 Model fits for chr13q22-32 deletions analysis at $\lambda_{FSR}$ for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.2. The estimates $\hat{\beta}(\lambda_{FSR})$ are plotted in red on the y-axis. The coefficients are ordered according as shown on the x-axis.

Figure B.10 Model fits for chr13q22-32 duplications analysis at $\lambda_{BIC}$ with thresholding for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.2. The estimates $\hat{\beta}(\lambda_{BIC})$ are plotted in red on the y-axis, and the black line is the threshold. The coefficients are ordered according as shown on the x-axis.

Figure B.11 Model fits for chr13q22-32 duplications analysis at $\lambda_{AIC}$ with thresholding for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.2. The estimates $\hat{\beta}(\lambda_{AIC})$ are plotted in red on the y-axis, and the black line is the threshold. The coefficients are ordered according as shown on the x-axis.
Figure B.12 Model fits for chr13q22-32 duplications analysis at $\lambda_{FSR}$ for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.2. The estimates $\hat{\beta}(\lambda_{FSR})$ are plotted in red on the y-axis. The coefficients are ordered according as shown on the x-axis.
As innovations in high-throughput sequencing technology have accelerated the processing and output of genetic data, methodological development has had to keep pace. New technology has brought increasing data dimensionality, larger samples of individuals with sequenced data, and greater sensitivity in variant detection as well as finer resolution in the variant features. The responding methods grapple with $p \gg n$ problems, multiple testing corrections, rare or low frequency data, and many other issues that make genetic data uniquely difficult and interesting to work with.

In this work, we propose statistical models relating genetic variants to phenotypes using kernel machine regression and lasso methods with the aim of making inference on the nature and source of phenotype-variant associations.

In this study we propose methods for a type of structural genetic variation called a copy number variant (CNV). A copy number variant (CNV) is a variation in the number of times a segment of DNA is duplicated or deleted in comparison to a reference genome and comprise a large proportion of variation in human genomes. CNVs are characterized by the number of copies present, called the dosage, and their length measured in units of base-pairs (bp) and multiples thereof. Increased burden of CNVs has been associated with many phenotypes including bipolar disorder, ADHD, and childhood obesity [Mal11] [Mar15] [Whe13]. CNVs can affect phenotypic variability and severity by influencing gene expression and interrupting regulatory pathways and can have additive or interactive effects with other CNVs. Furthermore, different types of CNVs (duplications versus deletions) commonly have different effects on phenotypes, and may interact with CNV length and dosage effects.

It is only recently that we discovered how widespread CNVs are (on average $>1000$ CNVs in the genome) [Mal12] and attempted a comprehensive map of copy number variation in the human genome.
genome [Red06]. Early CNV analysis relied on easily detectable large CNVs with substantial effects on phenotypes. For CNVs appearing in a large enough proportion of individuals, inference at the locus level is possible using methods such as chi-square and Fisher's exact tests to compare CNV burden in cases versus controls. The enrichment test of Raychaudhuri et al. [Ray10] applies the logistic model, using asymptotic or permutation-based p-values, to common CNV analysis to again compare cases and controls while adjusting for covariates and other CNV features. However, as advancements in sequencing technology have improved CNV detection and produced larger samples of genotyped individuals, analytical methods have evolved to accommodate shorter, rarer CNVs with smaller or more heterogeneous effects.

In this work we propose two methods for CNV analysis on the global and local scale. First we develop a novel kernel-based association test for rare CNV analysis across pathways or genomes. This works also explores a novel function called the “CNV profile curve” which captures CNV dosage and length information as well as location on the genome. The method shows good performance in simulations and in both chromosome and pathway analysis of real CNV data. Second, we propose a lasso-based framework for selection and estimation of CNV effects in smaller regions. The method identifies key causal subregions and estimates the CNV effects arising from them. Moreover, our accommodates rare CNVs through information sharing with neighbors using a generalized lasso.

In Chapter 1 we review the statistical methodology behind our proposed CNV analysis frameworks, including kernel machine regression and lasso methods, as well as the relevant literature. In Chapter 2 we present the CONCUR association test and CNV profile curve concept. In chapter 3 we propose a generalized lasso model for the post-hoc analysis of rare and common CNVs. We offer conclusions and discuss ongoing and future work in Chapter 4.

1.1 Kernel Machine Regression for Genetic Data Analysis

In early generations of rare CNV analysis, burden-style methods such as the enrichment test [Ray10] were proposed to model summary measures such as mean CNV length, count of CNVs, and CNV gene interruption as fixed effects in a linear regression framework. However, this approach was found to be underpowered in settings with between-locus effect heterogeneity, where variants at different loci have different effect sizes or directions. More recently, kernel-based association tests have emerged as a powerful tool in rare CNV analysis [Tze15; Zha16; Bru20] that can account for this between-locus effect heterogeneity as well as nonlinear effects, and interactions across loci using a kernel machine regression architecture.

Kernel-based methods for variant analysis were first applied to single nucleotide polymorphism (SNP) data [Kwe08; Wu10]. Wu et al. pioneered the use of kernel methods for rare variant analysis with the Sequence Kernel Association Test (SKAT) for rare SNPs [Wu11]. SKAT models the aggregated effects of rare SNPs across a target region using a kernel machine regression framework for Gaussian and binary phenotypes. As with the previous kernel-based tests, SKAT leveraged the results from Liu and Li [Liu07] deriving the connections between KMR and linear mixed models. Specifically, the
test for association between the phenotype and the kernel-based nonparametric effect in a KMR can be reframed as a variance components test for a random effect in a generalized linear model, which has a closed form score statistic with a specified exact or asymptotic distribution.

In 2015, Tzeng et al. proposed a kernel-based method for rare CNV analysis [Tze15]. The CNV Collapsing Random Effects Test (CCRET) models CNV length, dosage, and gene interruption features as fixed effects or random effects with covariance structure defined by a kernel function. The CCRET framework also proposes a CNVR-based codification of CNV features from which linear and polynomial kernels may be constructed. More recently, Zhan et al. introduced the CNV Kernel Association Test (CKAT) which captures CNV dosage and length similarity in a single kernel. Different from CCRET, CKAT allows CNVs from different loci to contribute to the measure of genetic similarity and is free from the concept of CNVRs. We discuss these methods in more detail in the sections below.

In the sections below we review least squares kernel machine regression, its connections to linear mixed models, and the resulting hypothesis testing structure for kernel-based association tests.

1.1.1 Basic kernel machine

For individuals $i = 1, \cdots, n$, let $y_i$ be a quantitative trait and $z_i$ be a $r \times 1$ multi-dimensional attribute. Then for an unknown function $h(\cdot): \mathbb{R}^r \rightarrow \mathbb{R}$, we take

$$ y_i = h(z_i) + \epsilon_i, \quad (1.1) $$

for $i = 1, \cdots, n$, where $\epsilon_i$ are independent and follow $\sim N(0,\sigma^2)$. The $h(\cdot)$ function maps the $z_i$ features to a higher dimensional space in which a linear fit to the transformed values versus the response is appropriate. The transformed feature space is defined by $L$ basis functions $\{\phi_1(\cdot), \phi_2(\cdot), \cdots\}$, where the functions $\phi_l(\cdot)$ also map from $\mathbb{R}^r \rightarrow \mathbb{R}$ and $L$ may be infinite. Then taking $\phi(z_i) = [\phi_1(z_i), \phi_2(z_i), \cdots]^T$, we have

$$ h(z_i) = \sum_{l=1}^{L} \phi_l(z_i) \omega_l = \phi(z_i)^T \omega, \quad (1.2) $$

where $\omega_l$ are scalar weights applied to each basis function and $\omega = [\omega_1, \cdots, \omega_L]^T$. This is called the primal representation of the nonparametric function. Then, in matrix form we can express Eq. 1.1 as

$$ y = Z_{\phi} \omega + \epsilon, \quad (1.3) $$

where $y = [y_1, \cdots, y_n]^T$, $Z_{\phi} = [\phi(z_1), \cdots, \phi(z_n)]^T$, and $\epsilon = [\epsilon_1, \cdots, \epsilon_n]^T$. Now, if $L = \infty$ or is larger than $n$, the estimate $\omega_{OLS}$ derived from minimizing the squared error loss $J(\omega) = \sum_{i=1}^{n} (y_i - \phi(z_i)^T \omega)^2$ is not unique and overfits the model. In this case of an under-determined system, the
standard approach is to introduce a ridge regression penalty, which yields the objective function

$$J(\omega) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \phi(z_i)^T \omega)^2 + \frac{\lambda}{2} \|\omega\|_2^2$$

(1.4)

where $\lambda$ is a parameter $\geq 0$ and $\|\omega\|_2^2 = \omega^T \omega$. To find the minimizer of this function, we take its derivative and set it equal to zero. This gives a solution of the form

$$\hat{\omega}_{L \times 1} = (Z_\phi^T Z_\phi + \lambda I_{L \times L})^{-1} Z_\phi^T y.$$  

(1.5)

Now, in the case that $L$ is infinite, it is not possible to calculate this estimate. However, we may alternatively express $\hat{\omega}$ as

$$\hat{\omega} = (Z_\phi^T Z_\phi + \lambda I_{n \times n})^{-1} Z_\phi^T y, \quad \text{and} \quad \hat{\alpha} = [\hat{\alpha}_1, \cdots, \hat{\alpha}_n]^T = (Z_\phi Z_\phi^T + \lambda I_{n \times n})^{-1} y.$$  

(1.6)

where $\hat{\alpha} = [\hat{\alpha}_1, \cdots, \hat{\alpha}_n]^T = (Z_\phi Z_\phi^T + \lambda I_{n \times n})^{-1} y$. This follows from

$$Z_\phi^T (Z_\phi Z_\phi^T + \lambda I_{n \times n})^{-1} y = Z_\phi^T y = Z_\phi^T (Z_\phi Z_\phi^T + \lambda I_{n \times n})^{-1} y$$

(1.7)

and the invertibility of $(Z_\phi^T Z_\phi + \lambda I_{p \times p})$ and $(Z_\phi Z_\phi^T + \lambda I_{n \times n})$ since $\lambda \geq 0$. Continuing, we note that the $(i, j)$th element in $Z_\phi Z_\phi^T$ is the dot product between $\phi(z_i)$ and $\phi(z_j)$. We define a kernel function

$$k(z_i, z_j) = \langle \phi(z_i), \phi(z_j) \rangle,$$  

(1.8)

and define $K = Z_\phi Z_\phi^T$, a kernel matrix with $(i, j)$th element $k(z_i, z_j)$. Note that $K$ is symmetric and positive semi-definite. It is at this stage that we can estimate the non-parametric effect $h(z_i)$ without any terms from potentially infinite spaces $\mathbb{R}^L$ and $\mathbb{R}^{L \times L}$. Plugging our estimates into Eq. 1.3, we now have

$$\hat{h}(z_i) = \phi(z_i)^T \hat{\omega}$$

$$= \phi(z_i)^T Z_\phi \hat{\alpha}$$

$$= \phi(z_i)^T Z_\phi^T (K + \lambda I_{n \times n})^{-1} y$$

$$= \kappa^T (K + \lambda I_{n \times n})^{-1} y.$$

(1.9)
where $\kappa^T = \phi(z_j)^T Z^T_\phi = [k(z_1, z_i), \ldots, k(z_n, z_i)]$. The expression $h(z_i) = \kappa^T \alpha$ for some true $\alpha$ is the dual representation of the $h(\cdot)$ function. The “kernel trick” refers to the reformulation of $Z^T_\phi \hat{\alpha}$ into $\kappa^T \hat{\alpha}$, giving us access to a possibly infinite-dimensional feature space $\phi(\cdot)$ through a kernel function. The only requirements are that the kernel function be non-negative definite and it must collapse a pair of $r \times 1$ attributes to a scalar. In this work, we will refer to methods based on the two types of kernel functions below:

- **Linear kernel**: $k(z_i, z_j) = z_i^T z_j$
- **Gaussian kernel**: $k(z_i, z_j) = \exp\left(-\frac{\|z_i - z_j\|^2}{\rho}\right)$

The kernel function implicitly defines the feature space $\phi$, where an estimate of effect $h(\cdot)$ is derived as a linear combination of the feature space basis functions. Poor choices of kernel functions will result in poor estimates of $h(\cdot)$. We can avoid this problem by using flexible kernels, like the polynomial kernel $k(z_i, z_j) = (z_i^T z_j + 1)^d$ that is a generalized version of the linear kernel, or by designing customized kernels.

### 1.1.2 Kernel machine regression for Gaussian response

We consider the same framework as Eq. 1.1 but now with $p$ covariates $x_i$, for individuals $i = 1, \ldots, n$, with linear effects on $y_i$ through $\beta$. This gives us the semiparametric model

$$y_i = x_i^T \beta + h(z_i) + \epsilon_i. \quad (1.10)$$

In this setting, we assume that $h$ lies in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with inner product $(\cdot, \cdot)_{\mathcal{H}}$. The RKHS is uniquely determined by a non-negative kernel function $k(\cdot, \cdot)$ and ensures that the primal and dual representations of the functions exist. Then we seek to estimate $\beta$ and the function $h$ by minimizing the objective function

$$J(\beta, h) = \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i^T \beta - h(z_i))^2 + \frac{\lambda}{2} \|h\|^2_{\mathcal{H}}, \quad (1.11)$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm induced by the inner product on the RKHS, and $\lambda$ is a tuning parameter that controls the complexity of the fit. The representer theorem gives us the form of the minimizer to Eq. 1.11 as

$$h(\cdot) = \sum_{j=1}^{n} \alpha_j k(\cdot, z_j), \quad (1.12)$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)^T$ are unknown parameters [Kim71]. Substituting this into Eq. 1.11 becomes

$$J(\beta, \alpha) = \frac{1}{2} \sum_{i=1}^{n} \left( y_i - x_i^T \beta - \sum_{j=1}^{n} \alpha_j k(z_i, z_j) \right)^2 + \frac{\lambda}{2} \|\alpha\|^2_{\mathcal{H}}$$

$$= \frac{1}{2} (y - X\beta - K\alpha)^T (y - X\beta - K\alpha) + \frac{\lambda}{2} \alpha^T K\alpha, \quad (1.13)$$
where $X = [x_1, \cdots, x_n]^T$ and $K = \{k(z_i, z_j)\}_{n \times n}$ is the kernel matrix built from the kernel function.

To minimize $J(\beta, \alpha)$, we take its derivative with respect to $\beta$ and $\alpha$ to obtain the first order equations

\[
\begin{bmatrix}
X^T X & X^T K \\
K^T X & K^T K + \lambda K
\end{bmatrix}
\begin{bmatrix}
\beta \\
\alpha
\end{bmatrix}
= 
\begin{bmatrix}
X^T y \\
K^T y
\end{bmatrix}.
\]

(1.14)

The solutions to the first order equations are

\[
\hat{\beta} = (X^T (K + \lambda I)^{-1} X)^{-1} X^T (K + \lambda I)^{-1} y
\]

(1.15)

\[
\hat{\alpha} = (K + \lambda I)^{-1} (y - X \hat{\beta})
\]

(1.16)

Plugging this back into Eq. 1.12 gives us the estimate $\hat{h}(z_i) = \kappa^T \hat{\alpha}$ for $i = 1, \cdots, n$ and where $\kappa = \{k(z_i, z_1), \cdots, k(z_i, z_n)\}^T$ as above. Across all individuals, we have $\hat{h} = K \alpha$ where $\hat{h} = [\hat{h}(z_1), \cdots, \hat{h}(z_n)]^T$.

With the estimates derived above, we can obtain predicted values of $y$ as well as estimates of the linear effects $\hat{\beta}$. In practice, researchers may fit the model at a sequence of $\lambda$’s and choose the value that minimizes an information criterion or cross-validation loss functions.

### 1.1.3 Connection to Linear Mixed Models

The kernel-based association methods discussed later in this section test the hypothesis $H_0 : h(\cdot) = 0$; that is, whether the collapsed measure of variant activity is associated with the phenotype. Liu, Lin, and Ghosh [Liu07] demonstrated the connection between KMR and linear mixed models which powerfully allows us to test $H_0$ using a score-based variance components test. Liu et al. show that the estimates of $\beta$ and $\alpha$ obtained in Eq. 1.16 can be obtained from the first order equations for a linear model with random effects $h$. The linear mixed model takes the form

\[
y = X \beta + h + \epsilon,
\]

(1.17)

for $y$, $X$, $\beta$, and $\epsilon$ defined as above, and $h \sim N(0, \tau D)$ and $h \perp \epsilon$, where $\tau$ is a variance component $\geq 0$ and $D$ is a $n \times n$ covariance matrix for the random effects. The linear mixed model has first order equations

\[
\begin{bmatrix}
X^T R^{-1} X & X^T R^{-1} \\
R^{-1} X & R^{-1} + (\tau D)^{-1}
\end{bmatrix}
\begin{bmatrix}
\beta \\
h
\end{bmatrix}
= 
\begin{bmatrix}
X^T R^{-1} y \\
R^{-1} y
\end{bmatrix}
\]

(1.18)

where $R = \sigma^2 I$. If we take $\tau = \lambda^{-1} \sigma^2$ and $D = K$, these equations produce estimates $\hat{\beta}$ and $\hat{h}$ that are equivalent to those under the KMR model.

Using the linear mixed models framework, we perform an equivalent test to $H_0 : h(\cdot) = 0$ using $H_0 : \tau = 0$. Liu et al. propose the score statistic

\[
S = \frac{1}{2\sigma^2} (y - X \hat{\beta})^T K (y - X \hat{\beta}) - \text{tr} \{P_0 K\}
\]

(1.19)
where $\hat{\beta}_0$ and $\hat{\sigma}_0^2$ are the MLEs under the null model $y = X\beta_0 + \epsilon_0$, $\epsilon_0 \sim N(0, \sigma_0^2)$, and $P_0 = I - X(X^TX)^{-1}X$. To obtain a p-value we calculate $\Pr(S > s)$, where $s$ is the observed score statistic and $\Pr(\cdot)$ is with respect to the distribution of the score statistic. To calculate this p-value we drop the second term which is constant and note that the first term, call it $Q$, asymptotically follows a weighted chi-square distribution $\sum_{i=1}^n \lambda_i Z_i^2$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $(P_0 K P_0)/2$ and $Z_1, \ldots, Z_n$ are i.i.d. variables $\sim N(0, 1)$.

When $n$ is small this test is often overly conservative, which motivated Chen et al. [Che16] to derive an exact test. They define the alternate score statistic

$$R = \frac{e^TP_0K P_0 e}{e^TP_0 e} = \frac{(y - X\hat{\beta}_0)^T K(y - X\hat{\beta}_0)}{(y - X\hat{\beta}_0)^T P_0(y - X\hat{\beta}_0)}$$

(1.20)

which is proportional to $Q$. Then, they calculate the observed statistic

$$r = \frac{(y - X\hat{\beta}_0)^T K(y - X\hat{\beta}_0)}{(y - X\hat{\beta}_0)^T (y - X\hat{\beta}_0)}.$$

(1.21)

This yields

$$\Pr(Q \leq q) = \Pr(R \leq r)$$

$$= \Pr\left(\frac{e^TP_0K P_0 e}{e^TP_0 e} \leq r\right)$$

$$= \Pr\left(\frac{e^T(P_0 K P_0 - r P_0) e}{\sigma^2} \leq 0\right)$$

(1.22)

which can be calculated using an exact weighted chi-square distribution with weights being the eigenvalues of $(P_0^{1/2} K P_0^{1/2} - r I)$. Davies’ method [Dav80] provides an algorithm for calculating p-values for weighted chi-square distributions which we use in our work.

### 1.1.4 Kernel machine regression for exponential family responses

We can also use KMR for binary and other non-Gaussian responses that follow a distribution in the exponential family. Recall that for $y_i$ with an exponential family distribution, the probability density function is

$$p(y_i|\eta_i, \phi) = h(y_i, \phi) \exp\left\{ \frac{1}{\phi} [\eta_i y_i - a(\eta_i)]\right\}$$

(1.23)

where $\eta_i$ is the natural parameter, $\phi$ is a dispersion parameter, and $a(\cdot)$ and $h(\cdot)$ are known functions. These terms relate to the mean and variance of $y_i$ through $\mu_i = E(y_i|\eta_i) = a'(\eta_i)$ and $\text{var}(y_i|\eta_i, \phi) = \phi a''(\eta_i) = \phi v(\mu_i)$. In a generalized KMR (GKMR) model we connect the responses to the covariates and attributes through

$$g(\mu_i) = x_i^T \beta + h(z_i)$$

(1.24)
for \( i = 1, \cdots, n \), where \( x_i, \beta, h(\cdot) \), and \( z_i \) are as defined above. We let \( g \) be the canonical link, which means that \( g^{-1}(\cdot) = a'(\cdot) \) and thus \( \eta_i = x_i^T \beta + h(z_i) \).

For simplicity, we assume \( \phi \) is known (e.g., \( \phi = 1 \) canonically for binomial and Poisson distributions) or has an estimate \( \hat{\phi} \). Then for GKMR, the objective function for \((\beta, h)\) derived from the negative log-likelihood is

\[
J(\beta, h) = -\sum_{i=1}^{n} \frac{[\eta_i y_i - a(\eta_i)]}{\phi} + \frac{\lambda}{2} \| h \|^2_{\mathcal{H}}.
\]

By the representer theorem [Kim71], the minimizer to this function with respect to \( h \) takes the form \( h(\cdot) = \sum_{j=1}^{n} a_j k(\cdot, z_j) \) so that we have

\[
J(\beta, \alpha) = -\sum_{i=1}^{n} \frac{[\eta_i y_i - a(\eta_i)]}{\phi} + \frac{\lambda}{2} \alpha^T K \alpha,
\]

where \( \eta_i = x_i^T \beta + \kappa_i^T \alpha, \alpha = (\alpha_1, \cdots, \alpha_n)^T \), and \( \kappa = [k(z_1, z_i), \cdots, k(z_n, z_i)]^T \).

The \((\alpha, \beta)\) minimizers of Eq. 1.9 do not have a closed form in general. In practice, we solve for \( \alpha \) and \( \beta \) using iteratively reweighted least squares (IRLS). Given the \( k \)th iteration results, the \((k+1)\)th estimates of \( \beta \) and \( \alpha \) solves the system of equations

\[
\begin{bmatrix}
X^T W^{(k)} X & X^T W^{(k)} K \\
W^{(k)} X & \lambda I + W^{(k)} K
\end{bmatrix}
\begin{bmatrix}
\beta^{(k+1)} \\
\alpha^{(k+1)}
\end{bmatrix}
= \begin{bmatrix}
X^T W^{(k)} \tilde{y}^{(k)} \\
W^{(k)} \tilde{y}^{(k)}
\end{bmatrix},
\]

where \( W^{(k)} \) is a \( n \times n \) diagonal matrix with diagonal elements \( w_{ii} = \left[ g'(\mu_i^{(k)}) \phi \right]^{-1} \) for \( i = 1, \cdots, n \), and \( \tilde{y}^{(k)} \) is the working response such that

\[
\tilde{y}_i^{(k)} = x_i^T \beta^{(k)} + \kappa_i^T \alpha^{(k)} + g'\left(\mu_i^{(k)}\right)(y_i - \mu_i^{(k)}),
\]

with \( \mu_i^{(k)} = g^{-1}\left(x_i^T \beta^{(k)} + \kappa_i^T \alpha^{(k)}\right) \).

As with KMR and linear mixed models, there is a connection between GKMR and generalized linear mixed models [Liu08a]. Specifically, we can test for \( H_0: h(\cdot) = 0 \) using a generalized linear mixed model (GLMM)

\[
g(\mu_i) = x_i^T \beta + h_i
\]

where \( h \sim N(0, \tau K) \), where \( K \) is the kernel matrix of the corresponding GKMR problem and \( \tau^2 = \lambda^{-1} \phi \). To test \( H_0: \tau = 0 \), researchers have proposed several variance components tests based on results from Zhang and Lin [Zha03]. Here we present the variance components score test of Tzeng and Zhang [Tze07], which for binomial responses with no overdispersion agrees with the tests in [Liu08a] and [Wu10]. First we consider the null model \( g(\mu_{0i}) = x_i^T \beta_0 = \eta_{0i} \) for \( i = 1, \cdots, n \) and let \( \phi_0 \) be the dispersion under \( H_0 \). We further define the \( n \times n \) diagonal matrix \( \Delta_0 \) with \( \delta_{ii} = g'(\mu_{0i}), n \times n \) diagonal matrix \( W' \) with \( w_{ii} = \left[ \phi_0 a''(\eta_{0i})|g'(\mu_{0i})|^2 \right]^{-1} \). A score statistic for \( H_0: \tau = 0 \) resulting from this GLMM is

\[
Q = \frac{(y - \mu_0)^T \Delta W K W \Delta(y - \mu_0)}{2} \bigg|_{\mu_0 = \mu_0, \phi_0 = \phi_0}.
\]
This statistic asymptotically follows a weighted chi-square distribution \( \sum_{i=1}^{n} \lambda_i Z_i^2 \) where i.i.d. \( Z_i \sim N(0, 1) \) and \( \lambda_1, \cdots, \lambda_n \) are the eigenvalues of \( \left(W^{1/2} P_0 K P_0 W^{1/2}\right)/2 \).

In this work, we employ a test developed in [Che16] for binomial responses that specifies an asymptotic distribution of the score statistic that accounts for potential overdispersion \( (\phi \neq 1) \). The test statistic proposed by Chen et al. is again derived from a model fit using iteratively reweighted least squares (IRLS). We apply IRLS to estimate \( \hat{\beta}_0 \) and \( \hat{\phi}_0 \) in the null model \( (h \equiv 0) \). In a given iteration \( k \), we let \( \tilde{y}^{(k)} \) be the working response and \( W^{(k)} \) be the working weight matrix with diagonal elements \( w_{ii} = \mu_{0i}^{(k)}(1-\mu_{0i}^{(k)}) \), with \( \mu_{0i}^{(k)} = \expit(x_i^T \hat{\beta}_0^{(k)}) \). Then IRLS iteratively solves for \( \hat{\beta}_0^{(k+1)} \) and \( \hat{\phi}_0^{(k+1)} \) under the model

\[
\tilde{y}^{(k)} = X \hat{\beta}_0^{(k+1)} + \epsilon^{(k)}
\]

where \( \epsilon^{(k)} \sim N \left( 0, \phi_0^{(k+1)} \left( W^{(k)} \right)^{-1} \right) \). Then, given converged terms \( \tilde{y}, \hat{\beta}_0, \tilde{W}, \) and \( \tilde{\phi}_0 \), the score statistic is

\[
\tilde{Q} = \frac{(\tilde{y} - X \hat{\beta}_0)^T D K D (\tilde{y} - X \hat{\beta}_0)}{\tilde{\phi}_0},
\]

where \( D = \tilde{W} \). Chen defines \( \tilde{y}^* = \tilde{W}^{1/2} \tilde{y}, X^* = \tilde{W}^{1/2} X, \) and \( \epsilon^* = \tilde{W}^{1/2} \epsilon_0, \) where \( \epsilon_0 = y - X \beta_0, \) which gives us

\[
\tilde{Q} = \frac{(\tilde{y}^* - X^* \hat{\beta}_0)^T D^{1/2} K D^{1/2} (\tilde{y}^* - X^* \hat{\beta}_0)}{\tilde{\phi}_0}.
\]

Taking \( P_0 = I - X^*(X^*T X^*)^{-1} X^*T, \) we have \( (\tilde{y}^* - X^* \hat{\beta}_0) = P_0 \epsilon^* \). We note that \( \epsilon^* \) is approximately \( \sim N(0, \phi I) \). Finally, we define the theoretical term

\[
R = \frac{\epsilon^T P_0 D^{1/2} K D^{1/2} P_0 \epsilon^*}{\epsilon^T P_0 \epsilon^*},
\]

which is proportional to Eq. 1.32, and the observed statistic

\[
r = \frac{(\tilde{y} - X \hat{\beta}_0)^T D K D (\tilde{y} - X \hat{\beta}_0)}{(\tilde{y} - X \hat{\beta}_0)^T W (\tilde{y} - X \hat{\beta}_0)}. \tag{1.34}
\]

Then we can obtain a p-value through

\[
\Pr(R \leq r) = \Pr \left( \frac{\epsilon^T (P_0 D^{1/2} K D^{1/2} P_0 - r P_0) \epsilon^*}{\phi} \leq 0 \right),
\]

where the distribution of the LHS term is approximately a weighted chi-square with weights being the eigenvalues of \( (P_0 D^{1/2} K D^{1/2} P_0 - r I) \). Davies’ method again may be used to calculate the p-value here.

### 1.1.5 Kernel-based association tests

The kernel machine regression and hypothesis testing methods above can be used in a variety of settings, but we turn our focus to their application in rare variant analysis. Here we overview 3
association tests for rare SNP analysis (SKAT) and rare CNV analysis (CCRET and CKAT).

### 1.1.5.1 SKAT

Wu et al. pioneered the use of KMR for rare variant analysis with the Sequence Kernel Association Test (SKAT) [Wu11]. Here we overview the SKAT framework. For \( i = 1, \cdots, n \) individuals we have continuous or binary traits \( y_i \), \( p \) covariates \( x_i = [x_{i1}, \cdots, x_{im}] \), and genotypes \( g_i = [g_{i1}, \cdots, g_{ir}] \) observed at \( r \) variant sites. SKAT assumes the KMR model

\[
y_i = \beta_0 + x_i^T \beta + h(g_i) + \epsilon_i, \quad i = 1, \cdots, n
\]

for quantitative traits and

\[
\logit(\Pr(y_i = 1)) = \beta_0 + x_i^T \beta + h(g_i), \quad i = 1, \cdots, n
\]

for binary traits, where \( h(g_i) \) is a non-parametric function and \( \epsilon_i \) is a random variable with mean 0 and variance \( \sigma^2 \). For an individual's genotypes \( g_i \) across \( r \) SNPs, \( h(g_i) \) represents the net effect resulting from collapsing all individual SNP effects. Since \( h(\cdot) \) is defined so flexibly, it can potentially capture myriad signals ranging from numerous small effects across many sites, heterogeneous effects across sites among rare and common SNPs, to large effects from very rare CNVs. Wu et al. explored 3 kernel choices which define a functional basis for \( h(\cdot) \): a weighted linear kernel of the form

\[
k(g_i, g_j) = \sum_{k=1}^r w_k g_{ik} g_{jk},
\]

which assumes that the trait depends on the variants through a linear random effect; a weighted quadratic kernel

\[
k(g_i, g_j) = (1 + \sum_{k=1}^r w_k g_{ik} g_{jk})^2;
\]

and a weighted identity by state (IBS) kernel

\[
k(g_i, g_j) = \sum_{k=1}^r w_k IBS(g_{ik}, g_{jk}),
\]

where the IBS between two individuals' SNPs is the number of alleles that share IBS. The effect vector \( h = [h(g_1), \cdots, h(g_n)]^T \) is taken to follow a distribution with mean 0 and covariance \( \tau K \), where \( K = [k(g_i, g_j)] \) for the chosen kernel function. Then the proposed variance component score statistic for testing \( H_0 : \tau = 0 \) is

\[
Q = (y - \bar{\mu}_0)^T K (y - \bar{\mu}_0)
\]

where \( \bar{\mu}_0 \) is the predicted mean under \( H_0 \).

In simulations and real data analysis, SKAT showed substantial power gains over existing burden-style methods when significant heterogeneity was present, and thus established kernel methods as a promising tool for the analysis of rare variants. This and the smoothing and dimension reduction qualities of KMR further cemented its usefulness in the analysis of genetic data.

### 1.1.5.2 CCRET

Following SKAT, Tzeng et al. proposed the CNV Collapsing Random Effects Test (CCRET) for rare CNV analysis. CCRET examines associations between CNV dosage, length, and gene interruption by modeling one feature as a random effect and treating the others as fixed effects. Loci are defined as
CNV regions (CNVR) which are constructed by clustering CNV segments with at least 1 basepair (bp) overlap. For each individual within each CNVR, the longest CNV in the CNVR is saved and its length and dosage is stored in two respective \( n \times r \) design matrices. To examine gene interruption, \( G \) genes of interest are selected a priori, and the gene interruption in each individual’s CNV activity is noted. This yields three design matrices: \( n \times r \) matrix \( Z^{DS} \) capturing CNV dosage (1, 2, and 3) within each CNVR; \( n \times r \) matrix \( Z^{Len} \) recording CNV length within each CNVR; and \( n \times G \) matrix \( Z^{GI} \) indicating gene interruption by CNVs.

CCRET relates phenotypes to covariates and CNV data in a GLMM framework for traits distributed in an exponential family. We let \( g(\cdot) \) and \( \mu_i \) be defined as above for a trait \( y_i \) and note \( a(\cdot) \) and \( \phi \) for later use in the score test. Then the CCRET model is

\[
g(\mu_i) = x_i^T \beta + h_{DS}(z_{i}^{DS}) + h_{Len}(z_{i}^{Len}) + h_{GI}(z_{i}^{GI})
\]

for \( i = 1, \ldots, n \), where \( x_i \) is individual \( i \)'s \( p \times 1 \) vector of covariates with effects \( \beta \), \( z_i^{DS} \) is the \( r \times 1 \) vector of CNV dosage data for individual \( i \) and likewise for length and gene interruption, and the function \( h_{DS}(\cdot) \), \( h_{Len}(\cdot) \), and \( h_{GI}(\cdot) \) are smooth, possibly non-parametric functions. In practice, CCRET users ust select a CNV feature of interest to be modeled flexibly and the background CNV features are assumed to have fixed effects, i.e., \( h_f(z_f^i) = z_f^iT \beta_f \) for background feature \( f \), where \( \beta_f \) is fixed. Now letting \( f \) denote the feature of interest, \( h_f(\cdot) \) is taken to be a random effect \( h_f = [h_f(z_f^1), \ldots, h_f(z_f^n)] \sim N(0, \tau_f K_f) \) where \( K_f \) is a \( n \times n \) matrix.

The CCRET method constructs the covariance matrix \( K_f \) through the distance metric \( k(z_f^i, z_f^j) \) which quantifies similarity in a pair of individuals based on similarity in CNV feature \( f \). CCRET adopts some notation from similarity regression [Tze09] and variance components (VC) methods [Tze07], both of which are closely connected to KMR. For the choice of \( k(\cdot, \cdot) \), the CCRET authors propose a \( d \)th order polynomial function of the form

\[
k_f(z_f^i, z_f^j) = \left(1 + \sum_{k=1}^{r_f} w_k z_{ik}^f z_{jk}^f\right)^d
\]

where \( r_f = r \) for the dosage and length features and \( r_f = G \) for gene interruption, and \( w_k \) is a pre-specified weight for locus \( k \). Given \( K_f \) for the feature of interest, the CCRET method tests the hypothesis \( H_0 : \tau_f = 0 \) using the score statistic given in Eq. 1.29. Finally, p-values are calculated using Davies’ method [Dav80] or the moment-matching approach of [Duc10].

1.1.5.3 CKAT

Zhan et al. propose an alternative association test for rare CNVs called the CNV Kernel Association Test (CKAT) [Zha16]. In contrast to CCRET, CKAT is intended to be applied to one or multiple small genomic regions (CNVR) in which individuals may have multiple CNVs. The CKAT-defined CNVR is a region in which any two CNVs between two individuals have the potential to have non-zero similarity via the kernel function, regardless of their proximity to each other within the CNVR. In a
KMR framework, CKAT summarizes genetic similarity between individuals using kernels at the CNV and CNVR level. We denote the length and dosage of a CNV by \( X = (X^{(1)}, X^{(2)}) \). The CNV-level kernel function for two CNVs \( X_1 \) and \( X_2 \) is defined as

\[
k(X_1, X_2) = \exp \left\{ -\frac{\left( X^{(1)}_1 - X^{(1)}_2 \right)^2}{\rho} \times \left[ I(X^{(2)}_1 = X^{(2)}_2) + 1 \right] \right\},
\]

(1.41)

where \( I(X^{(2)}_1 = X^{(2)}_2) \) takes value 1 if the two CNVs have the same dosage and 0 otherwise. All CNVRs are standardized to have length 1 and the choice of \( \rho \) controls how similarity in length contributes to the overall similarity between two CNVs.

In a fixed CNVR, we let \( R_i = (X^{(i)}_1, \cdots, X^{(i)}_{p_i}) \) represent the \( p_i \) CNVs in individual \( i \)'s genome in that CNVR sorted according to their positions. We take \( R_j \) to be an analogous array for individual \( j \). Then the CNVR-level kernel function for these two individuals in a fixed CNVR is defined as

\[
k_R(R_i, R_j) = \begin{cases} 
0 & \text{if } p_i p_j = 0 \\
\max_{l=0,1,\cdots,p_i-p_j} \sum_{t=1}^{p_i} k(X^{i}_{t+l}, X^{j}_{t}) & \text{if } p_i \geq p_j > 0, \\
\max_{l=0,1,\cdots,p_j-p_i} \sum_{t=1}^{p_j} k(X^{i}_{t+l}, X^{j}_{t}) & \text{if } p_j > p_i > 0.
\end{cases}
\]

(1.42)

This kernel function performs a scanning “alignment” of the two sets of CNVs based on their ordinal position within the CNVR. The function returns the maximum sum of the CNV-level kernel function values evaluated at all possible alignments. We illustrate this process in Fig. 1.1.

The CKAT method is designed to associate CNV behavior with normal or binomial traits in a generalized KMR framework. In a fixed CNVR and for individuals \( i = 1, \cdots, n \), we let \( y_i \) be a binary phenotype, \( x_i \) be a vector of \( p \) covariates with fixed effects \( \beta \), and \( R_i = [X^{(i)}_1, \cdots, X^{(i)}_{p_i}] \) capture the dosage and length data of all CNVs in the CNVR for individual \( i \). With \( g(\mu_i) = \mu_i \) for normal traits and \( g(\mu_i) = \logit(\mu_i) \) for binary traits, CKAT assumes the model

\[
g(\mu_i) = \beta_0 + x_i^T \beta + f(R_i),
\]

(1.43)

where \( \beta_0 \) is an intercept term and \( f(\cdot) \) is a function in the space spanned by the CNVR kernel \( k_R(\cdot, \cdot) \). CKAT performs the hypothesis test for \( f(\cdot) = 0 \) using Chen's small sample tests [Che16] presented above.

Here we detailed the respective kernel design and collapsing approaches of CCRET and CKAT since these are the key features that distinguish the two methods. In Chapter 2, we further contrast these methods, discuss their limitations, and propose a new kernel-based association test for rare CNV analysis.
Figure 1.1 CKAT CNVR-level kernel scanning alignment. The CKAT CNVR kernel evaluates the 3 possible alignments of CNVs between two individuals $i$ and $j$ with $p_i = 4$ and $p_j = 2$. For the $R_i = \{X_i^1, X_i^2, X_i^3, X_i^4\}$ and $R_j = \{X_j^1, X_j^2\}$ shown here, the CNVR-level kernel function $k(R_i, R_j)$ is the maximum value of the 3 CNV-level kernel sums to the right of the illustrations.

1.2 Lasso Methods for Genetic Data Analysis

The second method proposed in this work is a generalized lasso model for rare and common CNV analysis. The lasso is well suited for high-dimensional data and has been successfully applied in many genomic data analysis settings including genome wide association studies [Wu09; Hof13], integrative analysis [Man11; Pin15], microbiome compositional data [Fan15], and CNV detection [Tib08; Now11; Zha12]. While the standard lasso is ideal for detecting sparse, diffuse signals from well-populated predictor data, it is low-powered to estimate rare variant effects and will poorly capture smooth signals arising from clustered or correlated predictors. In these settings, other lasso methods such as the network-constrained lasso, sparse group lasso, and fused lasso can reflect the structure of the predictors and borrow information across them. Below we discuss applications of these methods that leverage regulatory pathway information in gene expression analysis, gene membership in rare SNP variable selection, and probe location in CNV calling.

In 2007, Li and Li introduced a network-constrained regularization model which incorporates graph structure into the lasso procedure [Li08]. In their study the graph nodes were genes and the edges (possibly weighted) were based on regulatory relationships between genes. The Laplacian $L$ of the graph modified the gene effects $\beta$ in the penalized term $\beta^T L \beta$ to produce smoothed coefficient profiles and encourage similar effects in connected genes. Given edge weight $w(u, v)$ for connected coefficients $\beta_u$ and $\beta_v$, the Laplacian matrix translates to the element-wise term...
\[ w(u, v) \left( \frac{\beta_u}{k_u} - \frac{\beta_v}{k_v} \right)^2 \]

where \( k_u \) and \( k_v \) are scaling factors that may equal 1; thus, the weighted squared difference in connected coefficients is shrunk towards zero in the network-constrained lasso. We note that this method was not explicitly designed for rare variants but to encourage more biologically interpretable results; however, the \( \ell_2 \) norm applied to pairs of coefficients does induce some degree of information-sharing.

Zhou et al. proposed a sparse group lasso for rare SNP analysis [Zho10]. Their method boosts power in pathway analysis by allowing rare SNPs to borrow information with rare and common SNPs falling in the same gene. This is achieved with a group lasso penalty, as introduced in [Yua06], which encourages all or no predictors in a group to enter the model. With the addition of a standard lasso penalty, the sparse group lasso of Zhou et al. shrinks the effects of neutral SNPs within causal genes to zero.

In the domain of CNV analysis, researchers have applied the fused lasso to estimate CNVs from array comparative genomic hybridization (aCGH) data. The fused lasso is suitable for ordered predictors when the expected or desired solution is piecewise constant. Neighboring predictors are encouraged to have identical effects through a penalty on the fusion term \( |\beta_{j+1} - \beta_j| \) for \( p \) coefficients and \( j = 1, \cdots, p - 1 \). Tibshirani et al. [Tib08] applied the fused lasso to the \( \log_2 \) relative intensity measurements from single-sample aCGH data indicating the number of DNA copies present across genomic loci. Their approach is a specific flavor of the fused lasso called the fused lasso signal approximator (FLSA) [Tib05; Fri07] suitable for one-dimensional signals, with the feature matrix \( X \) being the identity matrix. This induced a locally constant or “fused” estimate of the number of DNA copies present at the probe sites, which could be further processed to obtain integer calls.

Nowak et al. [Now11] extended this method to multi-sample aCGH data to identify shared CNV regions that may not be detectable in individual analysis. Zhang et al. [Zha12] further applied a Euclidean penalty to pairs of across-sample same-locus coefficients to encourage CNV alignment. More recently, Cheng et al. proposed a doubly penalized model to simultaneously align intensity change points across samples and estimate the intensities’ linear association with a phenotype [Che18]. This method makes use of the mapping from a lasso problem in \( \beta \) with a cleverly constructed design matrix \( X \) to a fused lasso problem in the mean signal \( \mu \) as shown in [Har07].

The group lasso, the network-constrained lasso, and the fused lasso approaches are all candidates for rare variant analysis. The methods have power to estimate rare variant effects by borrowing information from connected rare or common variants with differences in the nature of the estimated grouped effects, i.e., all-or-none, smoothed, or fused solutions. In our work here, we propose a novel method for the downstream pathway and chromosome analysis of rare CNVs using a generalization of the fused lasso called the generalized lasso. We pursue this approach due to its piecewise constant solution, which is well studied in upstream aCGH data processing and biologically plausible for signals arising from copy number as discussed above. Like the network-constrained lasso of Li and Li [Li08], the generalized lasso can incorporate a graph structure connecting the predictors, which we design to differentially weight pairs of coefficients in our proposed method. The generalized lasso penalizes differences in pairs of coefficients in a \( \ell_1 \) penalty not \( \ell_2 \), resulting in fusion as opposed to
more general smoothing.

In Chapter 3 we discuss the proposed generalized lasso method further and show its usefulness as a selection and estimation tool. Here we overview several flavors of lasso including the standard lasso, fused, generalized, and constrained lasso models and discuss the topics of unpenalized predictors, model selection, and degrees of freedom.

1.2.1 Standard lasso for linear models

We first review the standard lasso introduced by Tibshirani in [Tib96]. Let \( y \) be a \( n \times 1 \) vector of centered normal responses, \( X \) be a \( n \times p \) predictor matrix, and \( \beta = (\beta_1, \cdots, \beta_p)^T \) be a vector of coefficients. Define \( \|z\|_2 = \sqrt{\sum_{j=1}^p z_j^2} \) and \( \|z\|_1 = \sum_{j=1}^p |z_j| \) for vector \( z \in \mathbb{R}^p \). Then we obtain the lasso estimate \( \hat{\beta} \) by solving

\[
\min_{\beta} \left\{ \frac{1}{2} \|y - X\beta\|^2_2 + \lambda \|\beta\|_1 \right\},
\]

where \( \lambda \) is a scalar \( \geq 0 \). We take \( X \) to have centered and scaled columns, to protect against disproportionate penalization of predictors with small coefficients arising from the scale on which they are reported or measured.

The penalty parameter \( \lambda \) balances shrinkage versus saturation, such that estimates approach the null solution \( (\beta \equiv 0) \) as \( \lambda \to \infty \) and the least squares solution as \( \lambda \to 0 \). For Gaussian responses, the LARS algorithm [Efr04] provides an exact solution to the lasso problem as a function of \( \lambda \), which exists because the path is piecewise linear in \( \lambda \). For large \( n \) or non-Gaussian responses, pathwise coordinate descent [Fri10] can supply approximate solutions at a user-specified sequence of \( \lambda \) values. A computationally efficient implementation of the algorithm is available in the \texttt{glmnet} R package. We review the pathwise coordinate descent algorithm in Section 1.2.5.2 below.

1.2.2 Fused lasso and generalized lasso for linear models

The fused lasso assumes that the order of the predictors is meaningful and penalizes differences in neighboring coefficients to produce locally smooth "fused" estimates [Tib05]. Here we take the fused lasso to include both a fusion term and lasso term, which in some contexts is referred to as the "sparse fused lasso" [Tib11]. The fused lasso objective function is

\[
\min_{\beta} \left\{ \frac{1}{2} \|y - X\beta\|^2_2 + \lambda_1 \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j| + \lambda_2 \|\beta\|_1 \right\}
\]

where \( \beta_j \) and \( \beta_{j+1} \) are consecutive elements in the \( \beta \) vector of coefficients, and \( \lambda_1 \) and \( \lambda_2 \) are scalar penalty parameters \( \geq 0 \). As above we omit an intercept for simplicity, but it can be added along with unpenalized covariates as shown in Section 1.2.5.1. Also note that in contrast to the standard lasso, predictors are typically not scaled in the fused lasso unless smoothing on the normalized scale is desired. In this representation we assign different penalties to the fusion and lasso terms, which allows for the two forces to be differentially exerted. In some implementations, the two penalties
are related using a mixing or scaling parameter to reduce the λ space over which the lasso is solved, e.g., replace .

The fused lasso is a specific case of a broader formulation called the generalized lasso [Tib11]. In the generalized lasso framework, coefficients are penalized according to a contrast matrix $D$ in an $\ell_1$ penalty term as follows:

$$\min_\beta \left\{ \frac{1}{2} \| y - X \beta \|_2^2 + \lambda \| D \beta \|_1 \right\}$$

(1.46)

where $D$ is $m \times p$ with $m \leq p$. In the case of the fused lasso presented above,

$$D = \left( \begin{array}{c} D_0 \\ \frac{\lambda}{2} I_{p \times p} \end{array} \right)$$

(1.47)

where

$$D_0 = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\
0 & -1 & 1 & \cdots & 0 & 0 \\
& & \cdots \\
0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

(1.48)

is $m_0 \times p$ and imposes the fusion penalty, and $I_{p \times p}$ imposes the lasso penalty on the coefficients. The generalized lasso encompasses a variety of problems such as signal approximation, 1d and 2d fused lasso, and linear and polynomial trend filtering. Like the network-constrained lasso [Li08], the generalized lasso can incorporate a graph structure connecting the predictors. In this case, each row of $D_0$ is a contrast with two nonzero elements indicating nodes connected by an edge. For example, give a pair of coefficients $\beta_u$ and $\beta_v$ with edge weight $w(u, v)$, we can represent the connection as a row in $D_0$ with $u$th element $-w(u, v)$ and $v$th element $w(u, v)$.

As with the fused lasso, the effect of the $\ell_1$ penalty applied to pairs of coefficients is to not just smooth them towards each other but fuse them together. Tibshirani et al. [Tib11] derived an exact solution for the generalized lasso for Gaussian responses, with close connections to the LARS [Efr04] algorithm. In Chapter 3, we utilize this algorithm to implement a “weighted fused lasso” for CNV analysis.

### 1.2.3 Constrained lasso for linear models

To solve the generalized lasso in settings with binomial responses as well as continuous, we consider the constrained lasso [Jam13]. The constrained lasso problem for a Gaussian response takes the form

$$\min_\beta \left\{ \frac{1}{2} \| y - X \beta \|_2^2 + \lambda \| \beta \|_1 \right\}$$

subject to $A \beta = b$ and $C \beta \leq d$

(1.49)
for full row rank matrices $A$ and $C$. Following the results in [Tib11] and [Jam13], Gaines et al. showed the connection between the generalized lasso and the constrained lasso in [Gai18]. The connection to the constrained lasso is useful for our purposes due to the cost-effective algorithms solving the constrained lasso for generalized linear models [REF]. We begin with the constrained lasso for a Gaussian response to show its connection to the generalized lasso.

In the generalized lasso problem, we seek to minimize the objective function

$$
\min_{\beta} \left\{ \frac{1}{2} \| y - X \beta \|^2_2 + \lambda \| D \beta \|_1 \right\}
$$

for general $m \times p$ regularization matrix $D$. In the proposed method of Chapter 3, we have full column rank $D$; that is, $\text{rank}(D) = r = p$. This allows us to decompose $D$ using singular value decomposition (SVD) as

$$
D = U \Sigma V^T = (U_1, U_2) \begin{pmatrix} \Sigma_1 & \Sigma_2 \\ 0 & 0 \end{pmatrix} V^T
$$

where $U_1 \in \mathbb{R}^{m \times r}$, $U_2 \in \mathbb{R}^{m \times (p-r)}$, $\Sigma_1 \in \mathbb{R}^{r \times r}$, and $V \in \mathbb{R}^{p \times p}$. We define

$$
\alpha = D \beta = U_1 \Sigma_1 V^T \beta
$$

where $\alpha \in \mathbb{R}^m$. Then we have

$$
\beta = (D^T D)^{-1} D^T \alpha
= (D^T D)^{-1} V \Sigma_1 U_1^T \alpha
= V \Sigma_1^{-1} U_1^T \alpha
= D^+ \alpha
$$

where $D^+$ is the Moore-Penrose inverse of $D = U_1 \Sigma_1 V^T$. Gaines et al. [Gai18] show that $\beta$ is uniquely determined if and only if $U_2^T \alpha = 0_{m-r}$. This yields the constrained lasso problem

$$
\min_{\alpha} \left\{ \frac{1}{2} \| y - X D^+ \alpha \|^2_2 + \lambda \| \alpha \|_1 \right\}
\text{subject to } U_2^T \alpha = 0
$$

(1.54)

After solving Eq. 1.54 for $\hat{\alpha}$, we can calculate $\hat{\beta}$ through

$$
\hat{\beta} = D^+ \hat{\alpha}.
$$

Solutions to the constrained lasso problem are obtainable through quadratic programming methods and coordinate descent [Jam12; Jam13]. Additionally, Zhou et al. introduce an algorithm for the complete solution path [Zho14; Zho13; Gai18] which Jeon et al. customize to the compositional data setting and extend to almost quadratic loss functions such as Huber loss [Jeo20]. The solution of

17
Zhou et al. is piecewise linear, where the solution path within each linear segment is determined by a system of ordinary differential equations (ODE). Inflection points on the solution path correspond to constraints become active or inactive, at which point a new ODE system governs the linear path taken in the next segment. The solution path of Jeon et al. shares this quality but additionally can accommodate $n < p$ without a smoothing ridge penalty.

1.2.4 Constrained lasso for logistic regression

The generalized lasso for logistic regression has the optimization function

$$\min_{\beta} \left\{ -\frac{1}{n} \sum_{i=1}^{n} y_i (x_i^T \beta) - \log \left( 1 + e^{(x_i^T \beta)} \right) + \lambda \| D \beta \|_1 \right\}.$$  \hspace{1cm} (1.56)

where $y_i \in \{0, 1\}$ for $i = 1, \ldots, n$. We can reformulate this as a constrained lasso problem as in Eq. 1.54. Define $\tilde{X} = XD^+$ where $X$ and $D$ are defined in Section 1.2.3. Let $U_2$ be as defined above and $\tilde{x}_i$ be the $i$-th row of $n \times m$ matrix $\tilde{X}$. Then the constrained lasso formulation of Eq. 1.56 is

$$\min_{\alpha} \left\{ -\frac{1}{n} \sum_{i=1}^{n} y_i (\tilde{x}_i^T \alpha) - \log \left( 1 + e^{(x_i^T \alpha)} \right) + \lambda \| \alpha \|_1 \right\}$$

subject to $U_2^T \alpha = 0$  \hspace{1cm} (1.57)

The path-following algorithm of Zhou et al. can also accommodate the twice-differentiable loss function in logistic regression. For a full row-rank penalty matrix $D$, a simplified GLM solution exists and is implemented in the SparseReg Matlab package [Gai17]. The work of Jeon et al. is also applicable to GLM loss functions, and their software is under development at https://github.com/jenjong/ComLasso.

1.2.5 Additional topics

1.2.5.1 Unpenalized covariates for linear models

We now present how to include unpenalized covariates in the lasso procedure, such as age, sex, population structure, and other relevant factors in CNV analysis. We decompose the features and coefficients into penalized and unpenalized subsets, with $X = (X_1, X_2)$ and $\beta = (\beta_1, \beta_2)^T$, and exclude the unpenalized coefficients from the lasso penalty. In this example we consider the generalized lasso setting.

For normally distributed response $y$, penalized features $\{X_1, \beta_1\}$, and unpenalized features $\{X_2, \beta\}$, the objective function is

$$O(\beta_1, \beta_2) = \frac{1}{2} \| y - X_1 \beta_1 - X_2 \beta_2 \|_2^2 + \lambda \| D \beta_1 \|_1.$$  \hspace{1cm} (1.58)
Assume that $X_2$ is full rank, and define $H_2 = X_2 (X_2^T X_2)^{-1} X_2^T$, a projection matrix onto the column space of $X_2$. Then we manipulate Eq. 1.58 to obtain

$$
onumber O(\beta_1, \beta_2) = \frac{1}{2} \left\| (H_2 + I - H_2)(y - X_1 \beta_1) - X_2 \beta_2 \right\|_2^2 + \lambda \left\| D \beta_1 \right\|_1$$

$$
= \frac{1}{2} \left\| H_2(y - X_1 \beta_1) + (I - H_2)(y - X_1 \beta_1) - X_2 \beta_2 \right\|_2^2 + \lambda \left\| D \beta_1 \right\|_1
$$

$$
= \frac{1}{2} \left\{ \left\| H_2(y - X_1 \beta_1) - X_2 \beta_2 \right\|_2^2 + \left\| (I - H_2)(y - X_1 \beta_1) \right\|_2^2 + \text{crossproduct} \right\} + \lambda \left\| D \beta_1 \right\|_1 \quad (1.59)
$$

The crossproduct is given by

$$
2 \left\{ H_2(y - X_1 \beta_1) - X_2 \beta_2 \right\}^T \left\{ (I - H_2)(y - X_1 \beta_1) \right\} = 2 \left\{ H_2(y - X_1 \beta_1 - X_2 \beta_2) \right\}^T \left\{ (I - H_2)(y - X_1 \beta_1) \right\}
$$

$$
= 2(y - X_1 \beta_1 - X_2 \beta_2)^T H_2^T (I - H_2)(y - X_1 \beta_1)
$$

$$
= 0 \quad (1.60)
$$

since $H_2$ is a projection matrix which satisfies $H_2^T = H_2$ and $H_2^2 = H_2$. We can express the joint minimization of the objective function over $\beta_1$ and $\beta_2$ as the sequential minimization of the function with respect to $\beta_2$ and then $\beta_1|\beta_2$:

$$
\min_{\beta_1, \beta_2} \left\{ \frac{1}{2} \left\| H_2(y - X_1 \beta_1) - X_2 \beta_2 \right\|_2^2 + \frac{1}{2} \left\| (I - H_2)(y - X_1 \beta_1) \right\|_2^2 + \lambda \left\| D \beta_1 \right\|_1 \right\}
$$

$$
\min_{\beta_1} \min_{\beta_2} \left\{ \frac{1}{2} \left\| H_2(y - X_1 \beta_1) - X_2 \beta_2 \right\|_2^2 + \frac{1}{2} \left\| (I - H_2)(y - X_1 \beta_1) \right\|_2^2 + \lambda \left\| D \beta_1 \right\|_1 \right\} . \quad (1.61)
$$

The inner minimization with respect to $\beta_2$ relies on only the first term of Eq. 1.61, so we have

$$
\hat{\beta}_2 = \arg\min_{\beta_2} \left\{ \left\| H_2(y - X_1 \beta_1) - X_2 \beta_2 \right\|_2^2 \right\} . \quad (1.62)
$$

Taking the derivative of Eq. 1.62 and setting it equal to zero gives us

$$
\hat{\beta}_2 = (X_2^T X_2)^{-1} X_2^T H_2(y - X_1 \beta_1)
$$

$$
= (X_2^T X_2)^{-1} X_2^T (y - X_1 \beta_1) . \quad (1.63)
$$

Plugging in $\hat{\beta}_2$ to Eq. 1.61 makes the first term disappear, so that the minimization of $\beta_1$ conditional on $\beta_2$ is

$$
\min_{\beta_1|\beta_2} \left\{ \left\| (I - H_2)(y - X_1 \beta_1) \right\|_2^2 + \lambda \left\| D \beta_1 \right\|_1 \right\}
$$

$$
= \min_{\beta_1|\beta_2} \left\{ \left\| \hat{y} - \hat{X}_1 \beta_1 \right\|_2^2 + \lambda \left\| D \beta_1 \right\|_1 \right\} \quad (1.64)
$$

which is a standard generalized lasso problem in $\beta_1$. 

19
1.2.5.2 Unpenalized covariates for generalized linear models

For lasso in the GLM setting, coefficient estimates for unpenalized covariates are iteratively updated along with those for penalized covariates. The nature of the unpenalized coefficient updates depends on the algorithm in use, but most methods rely on penalty scaling parameters as illustrated below. We consider \( p_1 \) penalized predictors and \( p_2 = p - p_1 \) unpenalized predictors, and we identify them with the \( p \times 1 \) vector \( \nu = (\nu_1, \cdots, \nu_p) \) which takes value 1 for the \( p_1 \) penalized features and value 0 for the \( p_2 \) unpenalized features. Then minimization problem for penalized logistic regression is

\[
\min_{\beta} \left\{ -\frac{1}{n} \sum_{i=1}^{n} y_i (x_i^T \beta) - \log(1 + \exp(x_i^T \beta)) \right\} + \lambda \sum_{j=1}^{p} \nu_j |\beta_j| \tag{1.65}
\]

where \( x_i^T \) is the \( i \)th row of \( X \) for \( i = 1, \cdots, n \) and \( \nu_j \in \{0,1\} \) depending on whether \( \beta_j \) is penalized or not. For example, the cyclic coordinate descent of Friedman et al. [Fri10] applied to the penalized logistic regression problem above updates coefficient estimates according to

\[
\tilde{\beta}_j \leftarrow S\left( \frac{1}{n} \sum_{i=1}^{n} x_{ij} w_i (z_i - \sum_{k \neq j} x_{ik} \hat{\beta}_k), \lambda \nu_j \right) \frac{\sum_{i=1}^{n} x_{ij}^2 w_i}{\sum_{i=1}^{n} w_i x_{ij}^2}, \tag{1.66}
\]

where \( w_i = \expit(x_i^T \hat{\beta}) \ast (1 - \expit(x_i^T \hat{\beta})) \), and the soft threshold operator \( S(z, \delta) \) is defined as

\[
S(z, \delta) = \text{sign}(z) \cdot \left( |z| - \delta \right)_+ \tag{1.67}
\]

with \( (z)_+ = \max\{z, 0\} \). Thus, for an unpenalized predictor with \( \nu_j = 0 \), the update simplifies to

\[
\tilde{\beta}_j \leftarrow \frac{1}{n} \sum_{i=1}^{n} x_{ij} w_i (z_i - \sum_{k \neq j} x_{ik} \hat{\beta}_k) \frac{\sum_{i=1}^{n} x_{ij}^2 w_i}{\sum_{i=1}^{n} w_i x_{ij}^2}, \tag{1.68}
\]

which is the standard IRLS and Newton-Raphson update for unpenalized logistic regression.

1.2.5.3 Choosing \( \lambda \)

The behavior of the estimates \( \hat{\beta}(\lambda) \) as \( \lambda \) varies, i.e., the solution path, may be of interest and can be visualized in trace plots which show \( \hat{\beta}(\lambda) \) on the y-axis versus \( \lambda \) on the x-axis. Traceplots are particularly revealing in fused and group lasso problems as in [Zho10] where clusters of predictors enter and exit the model together. However, researchers often seek to make inference based on the solutions at a well chosen \( \lambda \). Below we discuss different approaches for selecting that value.

One approach for selecting an optimal \( \lambda \) value involves minimizing a loss function. A common choice of loss function when cross validation is possible is the mean cross-validated error: \( CV(\lambda) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in F_k} (y_i - \hat{y}_i)^2 \) for \( K \) folds. Then the optimal \( \lambda \) is defined as \( \lambda_{MCVE} = \arg \min CV(\lambda), \) out of a user specified vector or inflection points on the exact solution path \( \{ \lambda_1, \cdots, \lambda_L \} = \Lambda \). As standardized in the \texttt{glmnet} R package [Fri10] and [Fri01], the \( \lambda \) resulting in mean CV error within 1 standard error
of the optimal mean CV, called $\lambda_{1SE}$, may be considered as a sparser alternative to $\lambda_{MCV}$ and thus more protected against overfitting.

In our application in Chapter 3, cross validation is less feasible due to the very small proportion of non-zero values in each predictor, in the range of $0.005 - 5\%$ non-zero values by predictor in the simulations and real data analysis. We instead use information criteria to select an optimal value of $\lambda$ based on all data considered together. The Akaike information criterion (AIC) [Sak86] and Bayesian information criterion (BIC) [Sch78] balance the goodness of fit of the model against its complexity. The AIC tends to select the model that is better at predicting and the BIC will select the true model if it is among those tested. The BIC’s optimal selection relies on the presence of the true model among the models tested. Whereas the AIC’s selection of the optimal model assumes the data are generated from a density not represented among the tested models. The AIC is asymptotically equivalent to leave-one-out cross-validation for i.i.d samples under a deviance loss function [Sto77] and the BIC is equivalent to leave-$v$-out cross-validation where $v = n \left[ 1 - 1/(\log(n) - 1) \right]$ [Sha93].

Both the AIC and BIC require that a parametric distribution be specified for the observed data. Then given the log-likelihood $l(\beta)$, the AIC for a fixed $\lambda$ is

$$AIC(\lambda) = 2 \times df(\lambda) - 2l(\hat{\beta}(\lambda)),$$

where $df(\lambda)$ is the degrees of freedom at $\lambda$. The BIC is given by the similar expression

$$BIC(\lambda) = \log(n) \times df(\lambda) - 2l(\hat{\beta}(\lambda))$$

where $n$ is the number of observations. The AIC- and BIC-guided choices of $\lambda$ are then

$$\lambda_{AIC} = \arg \min_{\lambda \in \Lambda} AIC(\lambda)$$

$$\lambda_{BIC} = \arg \min_{\lambda \in \Lambda} BIC(\lambda).$$

As $\lambda$ decreases towards 0, the degrees of freedom $df(\lambda)$ typically increases and the $-2l(\hat{\beta}(\lambda))$ decreases. It is easy to see that for $n > 100$, the BIC weights the increasing degrees of freedom function more heavily than the AIC, thus favoring a sparser model than the AIC for large $n$.

The degrees of freedom for the lasso depends on the flavor of lasso, but in general it can be expressed as a function of the cardinality of the active set as first proposed by Efron et al. [Efr04]. In the standard lasso, an unbiased estimate for the degrees of freedom at a fixed $\lambda$ is the number of non-zero coefficients [Zou07] for full rank $X$ and Gaussian $y$, with results for a general $X$ given in [Tib12]. For the fused lasso, the degrees of freedom is estimated by the number of fused groups [Tib11], and for the sparse fused lasso by the number of non-zero fused groups [Tib05]. The degrees of freedom for the generalized lasso is derived by Tibshirani et al. in [Tib11] for a Gaussian response and full rank design matrix $X$. The degrees of freedom for the generalized lasso with general penalty matrix $D$ cannot be expressed qualitatively but is estimated from the size of the active set; likewise for the constrained lasso [Zho14]. Park and Hastie justify applying degrees of freedom estimates for
linear models to GLM problems in [Par07], invoking the local approximation of GLM log-likelihoods by weighted least squares.
ASSOCIATION TEST USING COPY NUMBER PROFILE CURVES (CONCUR) ENHANCES POWER IN RARE COPY NUMBER VARIANT ANALYSIS

*previously published in PloS Computational Biology (Brucker et al., 2020)

2.1 Introduction

Copy number variants (CNVs) are unbalanced structural variants that are typically 1 kilobase pair (kb) in size or larger and are comprised of more or fewer copies of a region of DNA with respect to the reference genome. CNVs are typically characterized by two descriptive features. The first feature is CNV dosage, or the total number of copies present, with $>2$ copies corresponding to duplications and $<2$ copies corresponding to deletions. The second is the CNV length, typically measured in base pairs (bp) or kilobase pairs. CNVs are important risk factors for many human diseases and traits, including Crohn's disease, HIV susceptibility, and body mass index [McC08; Liu10; Mac17]. Large and rare CNVs are particularly implicated in neuropsychiatric disorders including autism spectrum disorder, schizophrenia, bipolar disorder, and attention deficit disorder [Mal12]. For example, multiple studies have confirmed a greater burden of rare CNVs in schizophrenia cases compared with normal controls, both genome-wide and in specific neurobiological pathways.
important to schizophrenia (e.g., calcium channel signaling and binding partners of the fragile X mental retardation protein).

Rare CNVs (e.g., < 1% frequency) in the genome are intractable to test individually for disease association and instead are examined with collapsing methods. Collapsing methods summarize variant characteristics across multiple variants in a targeted region, such as a gene set, a chromosome or the whole genome, and perform a test of the collective CNV effects on the phenotype. By accumulating information across multiple rare variants, collapsing methods can have enhanced power to detect the effects of rare CNVs that are difficult to detect individually but collectively have a significant impact. Collapsing tests for rare CNVs are primarily built on the foundation of rare single nucleotide polymorphism (SNP) association tests but with additional complexity to accommodate the length and dosage features of CNVs. As with SNPs, the effects of CNVs can vary between loci, and collapsing methods demonstrate improved power as they account for this heterogeneity. However, CNV collapsing tests also need to account for within-locus heterogeneity due to differential dosage effects or length effects within a CNV region.

Similar to SNP collapsing tests, there are two families of tests for rare CNV analysis: burden-based methods and kernel-based methods. Burden-based tests, e.g., Raychaudhuri et al. [Ray10], summarize the CNV features of an individual via total CNV counts or average length and model the CNV effects as fixed effects, assuming etiological homogeneity of features across multiple CNVs in a targeted region. Kernel-based tests, such as CCRET [Tze15] and CKAT [Zha16], aggregate CNV information via genetic similarity based on certain CNV features and model CNV effects as random effects to account for the between-locus etiological heterogeneity. By design, burden tests are optimal when the association signal is driven by homogeneous effects across CNVs, and kernel-based tests are optimal in the presence of etiological heterogeneity. Burden testing often involves multiple analyses on subsets of CNVs according to their dosage (e.g., deletions only or duplications only) or size (e.g. > 100kb, > 500kb) to increase homogeneity; whereas, kernel-based tests do not have such requirements.

In this work, we focus on kernel-based methods, as etiological heterogeneity is becoming a more practically encountered scenario due to high-resolution CNV detection technologies permitting the detection of CNVs of smaller length. In kernel-based association tests, the association between CNVs and the trait is evaluated by examining the correlation between trait similarity and CNV similarity quantified in a kernel. For kernel construction, we can refer to kernel-based tests for SNPs. Since SNPs are evaluated at the same single base-pair position (referred to as a locus) across individuals, it is natural to assess similarity locus-by-locus and aggregate the locus-level similarity over all loci in the target region to obtain an overall measure of SNP similarity. A locus unit for CNVs, however, is not so obvious since CNVs span multiple base pairs and may overlap partially between individuals.

To address this issue, standard CNV kernel-based tests construct CNV regions (CNVRs). For example, the CNV Collapsing Random Effects Test (CCRET) [Tze15], developed previously by our group, defines CNVRs by a user-specified amount of CNV overlapping among different individuals (e.g., Fig 1 of Tzeng et al. [Tze15]). CNV similarity between an individual pair is quantified first
within each CNVR, and then CNVR-level similarity is summed over all CNVRs in the target region to characterize overall CNV similarity. However, a drawback of this approach is that CNVRs defined in this fashion are contingent on the unique CNV overlapping patterns among individuals in a study, and the defined CNVRs can vary from one study to another. The arbitrary choice of overlapping threshold also impacts the formation of locus units and consequently how the “between-locus” and “within-locus” heterogeneous effects of CNVs are accounted for.

To avoid the issues introduced by arbitrarily defined CNVRs as in CCRET, the CNV Kernel Association Test (CKAT) [Zha16] adopts a different strategy to quantify CNV similarity between two individuals. Specifically, CKAT allows users to define the CNVR as a biologically relevant region, e.g., a chromosome. CKAT also introduces a new kernel function to measure CNV similarity based on both dosage and length features between two CNV events. This CNV-level similarity is then aggregated to derive a measure of CNVR-level similarity using a shift-by-one scanning algorithm that “aligns” CNVs in two profiles based on their ordinal position. A multiple-testing correction is applied if multiple CNVRs are involved in the targeted region. Although the new strategy bypasses the need for an arbitrarily defined locus unit, the scanning alignment may yield unreliable results if CNVRs are too large. Additionally, CKAT aligns pairs of CNVs based on their ordinal position, which may or may not optimally capture similarity dependent on the CNV signal sources. There may also be computational considerations with a scanning algorithm when $n$ is large.

To address these challenges in quantifying CNV similarity using kernel-based methods, in this work we propose a new approach called the Copy Number profile Curve-based (CONCUR) association test. Based on the concept of copy number (CN) profile curves (Fig. 2.1), the CONCUR association test naturally incorporates both CNV dosage and length features and can capture their main effects as well as dosage-length interactions. Additionally, building the kernel based on CN profile curves permits the quantification of CNV similarity without the need of pre-specified locus units. Moreover, CNV length may be incorporated flexibly in units which are supported in good resolution by the sequencing technology or which are computationally stable. Like CCRET and CKAT, the test is built in the framework of kernel machine regression and is powerful under heterogeneous signals and can adjust for confounders. We use simulation studies to demonstrate the improved power CONCUR over existing kernel-based methods in a variety of settings and illustrate the practical utility of CONCUR by conducting pathway analysis on data from the Swedish Schizophrenia Study and the Taiwan Biobank.

2.2 Results

2.2.1 Overview of CONCUR

CONCUR assesses the collective effects of rare CNVs on a phenotype in a kernel machine regression framework where the kernel construction does not require a pre-defined CNV locus or region. As such, CONCUR is built on two major components: (a) the CN profile curve, with which we describe an individual's CNVs across the genome or a region of interest; and (b) the common area under the
curve (cAUC) kernel, with which we measure CNV similarity between individuals and characterize the CNV effects on the phenotype. In a CN profile curve (e.g., Fig. 2.1), CNV dosage is shown on the y-axis as jumps or troughs diverging from a baseline of 2 copies. The start and end points of the jumps and troughs correspond to the start and end locations of the CNV and are shown on the x-axis. At genomic locations where there are no CNV events, the y-axis (dosage) takes value 2 (i.e., the baseline value). CN profile curves are intended to be a visualization of CNV activity and concurrence across samples and contribute to the CONCUR method through the concept of cAUC.

By superimposing two CN profile curves, we identify regions of overlapping CNVs of the same type (i.e., deletion or duplication) and propose to use the common area under the curve (cAUC) to quantify CNV similarity between two individuals. To implement the idea, first the raw dosage values in the CN profile curve are centered and scaled to obtain the duplication profile curve and deletion profile curve. Let $d$ denote the dosage of a CNV. The scaling and centering can be achieved by the dosage transform functions: $a^{\text{Dup}}(d) = |2 - d|^c$ for duplications and 0 otherwise, and $a^{\text{Del}}(d) = |2 - d|^c$ for deletions and 0 otherwise, where $c$ is some pre-specified constant. Second, we superimpose the duplication profile curves of two individuals and note the overlapping regions where both curves are non-zero. Third, for each overlapping region, we multiply the minimum of the two respective transformed dosage values by the length of the overlap, and save this measure of “area of commonality”. Finally, we calculate the cAUC between two individuals as the sum of all such areas of commonality in their duplication profile curves plus the sum of all areas in their deletion profile curves. Fig. 2.1 illustrates the cAUCs between various pairs of individuals when setting $c = 1$ in the dosage transform functions, $a^{\text{Dup}}(d)$ and $a^{\text{Del}}(d)$. Generally speaking, the cAUC between individuals with overlapping CNVs of the same dosage $d$ is the overlapping length times $|d - 2|$. For example, for individuals with overlapping CNVs of dosage 0 (Fig. 2.1(b)), the cAUC is the overlapping length times $|0 - 2|$. The cAUC between individuals with overlapping CNVs of the same type but different dosages, $d_1$ and $d_2$, is the length of the overlap times $|d_1 - d_2|$ (e.g., dosages of 3 versus 4 in Fig. 2.1 (c)). If there are multiple overlaps in the individuals’ CN profile curves, the cAUC between two individuals is the sum of all areas of commonality (e.g., sum of shaded regions in Fig. 2.1(d)). The cAUC kernel measures similarity in both CNV length and dosage and hence characterizes the joint dosage and length effects. Using the semi-parametric kernel machine regression framework, CONCUR regresses the trait values on CNV effects captured by the cAUC kernel, and evaluates the association between traits and CNV profiles via a score-based variance component test.
<table>
<thead>
<tr>
<th>ID</th>
<th>CHR</th>
<th>BP1</th>
<th>BP2</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>200</td>
<td>400</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>100</td>
<td>500</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>600</td>
<td>800</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>200</td>
<td>400</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>100</td>
<td>500</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>600</td>
<td>800</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>200</td>
<td>700</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.1 Diagram of copy number profile curves and common area under the curve. (a) Example of CNV data in standard PLINK format describing profiles of individuals in a small region of chromosome 1. (b) Copy number (CN) profile curves illustrating the cAUC between individuals with overlapping deletions of dosage 0. (c) CN profile curves illustrating the cAUC between individuals with overlapping duplications of dosage 3 and 4. (d) CN profile curves which show 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.
2.2.2 Simulation Studies

We conducted three sets of simulations: whole genome analysis based on TwinGene pseudo-CNV (TGP) data (referred to as TGP-WG simulations); chromosome 1 analysis based on TGP data (referred to as TGP-Chr1 simulations); and chromosome 1 analysis based on Taiwan Biobank (TWB) CNV data (referred to as TWB-Chr1 simulations). The dosage values are integers in the TGP dataset and are continuous in TWB dataset. With these simulations, we evaluated the performance of CONCUR, CCRET, and CKAT under various signal patterns and different sources of effect heterogeneity.

To implement CCRET, we applied the functions from the CCRET package to convert the PLINK data to CCRET design matrices and computed the dosage kernel matrix. We used 1-bp overlapping of CNVs among different individuals to form CNVRs as in the CCRET paper [Tze15]; that is, as long as CNVs of different individuals overlapped by \( \geq 1 \) bp, it was considered an “overlap”. For CKAT, we designated a chromosome as a CNVR and performed an association test for each CNVR using the CKAT package. CNV lengths within each chromosome were scaled to be in \([0,1]\) by dividing by the range of CNV activity in each chromosome, i.e., the maximal ending position minus the minimal starting position of observed CNVs on each chromosome. The Gaussian kernel scaling parameter was set to be 1. In the TGP-WG simulations, as there were 22 CKAT p-values corresponding to the 22 chromosomes, we took the minimum p-value and used Bonferroni’s procedure to compute the adjusted p-value for multiple testing. Finally, we built the CCRET and CKAT kernels using CNVs’ categorical dosage (duplication, deletion, and normal) as required in the software packages. To assure comparability, CONCUR was built using categorical dosages (referred to as CONCUR_cat). We also implemented CONCUR using the original integer dosage values (referred to as CONCUR_int) in the TGP-Chr1 simulations, and using the original continuous dosages (referred to as CONCUR_cont) in the TWB-Chr1 simulation.

2.2.2.1 Datasets used for simulations

The TwinGene pseudo (TGP) CNV dataset of 2000 individuals is publicly available at https://www4.stat.ncsu.edu/~jytzeng/Software/CCRET/software_ccret.php. Autosomal-wide pseudo-CNV data were simulated by mimicking the CNV profiles of unrelated individuals in the TwinGene study [Bee03], and the details are described in Tzeng et al. [Tze15]. Briefly, the TwinGene study used a cross-sectional sampling design and included over 6,000 unrelated subjects born between 1911 and 1958 from the Swedish Twin Registry [Ped02; Lic06]. CNV calls were generated using the Illumina OmniExpress BeadChip for 72,881 SNP markers and using PennCNV (version June 2011) [Wan07] as the CNV calling algorithm with recommended model parameters. From the full callset, high quality rare CNVs (frequency \(< 1\% \) and size \(> 100\) kb) were extracted to form the simulation pool for the pseudo-CNV data. CNV dosages in this dataset are integers (0, 1, 2, 3, and 4). In the TGP whole genome data (chromosome 1 to chromosome 22), each of the 2000 individuals has at least one CNV, and in the chromosome 1 analysis, 291 individuals have CNV activity.

The Taiwan Biobank (TWB) CNV dataset is from the Taiwan Biobank project (https://www.
This data was studied in the real data analysis and further details regarding it are shared in Section 2.2.3.2. CNV dosage values in this dataset are continuous with dosages $\geq 2.3$ indicating duplications and $\leq 1.7$ indicating deletions. Out of the 11,664 individuals who were included in the real data analysis, we took a random sample of 2000 individuals and kept their data for all CNVs in chromosome 1. In this sample, 1432 individuals out of the 2000 had CNV activity in chromosome 1.

### 2.2.2.2 Simulation design

For the purpose of simulating phenotypes, we constructed "CNV segments" based on the CNVs in the focal dataset. The endpoints of the segments correspond to locations where a CNV in any one of the samples begins or ends, resulting in segments that contain either one or more intersecting CNVs. Within a segment, CNV dosage of an individual is a constant, and CNVs across individuals may have different dosages though they share the same starting and ending positions defined by the boundaries of the segment. Note that different segments will naturally have different lengths. In the simulation studies, we built design matrices $Z_{Dup}$, $Z_{Del}$, and $Z_{Len}$ which codified CNV features by segment in the CNV dataset. The dosage matrices, $Z_{Dup}$ and $Z_{Del}$, took value 0 for those individuals without CNVs in the segment and were coded as the number of additional or missing copies comprising the CNV otherwise. $Z_{Len}$ was the length of the CNV segment in kb for individuals with CNV events and was 0 for individuals without CNVs in the segment.

A case-control phenotype was generated from the logistic model

$$
\logit(Pr(Y_i = 1)) = \gamma_0 + \beta_X X_i + \sum_{j=1}^{R} \beta^D_{ij} Z_{ij}^{Dup} + \sum_{j=1}^{R} \beta^D_{ij} Z_{ij}^{Del} + \sum_{j=1}^{R} \beta^L_{ij} Z_{ij}^{Len} \\
+ \sum_{j=1}^{R} \beta^{Dup\times Len}_{ij} Z_{ij}^{Dup} Z_{ij}^{Len} + \sum_{j=1}^{R} \beta^{Del\times Len}_{ij} Z_{ij}^{Del} Z_{ij}^{Len},
$$

(2.1)

where $Z_{ij}^*$ is the $(i, j)$ entry of matrix $Z^*$, $i = 1, \ldots, N$ indexes individuals, and $j = 1, \ldots, R$ indexes CNV segments. A binary covariate $X_i$ was simulated from Bernoulli(0.5) for each individual. $\beta^D_{ij}$ and $\beta^D_{ij}$ are the log odds ratios of segment $j$ for the presence of a CNV versus the absence. Likewise, $\beta^L_{ij}$ controls the effect of CNV length in segment $j$, and $\beta^{Dup\times Len}_{ij}$ and $\beta^{Del\times Len}_{ij}$ allow the effects of CNV length to differ by dosage. $\beta^L_{ij} > 0$ (or $< 0$) corresponds to a deleterious (or protective) CNV effect, and $\beta^L_{ij}$ was set to 0 in non-causal segments. We set $\beta_X = \log(1.1)$ and $\gamma_0 = -2$, which corresponds to a baseline disease rate of roughly 0.12. We also fixed $\beta^L_{ij} = 0$ to reflect the observation that length tends to act like an effect modifier of dosage effects.

In the TGP-WG simulations, we generated phenotypes from CNV dosage×length effects and from dosage-only effects. We chose these signals to roughly replicate the simulation settings applied to assess CKAT in [Zha16] (dosage×length signal) and CCRET in [Tze15] (dosage signal). In the TGP-Chr1 and TWB-Chr1 simulations, signals were generated from CNV dosage×length effects. Below we describe the settings for the Chr1-based simulations; the TGP-WG simulations basic
settings are similar and are detailed in Appendix A.2.

In the Chr1-based simulations, we considered three types of causal effects: causal effects from both duplications and deletions, causal effects from duplications only, and causal effects from deletions only. Under each effect type, we designated varying percentages of the causal segments to be deleterious (D) or protective (P). When both duplications and deletions were causal, the settings included $\left(D_{\text{Dup}}, P_{\text{Dup}}, D_{\text{Del}}, P_{\text{Del}}\right) = (90, 10, 90, 10)$, i.e., both causal duplications and causal deletions had 90% deleterious and 10% protective effects, as well as (90, 10, 10, 90) and (10, 90, 90, 10). In the scenarios where duplications or deletions alone were causal, settings included $(D_{\bullet}, P_{\bullet}) = (90, 10), (50, 50), \text{ and } (10, 90)$.

In the TGP-Chr1 simulations, we randomly selected 40 segments across chromosome 1 to be causal, comprised of 20 segments containing $\geq 1$ duplication and 20 segments containing $\geq 1$ deletion. As the segments were formed purely based on the relative CNV patterns among individuals and could be very short in length, we also required causal segments to be at least as long as the median length of all segments of that type (35kb for duplications, 46kb for deletions) to ensure that they had realistic lengths. We allowed for the possibility of duplication and deletion effects arising from the same location, and used categorical dosages to simulate the length-dosage effects. That is, when simulating phenotypes using Eq. 2.1, for individual $i$ and segment $j$, we set $Z_{ij}^{\text{Dup}} = 1$ if a duplication was present in the segment and 0 otherwise, and set $Z_{ij}^{\text{Del}} = 1$ if a deletion was present in the segment and 0 otherwise. We implemented CONCUR in all TGP-Chr1 simulations using the CONCUR_cat kernel as well as the CONCUR_int kernel. We refer to the three scenarios described here (i.e., effects from duplications only, from deletions only, and from combined effects) as TGP-Chr1(a) to distinguish it from the sensitivity analyses introduced below.

In the scenario with causal effects from duplications and deletions combined, we considered two additional scenarios as sensitivity analyses: TGP-Chr1(b) examines the methods’ performance under inaccuracy in the called end-points of CNVs; and TGP-Chr1(c) imposes a more rare baseline disease rate. In TGP-Chr1(b), we added random uniformly distributed errors to the endpoints (BP1 and BP2) of all CNVs. The CNV endpoints in the error-added data differed from the CNV endpoints in the true data used to generate phenotypes by up to $\pm 2.5\%$ of the total length of the CNV. In TGP-Chr1(c), we set $\gamma_0 = -3$ to lower the baseline disease rate to roughly 0.05.

In the TWB-Chr1 simulations, we randomly selected 600 CNV segments across chromosome 1 to be causal, comprised of 300 segments containing $\geq 1$ duplication and 300 segments containing $\geq 1$ deletion. We imposed similar criteria on the length of causal segments as in the TGP-Chr1 simulations, and we allowed for duplication and deletion effects from the same location. Unlike the TGP-Chr1 simulations, here we used continuous dosages to simulate the length-dosage effects in all three scenarios (referred to as TWB-Chr1(a) simulations). In the scenario with combined duplication and deletion causal effects, we also generated signals from categorical dosages (referred to as TWB-Chr1(b) simulations). When simulating phenotypes based on continuous dosage signals using Eq. 2.1, we constructed the dosage matrices such that for a CNV of dosage $d$ in segment $j$ for individual $i$, $Z_{ij}^{\text{Dup}} = |d - 2|$ if a duplication was present and 0 otherwise, and $Z_{ij}^{\text{Del}} = |d - 2|$ if
a deletion was present and 0 otherwise. We applied the CONCUR_cont kernel and CONCUR_cat kernel in both settings (a) and (b) of the TWB-Chr1 simulations to evaluate their robustness to signals arising from dosage values of the same type (categorical or continuous) or of an incongruent type.

We implemented case-control sampling to obtain 2000 cases and 2000 controls for each simulation replication. Type I error rates were evaluated based on 5000 replications, and power was estimated based on 300 replications at each effect size. For all methods, we adjusted for a simulated binary covariate as a fixed effect in the kernel machine regression. We employed the small-sample variance components test of Chen et al. [Che16] and obtained p-values using Davies’ method [Dav80] as implemented in the CKAT R functions.

### 2.2.2.3 Results of TGP-WG simulations

The type I error rates of CONCUR, CCRET, and CKAT were examined at nominal levels of 0.01, 0.05, and 0.1 in the TGP-WG simulations (Table 2.1). All methods had type I error rates around the nominal level.

#### Table 2.1 Type I error rates. Type I error rates of three CNV tests evaluated based on 5000 replications.

<table>
<thead>
<tr>
<th>Nominal level</th>
<th>CONCUR</th>
<th>CCRET</th>
<th>CKAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.010</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td>0.05</td>
<td>0.045</td>
<td>0.047</td>
<td>0.049</td>
</tr>
<tr>
<td>0.10</td>
<td>0.096</td>
<td>0.093</td>
<td>0.092</td>
</tr>
</tbody>
</table>

The power of the methods under causal dosage×length effects in TGP-WG is shown in Fig. A.1. We observe that CONCUR has the best or comparable power with the second best method (CCRET) across different patterns of deleterious-protective effects and in the duplication, deletion, and combined effects scenarios. Both CONCUR and CKAT are designed to detect dosage×length signals, but CKAT struggled to pick up the signals. One possible reason might be the multiple testing penalty applied to CKAT. In addition, the CNV signals in these simulations originate from aligned genomic regions; such signals may or may not be well-captured by CKAT, since its scanning algorithm may incorporate CNV similarity from off-position CNV events or same-position CNV events dependent on the data.

We note that we do not expect the methods to display similar relative performance in the duplication-only causal effects and deletion-only causal effects scenarios. This is because the causal segments for duplications versus deletions have different characteristics, due to differences in the patterns of duplications overlapping versus deletions overlapping in the data. In addition, the strength of the signal in the duplication-only and deletion-only simulations differs due to differences in the length of the segments as well as the frequency of CNVs in the protective versus deleterious
segments. For example, in the duplication-only effects \((D,P) = (10,90)\) setting, the combination of longer segments and more CNV activity in the protective segments leads to a proportionally larger protective signal than in the corresponding deletion-only setting. The result of these differences is asymmetry in the methods' performance in the two settings.

The power under causal dosage effects is shown in Fig. A.2. As expected, the dosage-based CCRET kernel performs the best, with CONCUR following CCRET or having comparable power.

### 2.2.2.4 Results of TGP-Chr1 simulations

The results of the TGP-Chr1(a) simulations are shown in Fig. 2.2. We observe that CONCUR has higher than or comparable power to the second best method, and here the second best method is CKAT or CCRET depending on the scenario. After further exploration of the TGP-Chr1(a) simulations as well as TWB-Chr1 dosage×length simulations, it appears that the relative performance between CKAT and CCRET depends more heavily on whether the causal signals can be well-captured by the CKAT kernel than what the patterns of causal effects are (i.e., causal effects from duplications, deletions, or both, or the proportion of deleterious vs. protective effects). The power from CONCUR_cat and CONCUR_int appear to be identical in all settings, which is not surprising given that there are very few CNVs of larger magnitude (0 or 4+) in the data.
Fig. 2.2 Power comparison between CONCUR, CCRET, and CKAT in the TGP-Chr1(a) simulations (causal dosage×length effects in chromosome 1 of TGP data). The top panel shows power under combined duplication and deletion effects, the middle panel shows power under effects from duplications only, and the bottom panel shows power under effects from deletions only. Different proportions of deleterious vs. protective effects are considered as indicated by \((D_{\text{Dup}},P_{\text{Dup}},D_{\text{Del}},P_{\text{Del}})\) with \(D_{\text{Dup}}\) and \(P_{\text{Dup}}\) reflecting the proportions of deleterious and protective segments among causal duplication segments, and with \(D_{\text{Del}}\) and \(P_{\text{Del}}\) defined similarly for causal deletion segments.

Fig. 2.3 shows the performance of the methods applied to data with inaccurate CNV endpoint information (top panel) and under a lower baseline disease rate (5%; lower panel). For both analyses, the top row of Fig. 2.2 is a useful reference showing the methods’ performance under error-free CNV data and a higher baseline disease rate (12%). The top panel of Fig. 2.3 shows that CONCUR still has higher power than the baseline methods, although the gap between CONCUR and CCRET...
is smaller compared to the error-free scenario. The lower panel of Fig. 2.3 demonstrates that the performance of the methods under a lower baseline disease rate is very comparable to that under a higher disease rate.

Figure 2.3 Power comparison between CONCUR, CCRET, and CKAT in TGP-Chr1(b) and TGP-Chr1(c) (causal dosage×length effects from both duplications and deletions in chromosome 1 of TGP data). The top panel shows power under TGP-Chr1(b), in which the kernels are built on CNV data with error added to the CNV endpoints, mimicking the scenario of inaccuracy end points of called CNVs. The bottom panel shows power under TGP-Chr1(c), in which the disease base rate is more rare (5%). Different proportions of deleterious vs. protective effects are considered as indicated by \((D_{\text{Dup}}, P_{\text{Dup}}, D_{\text{Del}}, P_{\text{Del}})\) with \(D_{\text{Dup}}\) and \(P_{\text{Dup}}\) reflecting the proportions of deleterious and protective segments among causal duplication segments, and with \(D_{\text{Del}}\) and \(P_{\text{Del}}\) defined similarly for causal deletion segments.

### 2.2.2.5 Results of TWB-Chr1 simulations

The results of TWB-Chr1(a) are shown in Fig. 2.4. We observe little difference in the power of the two CONCUR approaches. We also observe that CONCUR had stronger power than CCRET and CKAT, with the exception of some degenerate behavior in CONCUR under \((D_{\text{Dup}}, P_{\text{Dup}}, D_{\text{Del}}, P_{\text{Del}}) = (10, 90, 0, 0)\). In this setting, we suspect that CONCUR’s behavior is due to the combination of a couple factors. There are few duplication events to begin with in the TWB simulated data, and that combined with the 90% protective effects leads to the simulated controls being comprised
primarily of “random” controls with relatively few individuals carrying causal protective CNVs. This results in an extremely weak signal-to-noise ratio. All methods were affected in this setting, such that we needed to significantly boost the range of effect sizes to observe power in any method (e.g., 1.01-1.10 here vs. 1.0005-1.0035 under \((D_{\text{Dup}}, P_{\text{Dup}}, D_{\text{Del}}, P_{\text{Del}}) = (0, 0, 10, 90))\). We believe that CCRET and CKAT are more robust in this setting due to borrowing information across loci through CNVRs and through across-position alignment, respectively. Finally, as in the TGP-Chr1 simulations, under some scenarios CKAT had higher power than CCRET, e.g., in the settings of \((D_{\text{Dup}}, P_{\text{Dup}}, D_{\text{Del}}, P_{\text{Del}}) = (50, 50, 0, 0)\) and \((0, 0, 50, 50)\). The relative performance of the two methods is again likely dependent on whether the causal signals were well captured by the CKAT kernel.

Fig. 2.5 evaluates the ability of the CONCUR_cont approach to detect a signal from categorical dosages, on which CONCUR_cat, CCRET, and CKAT are built. We observe nearly identical power in CONCUR_cont and CONCUR_cat.
Figure 2.4 Power comparison between CONCUR, CCRET, and CKAT in TWB-Chr1(a) (causal continuous dosage × length effects in chromosome 1 of TWB data). The top panel shows power under combined duplication and deletion effects, the middle panel shows power under effects from duplications only, and the bottom panel shows power under effects from deletions only. Different proportions of deleterious vs. protective effects are considered as indicated by (D<sub>Dup</sub>,P<sub>Dup</sub>,D<sub>Del</sub>,P<sub>Del</sub>) with D<sub>Dup</sub> and P<sub>Dup</sub> reflecting the proportions of deleterious and protective segments among causal duplication segments, and with D<sub>Del</sub> and P<sub>Del</sub> defined similarly for causal deletion segments.

2.2.3 Real data application

In real data applications, we first conducted CNV association tests on a previously analyzed CNV dataset from the Swedish Schizophrenia Study as a proof of concept. We next conducted a CNV-triglyceride (TG) association analysis on data from the Taiwan Biobank. We constructed kernels for
Figure 2.5 Power comparison between CONCUR, CCRET, and CKAT for TWB-Chr1(b) (causal categorical dosage × length effects in chromosome 1 of TWB data). The panels show power under combined duplication and deletion effects. Different proportions of deleterious vs. protective effects are considered as indicated by \((D_{\text{Dup}}, P_{\text{Dup}}, D_{\text{Del}}, P_{\text{Del}})\) with \(D_{\text{Dup}}\) and \(P_{\text{Dup}}\) reflecting the proportions of deleterious and protective segments among causal duplication segments, and with \(D_{\text{Del}}\) and \(P_{\text{Del}}\) defined similarly for causal deletion segments.

all methods (CONCUR, CKAT and CCRET) based on categorical dosages, and dropped the “_cat” suffix in our discussion of the CONCUR method in this section for simplicity. We used the step-wise Holm method to adjust for multiple testing, which yields more powerful results than the Bonferroni procedure and remains valid when applied to dependent p-values [Hol79].

2.2.3.1 CNV analysis on schizophrenia in the Swedish Schizophrenia Study

We conducted pathway-based CNV analysis on data from the Swedish Schizophrenia Study [Rip13]. The Swedish Schizophrenia Study used a case-control sampling design. Genotyping was done in six batches using Affymetrix 5.0 (3.9% of the subjects), Affymetrix 6.0 (38.6%), and Illumina OmniExpress (57.4%). PennCNV [Wan07] was used to generate CNV calls. After quality control, we obtained a high quality rare CNV (frequency < 1% and size > 100kb) dataset in 8,457 subjects (3,637 cases and 4,820 controls) [Sza14]. Previous analyses of this data [Sza14] indicated significant associations of large rare CNVs with schizophrenia risk for both genome-wide dosage effects and gene intersecting effects of selected gene sets.

To evaluate the practical utility of the three kernel-based tests, we performed analysis on the gene sets previously examined in [Tze15], excluding the PSD pathway as it overlaps the other three PSD-related pathways considered. In the 8 gene sets, large (> 500kb) rare CNVs were found to be associated with schizophrenia by Szatkiewicz et al. [Sza14], and these associations were corroborated by Tzeng et al. [Tze15] in a gene-interruption analysis with CNVs > 100kb. In each pathway analysis, we performed association tests for joint dosage and length effects of rare CNVs > 100kb, using a fixed effect term to adjust for batch effects. CONCUR and CKAT kernels were constructed from the raw PLINK data and the CCRET dosage kernel was created using the functions available on the
CCRET website. For CKAT, we used pathways as the CNVR unit instead of chromosomes because there were multiple chromosomes with only one gene.

After applying Holm’s procedure to adjust for multiple testing on the 8 pathways, CONCUR and CKAT found significant associations in all 8 pathways, while CCRET identified 4 pathways as significant (Table 2.2). We observed stronger power in CKAT in the analysis here compared to the power observed in the simulation studies. CKAT and CONCUR are more sensitive to dosage-length effects while CCRET is more sensitive to dosage effects; thus, these results suggest significant CNV effects from dosage \times length or length affecting schizophrenia risk in several pathways.
Table 2.2 Association test results for the effects of CNVs with $> 100$kb in length on schizophrenia risk in the Swedish Schizophrenia Study. Raw p-values are reported for CONCUR, CCRET, and CKAT. Asterisks indicate p-values that were significant after a Holm multiple-testing adjustment. Pathways are ordered alphabetically.

<table>
<thead>
<tr>
<th>Gene-set Name</th>
<th># Genes</th>
<th># Genes Interrupted in Cases</th>
<th># Genes Interrupted in Controls</th>
<th>CONCUR</th>
<th>CCRET</th>
<th>CKAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytoplasm (Kirov et al. [Kir12])</td>
<td>266</td>
<td>28</td>
<td>32</td>
<td>0.00124</td>
<td>0.01408</td>
<td>0.00030</td>
</tr>
<tr>
<td>FMRP targets (Darnell et al. [Dar11])</td>
<td>810</td>
<td>149</td>
<td>152</td>
<td>2.29E-05</td>
<td>0.00044</td>
<td>0.00026</td>
</tr>
<tr>
<td>Mental Retardation</td>
<td>503</td>
<td>67</td>
<td>63</td>
<td>0.00164</td>
<td>0.10200</td>
<td>0.00350</td>
</tr>
<tr>
<td>PSD/mGluR5 (Kirov et al. [Kir12])</td>
<td>38</td>
<td>4</td>
<td>7</td>
<td>0.00040</td>
<td>0.10540</td>
<td>0.00129</td>
</tr>
<tr>
<td>PSD/NMDAR (Kirov et al. [Kir12])</td>
<td>61</td>
<td>12</td>
<td>12</td>
<td>0.00102</td>
<td>0.00922</td>
<td>0.00046</td>
</tr>
<tr>
<td>PSD/PSD-95 (Kirov et al. [Kir12])</td>
<td>65</td>
<td>13</td>
<td>10</td>
<td>0.00052</td>
<td>0.00144</td>
<td>0.00903</td>
</tr>
<tr>
<td>Synaptic genes (Ruano et al. [Rua10])</td>
<td>718</td>
<td>154</td>
<td>164</td>
<td>5.45E-06</td>
<td>0.02005</td>
<td>0.00766</td>
</tr>
<tr>
<td>Synaptic Proteome (G2Cdb)</td>
<td>1023</td>
<td>121</td>
<td>106</td>
<td>0.00067</td>
<td>0.00010</td>
<td>0.00736</td>
</tr>
</tbody>
</table>
2.2.3.2 CNV analysis on triglycerides in the Taiwan Biobank

We applied the proposed CONCUR test to the Taiwan Biobank (TWB) data (https://www.twbiobank.org.tw/new_web/) and conducted CNV association analysis with triglyceride (TG) levels on lipid-related pathways. The nationwide biobank project was initiated in 2012 and has recruited more than 15,995 individuals. Peripheral blood specimens were extracted from healthy donors and genotyped using the Affymetrix Genomewide Axiom TWB array, which was designed specifically for a Taiwanese population. The TWB array contains 653,291 SNPs and was used to generate calls for genome-wide CNVs in the following process. First, Affymetrix Power Tools version 1.18.0 was used to produce a summary file of the intensity values of all probes, and the file was input into the Partek Genomic Suite version 6.6 to call CNVs based on the following criteria: at least 35 consecutive SNP markers, p-values of different CN values between two consecutive segments < 0.001, and signal-to-noise ratio (SNR) ≥ 0.3. A duplication was called if its copy number was ≥ 2.3, and a deletion was called if its copy number was ≤ 1.7. Several previous studies [Lu11][Lai14] have demonstrated appropriate CNV calls with these parameters. After quality control, we obtained CNV data in 14,595 unrelated individuals. Our CNV association analyses focused on a subset of 11,664 individuals who had non-missing TG levels.

We referenced the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database [Kan16] to identify lipid-related pathways. Among the 17 pathways related to “lipid metabolism”, 15 pathways included genes intersected by the TWB CNV data and were selected as candidate pathways in our analysis. For each pathway analysis, we adjusted for sex, age, BMI, and the top 10 principal components representing the population structure as covariates with fixed effects. As before, CKAT was implemented with each pathway comprising a single CNVR.

Similar to the Swedish Schizophrenia Study analysis, after applying Holm’s procedure to adjust for multiple testing on the 15 pathways, all 15 pathways were found significant by CONCUR and CKAT, and 1 pathway was found significant by CCRET (Table 2.3). CKAT again demonstrated much better power than in the simulation studies. The relative performance among the three methods may be due to more dominant length or dosage×length signals.
Table 2.3 Association test results for the effects of CNVs on triglyceride levels in the Taiwan Biobank. Raw p-values are reported for CONCUR, CCRET, CKAT, and a negative control test in which CONCUR is applied to a randomly permuted response vector. Asterisks indicate p-values that were significant after a Holm multiple-testing adjustment. Pathways are ordered alphabetically.

<table>
<thead>
<tr>
<th>Gene-set Names</th>
<th># Genes</th>
<th># Genes Interrupted</th>
<th>CONCUR</th>
<th>CCRET</th>
<th>CKAT</th>
<th>Neg. Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa00061 (Fatty acid biosynthesis)</td>
<td>13</td>
<td>12</td>
<td>0.00171*</td>
<td>0.01187</td>
<td>0.00197*</td>
<td>0.89046</td>
</tr>
<tr>
<td>hsa00062 (Fatty acid biosynthesis)</td>
<td>30</td>
<td>26</td>
<td>0.00031*</td>
<td>0.01591</td>
<td>0.00508*</td>
<td>0.36231</td>
</tr>
<tr>
<td>hsa00071 (Fatty acid degradation)</td>
<td>44</td>
<td>43</td>
<td>0.00406*</td>
<td>0.01088</td>
<td>0.00631*</td>
<td>0.78391</td>
</tr>
<tr>
<td>hsa00072 (Synthesis and degradation of ketone bodies)</td>
<td>10</td>
<td>10</td>
<td>0.00008*</td>
<td>0.00459</td>
<td>0.00383*</td>
<td>0.37588</td>
</tr>
<tr>
<td>hsa00100 (Steroid biosynthesis)</td>
<td>19</td>
<td>16</td>
<td>0.01641*</td>
<td>0.00618</td>
<td>0.00906*</td>
<td>0.28308</td>
</tr>
<tr>
<td>hsa00120 (Primary acid bile biosynthesis)</td>
<td>17</td>
<td>17</td>
<td>0.00019*</td>
<td>0.00314*</td>
<td>0.00274*</td>
<td>0.56475</td>
</tr>
<tr>
<td>hsa00140 (Steroid hormone biosynthesis)</td>
<td>60</td>
<td>58</td>
<td>0.00030*</td>
<td>0.00623</td>
<td>0.00159*</td>
<td>0.25125</td>
</tr>
<tr>
<td>hsa00561 (Glycerolipid metabolism)</td>
<td>61</td>
<td>50</td>
<td>0.00430*</td>
<td>0.00494</td>
<td>0.00198*</td>
<td>0.46773</td>
</tr>
<tr>
<td>hsa00564 (Glycerophospholipid metabolism)</td>
<td>97</td>
<td>86</td>
<td>0.00322*</td>
<td>0.00398</td>
<td>0.00209*</td>
<td>0.52954</td>
</tr>
<tr>
<td>hsa00565 (Ether lipid metabolism)</td>
<td>47</td>
<td>43</td>
<td>0.00018*</td>
<td>0.00859</td>
<td>0.00439*</td>
<td>0.50786</td>
</tr>
<tr>
<td>hsa00590 (Arachnidonic acid metabolism)</td>
<td>63</td>
<td>62</td>
<td>0.00212*</td>
<td>0.00883</td>
<td>0.00211*</td>
<td>0.78048</td>
</tr>
<tr>
<td>hsa00591 (Linoleic acid metabolism)</td>
<td>29</td>
<td>29</td>
<td>0.00080*</td>
<td>0.01799</td>
<td>0.00291*</td>
<td>0.65068</td>
</tr>
<tr>
<td>hsa00592 (alpha-Linolenic acid metabolism)</td>
<td>25</td>
<td>25</td>
<td>0.00581*</td>
<td>0.00927</td>
<td>0.00273*</td>
<td>0.74765</td>
</tr>
<tr>
<td>hsa00600 (Sphingolipid metabolism)</td>
<td>47</td>
<td>43</td>
<td>0.00382*</td>
<td>0.00512</td>
<td>0.00789*</td>
<td>0.31366</td>
</tr>
<tr>
<td>hsa01040 (Biosynthesis of unsaturated fatty acids)</td>
<td>27</td>
<td>23</td>
<td>0.00012*</td>
<td>0.00394</td>
<td>0.00158*</td>
<td>0.77243</td>
</tr>
</tbody>
</table>
Next, we explored associations with TG levels in the TWB data on a by-chromosome basis (Table 2.4). Using the Holm method to adjust for multiple testing on 22 chromosomes, CONCUR found all 22 chromosomes significantly associated with TG, CKAT found 12 significant chromosomes, and CCRET found none. It is not unexpected to see all 22 chromosomes identified to be significantly associated with TG, since the genes in the 15 significant lipid metabolism pathways examined are located across all 22 chromosomes. For example, for the 12 chromosomes identified by both CONCUR and CKAT, the number of pathway genes intersected by CNVs ranges from 8 to 41, and for those chromosomes uniquely identified by CONCUR (excluding chromosome 13), the number of intersected genes ranges from 4 to 36. In chromosome 13, while there is only one pathway gene, all CNVs are located in chr13q, which is a well-studied region related to cholesterol metabolism [Kno00][AK02][Klo01]. Since cholesterol is strongly related to TG levels, CNVs in chr13q and chr13q22-q32 may impact TG levels by affecting the metabolism efficiency of TG and cholesterol. To further interpret the significant CONCUR test result, we examined the subregion chr13q22-q32 that is highlighted in [AK02] and contains or overlaps with the markers in [Kno00] and [Klo01]. By applying CONCUR, CKAT and CCRET to this subregion, we obtained the p-values as 0.0000143, 0.0004388 and 0.0242815, respectively. These results suggest a length or dosage×length signal arising from chr13q22-q32, which CONCUR and CKAT can detect with good power. This length-driven CNV signal is not well captured by CCRET in both of the subregion and chromosome-wide analyses, since CCRET does not account for CNV length features. The strength of the signal from chr13q22-q32 may be diluted for CKAT when the entire chromosome is treated as a CNVR.

As further assurance that these associations are less likely due to false positives, we conducted a CONCUR negative-control analysis by repeating the by-chromosome analysis using permuted TG levels. The resulting p-values are shown in Table 2.4; the p-value range of those chromosomes identified by both CONCUR and CKAT (i.e., 0.483 to 0.888) is similar to that of those uniquely identified by CONCUR (i.e., 0.485 to 0.963). In addition, we also examined the quantile-quantile plots (Q-Q plots) of CONCUR p-values from negative-control analyses, by generating (a) 20 TG-permuted datasets for each of the 15 pathways, and (b) 1000 TG-permuted datasets for chromosome 13. The Q-Q plots, shown in Fig. A.4 and Fig. A.5, suggest that the p-values follow the expected null distribution, Uniform(0,1).

Finally, we illustrate in Appendix A.4 possible CONCUR post-hoc analyses probing the potential sources of a CNV association identified at the aggregate-level. As an example, we looked more closely at pathway hsa01040 (biosynthesis of unsaturated fatty acids), for which both CONCUR and CKAT were significant but not CCRET. In short, we calculated summary statistics describing CNV length and dosage in hsa01040 for individuals with different levels of TG (low, medium, and high), and examined CNV features in all CNVs together and in duplications and deletions separately. We also used heatmaps to visualize CNVs in the 23 genes in hsa01040 (Fig. A.3), displaying the duplications and deletions intersecting genes in all CNV profiles categorized by their TG level. These exploratory analyses suggested that for duplications only, there may exist “promising” differences in CNV length and relatively weaker differences in dosage across TG levels. Because these “promising” associations

42
Table 2.4 Association test results for the effects of CNVs on triglyceride levels by chromosome in the Taiwan Biobank. Results from the CONCUR, CCRET, and CKAT association tests are shown. The results of the negative control analysis reflect the p-value from CONCUR applied to a randomly permuted response vector. Asterisks indicate p-values that were significant after a Holm multiple-testing adjustment. Pathways are ordered according to chromosome. For interpretation of the by-chromosome association tests, the number of genes from the 15 lipid metabolism pathways that are intersected by CNVs is given (# Genes Interrupted).

<table>
<thead>
<tr>
<th>Chromosomes</th>
<th>P-values</th>
<th>CONCUR</th>
<th>CCRET</th>
<th>CKAT</th>
<th>Neg. Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr # Genes Interrupted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>41</td>
<td>0.000018*</td>
<td>0.734017</td>
<td>0.000017*</td>
<td>0.739204</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>0.000007*</td>
<td>0.077438</td>
<td>0.000118*</td>
<td>0.790286</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>4.41E-08*</td>
<td>0.363502</td>
<td>0.000834*</td>
<td>0.483097</td>
</tr>
<tr>
<td>4</td>
<td>36</td>
<td>5.11E-08*</td>
<td>0.257951</td>
<td>0.035544</td>
<td>0.900595</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>1.25E-07*</td>
<td>0.153555</td>
<td>0.003703*</td>
<td>0.865209</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>4.00E-07*</td>
<td>0.080112</td>
<td>0.000105*</td>
<td>0.735159</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>0.000030*</td>
<td>0.216259</td>
<td>0.001868*</td>
<td>0.855037</td>
</tr>
<tr>
<td>8</td>
<td>14</td>
<td>0.000006*</td>
<td>0.627055</td>
<td>0.000229</td>
<td>0.781897</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td>0.000663*</td>
<td>0.109319</td>
<td>0.006223</td>
<td>0.962761</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
<td>0.000082*</td>
<td>0.256622</td>
<td>0.000027*</td>
<td>0.603895</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>0.000034*</td>
<td>0.002463</td>
<td>0.000557*</td>
<td>0.623104</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>0.000006*</td>
<td>0.316674</td>
<td>0.009471</td>
<td>0.484621</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0.000101*</td>
<td>0.195530</td>
<td>0.217677</td>
<td>0.883870</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>0.000070*</td>
<td>0.075694</td>
<td>0.000149*</td>
<td>0.608107</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>0.000263*</td>
<td>0.464109</td>
<td>0.000635*</td>
<td>0.768470</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>0.001748*</td>
<td>0.217819</td>
<td>0.008337</td>
<td>0.604125</td>
</tr>
<tr>
<td>17</td>
<td>22</td>
<td>0.001328*</td>
<td>0.024979</td>
<td>0.015718</td>
<td>0.658522</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>0.000009*</td>
<td>0.155285</td>
<td>0.017809</td>
<td>0.954988</td>
</tr>
<tr>
<td>19</td>
<td>21</td>
<td>0.012431*</td>
<td>0.514188</td>
<td>0.009205</td>
<td>0.663204</td>
</tr>
<tr>
<td>20</td>
<td>8</td>
<td>0.000079*</td>
<td>0.120546</td>
<td>0.001250*</td>
<td>0.888174</td>
</tr>
<tr>
<td>21</td>
<td>4</td>
<td>0.000511*</td>
<td>0.200590</td>
<td>0.005763</td>
<td>0.583767</td>
</tr>
<tr>
<td>22</td>
<td>13</td>
<td>0.011617*</td>
<td>0.946787</td>
<td>0.019274</td>
<td>0.786773</td>
</tr>
</tbody>
</table>

from a stratified analysis reflected only marginal associations of a CNV feature and did not account for the effect heterogeneity that motivates the application of kernel-based methods, we also applied CONCUR to duplications and deletions separately, and found a very significant association with TG in duplications (p-value < 1 × 10⁻⁸) and a weaker signal in deletions (p-value=0.0313).

2.3 Discussion

We introduce CONCUR to leverage the strength of kernel-based methods to assess the collective effects of rare CNVs on disease risk and incorporate several desired features. First, CONCUR permits the quantification of CNV similarity in an CNVR-free manner, avoiding the need of arbitrarily defining CNVRs as in current practice. Second, CONCUR incorporates both length and dosage information via the cAUC kernel, and is capable of detecting dosage, length and length-dosage
interaction effects. Third, as the technology for detecting smaller CNVs improves, we expect to observe more length variation in CNVs and an increasing need to accommodate length effects in CNV association studies. However, there are shortcomings in the standard kernel choices for handling CNV length. For example, a linear (or polynomial) kernel, which scores length similarity in a multiplicative fashion, cannot always reflect the true level of length similarity between an individual pair: e.g., a pair of CNVs of length 20 would be equally similar to two CNVs with lengths 1kb and 400kb (as $20 \times 20 = 1 \times 400$). The alternative Gaussian kernel as in CKAT would still require a pre-specified scaling factor. CONCUR addresses these issues by using the common AUC of the CN profile curves of an individual pair, quantifying CNV similarity in dosage and length simultaneously. Finally, unlike current kernel methods which require discretized copy numbers, CONCUR is directly applicable to continuous and discrete copy numbers. We implement the CONCUR test in the R package ‘CONCUR’.

CONCUR shares some philosophy with several CNV analysis strategies in the literature. For example, Aguirre et al. [Agu04] characterized the copy number changes in the pancreatic adenocarcinoma genome by detecting the minimum common regions (MCR) of recurrent copy number changes across tumor samples and using MCRs to prioritize genes that might be involved in pancreatic carcinogenesis. Harada et al. [Har08] also examined the minimal overlapping/common regions of frequent CNV activities among pancreatic cancer samples and among normal samples to identify candidate regions that might contain critical oncogenes or tumor suppressor genes. Furthermore, Mei et al. [Mei10] proposed algorithms for identifying common CNV regions across individuals of homogeneous phenotypes for downstream association analysis. Built on similar concepts to these “common regions”, CONCUR quantifies CNV similarity between sample pairs based on the “size” of the common regions as reflected in congruent location and dosage, and provides an association test to evaluate dosage and length effects.

In the analyses performed in this study, we calculated the cAUC using CNV dosage values transformed by the functions $a_{D_{up}}(d) = |d - 2|$ for duplications and 0 otherwise, and $a_{D_{el}}(d) = |d - 2|$ for deletions and 0 otherwise. That is, we used copy number 2 as a reference value, and defined CNV similarity as the overlapping CNV length scaled linearly according to the magnitude of dosage deviation from the reference value. As indicated in the method section, CONCUR can be flexibly extended to accommodate other schemes of quantifying common area by adopting different $a^*(\cdot)$ functions in the calculation of the cAUC, e.g., $a^*(d) = |d - 2|^c$ with $c \neq 1$. Finally, overlapping area may be further weighted by inverse frequencies or according to CNV type (e.g., deletion) when needed, to augment the contribution from overlapping regions from rarer CNVs or from CNVs with more severe impact, respectively.

We note that although both CONCUR and CKAT are designed to capture CNV dosage and length information, the two kernels are constructed based on different philosophies, and each method has sensitivity to certain effect mechanisms. CONCUR first quantifies the similarity in an individual pair within the same genomic locations, and then collapses the similarity information across different locations in a user-specified region (e.g., whole genome, chromosome, or gene set). In contrast,
CKAT treats the user-specified region as a CNVR, and collapses CNV information across different locations by incorporating off-position similarities and/or same-position similarities to obtain a CNVR-level measure of similarity.

The different philosophies of quantifying CNV similarity may also explain the differences in CKAT’s relative performance compared to CCRET here versus in the CKAT paper [Zha16]. Here we observe that CKAT has higher or comparable power compared to CCRET in some scenarios and lower power in other scenarios; however, in the CKAT paper, CKAT outperforms CCRET in the majority of the considered scenarios. This discrepancy is likely due to differences in the assumed effect mechanisms in our approach versus those in the CKAT paper. In the simulation study here, certain genomic regions are selected to be causal. Whereas, in the CKAT paper simulations, causal effects arise from CNVs, not genomic regions. Each individual has either 0 or 1 CNV with randomly generated endpoints; CNVs of the same type (i.e., duplication or deletion) are either all causal or all non-causal, depending on the scenario. Under this design, causal CNVs in different individuals may fall in different genomic locations, yet CNVs of the same type will have similar effects. CKAT powerfully detects these signals because its similarity quantification approach (i.e., based on pairs of CNVs rather than aligned CNV activity in a fixed genomic location) better captures the CNV-driven (as opposed to locus-driven) signal. Whereas, methods that quantify similarity based on aligned positions, such as CCRET and CONCUR, can suffer from power loss under this signal, as exemplified by the performance of CCRET in [Zha16].

2.4 Materials and methods

2.4.1 Ethics statement

In the Swedish Schizophrenia Study, all procedures were approved by ethical committees at the Karolinska Institutet (Dnr No. 04/-449/4 and No. 2015/2081-31/2) and University of North Carolina (No. 04-1465 and No. 18-1938). All subjects provided written informed consent (or legal guardian consent and subject assent). The Taiwan Biobank study was approved by the ethical committee at Taichung Veterans General Hospital (IRB TCVGH No. CE16270B-2). Consent was not obtained because the data were de-identified.

2.4.2 CONCUR method

For individual \( i, i = 1, \cdots, n \), denote \( Y_i \) the phenotype of individual \( i \). Codify the CNV information in matrix \( Z_i \) with dimension \( P_i \times 4 \) as in the standard PLINK format of CNV data, where \( P_i \) is the number of CNVs that individual \( i \) has, and each row of \( Z_i \) records four features of CNV \( p, p = 1, \cdots, P_i \): dosage (denoted as \( d_p \)) chromosome (denoted as \( C hr_p \)), start location (denoted as \( BP1_p \)), and end location (denoted as \( BP2_p \)). The dosage \( d_p \) can be integer, continuous or categorical values. For example, in the Swedish Schizophrenia pathway analysis, an individual might have between 1-88 CNVs, and CNV lengths might range from 100kb up to 7841kb. Let \( X_i = (X_{i1}, \cdots, X_{ir})^T \) be the \( r \)
covariates. Under the kernel machine regression framework, we model the association between phenotypes and CNVs as follows

\[ g(\mu_i) = \beta_0 + X_i^T \beta_X + h(Z_i), \]

(2.2)

where \( \mu_i = E(Y_i | X_i, Z_i) \), the conditional phenotype mean given covariates and CNVs; \( g(\cdot) \) is the canonical link, which transforms the conditional phenotype mean \( \mu_i \) so that the mean is on the same scale of the linear predictors formed by covariates and CNV data. For continuous phenotypes, \( g(\mu_i) = \mu_i \); for binary phenotypes, \( g(\mu_i) = \log([\mu_i/(1-\mu_i)]). \) \( h(Z_i) \) is an unknown smooth function of the variant features characterized by a kernel function \( k(\cdot, \cdot). \)

### 2.4.2.1 Profile curves

The proposed cAUC kernel is built on the concept of a CN profile curve as shown in Fig. 2.1. Consider the genomic location \( x \) from chromosome \( k \) for individual \( i \). Given the CN profile curve, we define the duplication profile curve, \( f_{ik}^{Dup}(x) \), and the deletion profile curve, \( f_{ik}^{Del}(x) \), which recenter and rescale the CN values in CN profile curves through the "dosage transform functions" as described below, and allow us to compute cAUC similarity from duplications and from deletions in a more flexible manner. Specifically, let \( q = 1, \cdots, P_k \) index the CN features \( (d_q, BP1_q, BP2_q) \) occurring on chromosome \( k \) of individual \( i \). Then we construct duplication and deletion profile curves respectively describing duplications and deletions on chromosome \( k \) for individual \( i \) as follows:

\[
\begin{align*}
    f_{ik}^{Dup}(x) &= \sum_{q=1}^{P_k} I(BP1_q \leq x \leq BP2_q) a^{Dup}(d_q) \\
    f_{ik}^{Del}(x) &= \sum_{q=1}^{P_k} I(BP1_q \leq x \leq BP2_q) a^{Del}(d_q)
\end{align*}
\]

(2.3)

(2.4)

where \( x \) is a location on the genome on the same scale as \( BP1_q \) and \( BP2_q \); \( I(\cdot) \) is the indicator function such that \( I(\cdot) = 1 \) if the condition contained within is satisfied and equals 0 if otherwise; and \( a^*(d) \) is a dosage transform function which determines the reference copy number value and controls how different copy number values contribute more or less to similarity in profiles. If an individual has no CNVs in chromosome \( k \), then their duplication and deletion profile curves are identically equal to zero, i.e., \( f_{ik}^{Dup}(x) = f_{ik}^{Del}(x) \equiv 0 \) for all \( x \). Although not explicitly shown, \( f_{ik}^{Dup} \) and \( f_{ik}^{Del} \) are functions of \( Z_i \) as the information of \( d_q, BP1_q, BP2_q \) and chromosome \( k \) for subject \( i \) is obtained from \( Z_i \).

In this study, we designated \( a^{Dup}(d_q) = |d_q - 2| \) if \( d_q \) is from a duplication and 0 otherwise. That is, for a given chromosome \( k \) and individual \( i \), the function \( f_{ik}^{Dup}(x) \) equals the magnitude of the duplication (i.e., number of additional copies compared to the reference copy number 2) for \( x \) inside a duplication and equals 0 otherwise. For deletions, \( a^{Del}(d_q) \) and \( f_{ik}^{Del}(d_q) \) can be obtained in an analogous way.
We propose to quantify the similarity between individuals $i$ and $j$ by comparing $f_{ik}^{\text{Dup}}$ vs. $f_{jk}^{\text{Dup}}$ and $f_{ik}^{\text{Del}}$ vs. $f_{jk}^{\text{Del}}$ over chromosomes $k = 1, \ldots, 22$ using the following kernel function:

$$k_{\text{cAUC}}(Z_i, Z_j) = \sum_{k=1}^{22} \int \left[ \min \left( f_{ik}^{\text{Dup}}(x), f_{jk}^{\text{Dup}}(x) \right) + \min \left( f_{ik}^{\text{Del}}(x), f_{jk}^{\text{Del}}(x) \right) \right] d\mu(x) \quad (2.5)$$

where $\min \left( f_{ik}^{*}(x), f_{jk}^{*}(x) \right)$ captures the minimum of the two functions evaluated at $x$ and $\mu(x)$ is the counting measure. We refer to the kernel function as the cAUC kernel as it computes the minimal common area under the two individuals’ duplication and deletion profile curves. The cAUC kernel matrix $K_{\text{cAUC}}$ is constructed such that its $(i, j)$th element is $k_{\text{cAUC}}(Z_i, Z_j)$. The cAUC kernel is a valid kernel as shown in Appendix A.2.

As an illustrating example, we calculate the cAUC between individuals 1 and 2 from Fig. 2.1. Both individuals have duplication profile curves on chromosome 1 as $f_{11}^{\text{Dup}}(x) = f_{21}^{\text{Dup}}(x) = 0$, since they have no duplications. For individual 1, the deletion profile curve is $f_{11}^{\text{Del}}(x) = |d - 2| = |0 - 2| = 2$ if $x \in [200, 400]$ on chromosome 1 and 0 otherwise; for individual 2, $f_{21}^{\text{Del}}(x) = 2$ if $x \in [100, 500]$, 1 if $x \in [600, 800]$, and 0 otherwise. To compute their cAUC, we characterize the individuals’ curves in 2 genomic regions: (1) $x \in [200, 400]$, in which $f_{11}^{\text{Del}}(x) = f_{21}^{\text{Del}}(x) = 2$; and (2) $x \notin [200, 400]$, in which $\min \left( f_{11}^{\text{Del}}(x), f_{21}^{\text{Del}}(x) \right) = 0$. This discretization allows us to compute the cAUC by multiplying the minimum value of the two curves in each region by the length of the region to obtain $k_{\text{cAUC}}(Z_1, Z_2) = (2 \times 200) + 0 = 400$.

The intuition of the cAUC kernel is to quantify similarity using the length of overlapping CNVs between two individuals, with dosage information of the two overlapping CNVs determining how the overlapping length is scaled. The minimum operator enforces that the overlapping length is scaled by the CNV of smaller magnitude in a pair with different magnitudes. The similarity between CNVs of different types (i.e., duplication vs. deletion) is 0. The similarity between CNVs of the same type depends on the copy number values via the dosage transform function, $a^*(d)$. Legal choices of $a^*(d)$ will upweight the contribution from similar CNVs of greater magnitude in duplication or deletion, which are often more rare and have higher impact. The family of dosage transform functions $a^*(d) = |d - 2|^c$ provides a spectrum of weighting schemes, with $c < 1$ down-weighting and $c > 1$ upweighting the contribution of higher magnitude CNVs. Across copy number data of varying types and varying sample-level characteristics, the $a^*(\cdot)$ dosage transform function allows for flexible scaling of dosage to appropriately customize the cAUC measure of similarity.

### 2.4.2.3 Association test

The association between phenotype and CNVs is examined by testing the hypothesis $H_0 : h(\cdot) = 0$. To do so, we define the vector of subject-specific CNV effects $H = (h(Z_1), \ldots, h(Z_n))$ and treat $H$ as random effects which follow $N(0, \tau K)$, where $\tau \geq 0$ is a variance component and $K$ is a $n \times n$ kernel matrix with its $(i, j)$th entry being $k(Z_i, Z_j)$. Following Liu et al. [Liu07] [Liu08b], testing $H_0 : h(\cdot) = 0$
is equivalent to testing \( \tau = 0 \) under a generalized linear mixed model. As in [Tze15][Zha16], we use a score-based test, which is

\[
T = \frac{1}{2\hat{\sigma}^2}(Y - \hat{\mu}_0)^T K (Y - \hat{\mu}_0)
\]  

(2.6)

for continuous phenotypes, and is

\[
T = \frac{1}{2}(Y - \hat{\mu}_0)^T K (Y - \hat{\mu}_0)
\]  

(2.7)

for binary phenotypes. In the score statistic \( T \), \( Y = (Y_1, \cdots, Y_n)^T \) and \( \sigma^2 \) is the variance of the continuous \( Y \). Define \( \mu = (\mu_1, \cdots, \mu_n)^T \) with \( i \)-th element \( \mu_i = E(Y \mid X_i, Z_i) \); then \( \mu_0 \) is the estimate of \( \mu \) in Eq. 2.2 under \( H_0 : h(\cdot) = 0 \). Specifically, \( \hat{\mu}_0 = \hat{\beta}_0 + X \hat{\beta}_X \) for continuous phenotypes and \( \hat{\mu}_0 = \text{logit}^{-1}(\hat{\beta}_0 + X \hat{\beta}_X) \) for binary phenotypes. The score statistic asymptotically follows a weighted chi-square distribution[Li07][Li08b]. Recently, Chen et al.[Che16] derived the corresponding small-sample distribution, which is used to calculate the p-value in this work.
WEIGHTED FUSED LASSO FOR LOCAL ANALYSIS OF COPY NUMBER VARIANTS

3.1 Introduction

Recent methodological developments have advanced the state of rare copy number variant analysis. Global association tests such as CCRET [Tze15], CKAT [Zha16], and CONCUR [Bru20] have allowed researchers to detect CNV-phenotype associations with power under low-frequency variants as well as multiple sources of etiological heterogeneity. These tests are suitable for rare CNV analysis on the scale of genome, chromosome, or pathway, since they rely crucially on aggregating rare CNV data over large regions. As such, the inference from these methods pertains to aggregate CNV behavior versus the phenotype. In post-hoc analysis, researchers may naively apply these global tests to smaller sub-regions based on observed features of the CNV data or through an exhaustive approach to identify causal sub-regions within the genome, chromosome, or pathway. However, this analysis suffers from lower power due to fewer CNVs to aggregate over and is fraught with multiple-testing issues.

As discussed above, no formal local analysis methods for rare CNVs have been proposed at the time of this work. However, post-hoc and localization methods designed for other types of genetic data, such as SNP, gene expression, and aCGH data, have introduced promising approaches that might be adapted to CNV analysis. In particular, several recent studies have applied lasso methods to the post-hoc effect localization problem. The lasso is a natural choice for this setting due to its tendency to produce sparse solutions and its compatibility with high dimensional data. Moreover, lasso variations with additional or modified ℓ₁ and ℓ₂ penalties can be customized to incorporate
known a priori biological structure in the variables, borrow information across rare features, and produce smooth or grouped effect estimates for greater interpretability. We refer to the work of Zhou et al. [Zho10], who proposed a sparse group lasso method to identify causal SNPs in a pathway. The sparse group lasso approach encourages all or none of the SNPs in a gene to have a non-zero effects with a group penalty, while effectively culling neutral SNPs with the standard lasso penalty. The smoothing provided by the group penalty boosts the power to detect signals from rare SNPs while protecting against noisy or unstable effect estimates that can also arise from non-causal rare SNPs. The $\ell_2$-norm group penalty also prevents the regularization from overly shrinking low frequency variants with moderate effects.

Outside of rare variant analysis, Li and Li [Li08] propose a lasso method which leverages a priori biological knowledge to localize genes and gene subnetworks related to a phenotype. Designed for gene expression data, their method incorporates the known network structure of gene regulatory relationships to encourage smoothing among related coefficients. The lasso framework has also previously been applied to aCGH data in order to call CNVs [Tib08]. Tibshirani et al. identified the fused lasso as being uniquely suited to detect a piecewise constant signal underlying the aCGH data, where each constant region corresponds to a single called CNV. This piecewise constant signal also evokes the copy number profile curve representation of CNVs as proposed in [Bru20]. Drawing inspiration from these 3 lasso methods, we identify the sparse solutions, fused or locally constant estimates, and incorporation of a priori and observed data structure as desirable features for a lasso-based framework for post-hoc CNV analysis.

In this work, we propose a novel method called the weighted sparse fused lasso which combines the sparsity, fusion, and biological interpretability qualities of existing lasso methods with the aim of detecting and localizing signals in post-hoc analysis of rare CNVs. The proposed method introduces a new approach for discretizing CNV activity, which allows us to represent the data in a matrix form compatible with the lasso framework. The use of the lasso moreover gives us access to powerful, computationally efficient optimization algorithms which allow the proposed method to accommodate large $p$ and large $n$ datasets. In this study, we explore several fusion weighting schemes which encourage slightly different fusion behaviors and examine the utility of the weighted sparse fused lasso compared to simpler approaches including the standard lasso and sparse fused lasso. We explore the method’s performance in a variety of simulation settings and apply it to the analysis of CNVs in the region chr13q22-32 and triglyceride levels in data from the Taiwan Biobank.

### 3.2 Methods

#### 3.2.1 Weighted sparse fused lasso

We propose a weighted sparse fused lasso that can be connected to the generalized lasso as shown in the following section. We define $y$ as a $n \times 1$ vector of normal responses, $X$ as a $n \times p$ matrix of unpenalized covariates with effects $\beta_X$, and $Z$ as a $n \times M$ matrix of penalized features with coefficients $\beta$. Then the weighted sparse fused lasso optimization problem is
\[
\min_{\beta} \left\{ \frac{1}{2} \| y - X\beta - Z\beta \|_2^2 + \lambda_1 \| \beta \|_1 + \lambda_2 \sum_{j=1}^{M-1} w_{(j,j+1)} \| \beta_{j+1} - \beta_j \| \right\}, \tag{3.1}
\]

with \( \ell_1 \) norm \( \| v \|_1 = \sum_{k=1}^{K} |v_k| \) and \( \ell_2 \) norm \( \| v \|_2 = \left( \sum_{k=1}^{K} v_k \right)^{1/2} \). The tuning parameters \( \lambda_1 \) and \( \lambda_2 \) control the degree of shrinkage and fusion, respectively. The variable selection, shrinkage, and smoothing effects of the lasso penalties apply only to the coefficients \( \beta \), and the unpenalized coefficients can be estimated at the end of the lasso process using the equations given in 1.

In our CNV analysis setting, we take \( Z \) and \( \beta \) to represent CNV activity and its effects, with covariates such as age, sex, and population structure being captured in \( X \). We propose mapping the functional or CNV-level data to the design matrix \( Z \) using the concept of CNV segments per Brucker et al. [Bru20]. A CNV segment is a genomic region in which the cross section of all \( n \) individuals’ CNV profile curves are constant. We construct \( Z \) to contain segment-wise “area under the profile curve” (AUC) values, a concept also borrowed from [Bru20] which captures dosage and length information. Specifically, the AUC is the length of a CNV times the deviation of its dosage from a baseline value of 2 (\( |d - 2| \)). Thus, the \((i, j)\)th element of \( Z \) reflects individual \( i \)’s duplication or deletion AUC in segment \( j \), and \( \beta_j \) is the effect of the AUC in segment \( j \).

In our simulation study, we consider CNVs falling inside a pathway. CNVs are truncated at pathway gene boundaries and segments are built on the truncated data with the gene location of each segment being noted. We index genes in the pathway with \( g = 1, \ldots, G \) and segments within each gene with \( m = 1, \ldots, M_g \). More generally, we take \( g \) to index CNV regions (CNVRs). For our purposes, a CNVR is a region of CNV activity which is separated from other CNVRs by regions of no activity. CNVRs contain groups of segments such that coefficients within a particular CNVR are not smoothed towards any coefficients in another CNVR but are eligible to be smoothed toward those within the same CNVR; that is, information sharing halts at the boundaries of CNVRs. In this sense a gene may be considered a CNVR, and in the foregoing sections we refer to the constructs indexed by \( g \) as CNVR.

The weight \( w_{(j,j+1)} \) assigned to each pair of \( \beta_j, \beta_{j+1} \) controls how strongly the two estimates are encouraged to fuse with each other. In the standard fused lasso, \( w_{(j,j+1)} = 1 \) for all \( j, j+1 \). In the proposed method we explore differential weighting based on the CNV activity spanning neighboring segments. Also, per the CNVR definition given above, weights for segments that are consecutive in index but which fall in separate CNVR are set to 0.

### 3.2.2 Generalized lasso formulation

We implement the weighted sparse fused lasso using its representation as a generalized lasso. The generalized lasso problem [Tib11] for Gaussian responses takes the form

\[
\frac{1}{2} \| y - X\beta - Z\beta \|_2^2 + \lambda \| D\beta \|_1, \tag{3.2}
\]

where matrix \( D \in \mathbb{R}^{r \times p} \) governs the magnitude and nature of penalties applied to individual coeffi-
cients and combinations of coefficients. For a CNVR $g$ with $M_g$ segments, we define the weighted fusion penalty submatrix $D_g$ as

$$D_g = \begin{pmatrix}
-w_{1,2} & w_{1,2} & 0 & \cdots & 0 & 0 \\
0 & -w_{2,3} & w_{2,3} & \cdots & 0 & 0 \\
 & & & \ddots & \ddots & 0 \\
0 & 0 & 0 & \cdots & -w_{M_g-1,M_g} & w_{M_g-1,M_g}
\end{pmatrix},$$

(3.3)

where $w_{M_g-1,M_g}$ corresponds to a non-negative weight for the fusion of coefficients $\beta_{M_g-1}$ and $\beta_{M_g}$ in CNVR $g$, for $g = 1 \cdots G$. Then we define $D_0$ as the block matrix

$$D_0 = \begin{pmatrix}
D_1 & 0 & \cdots & 0 \\
0 & D_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & 0 \\
0 & 0 & \cdots & D_G
\end{pmatrix},$$

(3.4)

which is dimension $(M-G) \times M$ where $M = \sum_{g=1}^{G} M_g$. Finally, we define $D = \left[D_0^T, \frac{\lambda_1}{\lambda} I_{p \times p}\right]^T$ where $\lambda > 0, \lambda_1 \geq 0$. Plugging this into Eq. 3.2 gives us

$$\frac{1}{2} \| y - X\beta_X - Z\beta \|_2^2 + \lambda_1 \| I_{p \times p} \beta \|_1 + \lambda \| D_0 \beta \|_1,$$

(3.5)

which can easily be shown to correspond to Eq. 3.1 with $\lambda_2 = \lambda$.

### 3.2.3 Choice of weights

When we apply the proposed method to CNV analysis, we construct fusion weights $w_{(j,j+1)}$ based on the CNV activity spanning pairs of segments. A well designed weighting scheme will fuse effects from rare variants with those from nearby common variant effects when the signal is indeed constant among all of them. In addition, our choice of weights should allow local heterogeneity to emerge if comes from a large enough signal or from regions with a large minor allele frequency (MAF). In general, effects from low-MAF segments should be more strongly fused with neighbors than those from segments containing common variants. Below we discuss two weighting schemes which we evaluate in the simulations and real data analysis.

We first propose the inverse frequency (IF) weighting scheme which fulfills the 3 desired characteristics outlined above. We define $n_{(j,j+1)}^{\text{CNV}}$ as the number of CNVs spanning consecutive segments $j$ and $j+1$. Then the IF weight between two neighboring segments is given by

$$w_{(j,j+1)} = \begin{cases}
\left( \frac{k}{n_{(j,j+1)}^{\text{CNV}}} \right)^T, & \text{if } n_{(j,j+1)}^{\text{CNV}} \geq 1 \\
0, & \text{if } n_{(j,j+1)}^{\text{CNV}} = 0,
\end{cases}$$

(3.6)

where $k$ is a scaling factor $\geq 0$ that transforms the frequencies $n_{(j,j+1)}^{\text{CNV}}$ to proportions which are then
We consider two neighboring segments falling in the same CNVR with vector representation $z$ where $r$, which ranges from 0 to 1 since $k = k_1$ and close to 0. We refer to the weighted sparse fused lasso method using the IF weighting scheme with $k = k$ and $r = r$ as $D_{IF}(k, r)$.

We additionally propose a weighting scheme which is based on cosine similarity, a natural measure of correlation since it takes value 0 for orthogonal vectors and value 1 for identical vectors. We consider two neighboring segments falling in the same CNVR with vector representation $z_j$ and $z_{j+1}$, both of which $\in \mathbb{R}^n$. These vectors contain the AUC of $n$ individuals in the $j$th and $(j+1)$th segments. Individuals with no CNV activity in these segments have values of 0. Then the cosine similarity between these two vectors is

$$s(z_j, z_{j+1}) = \frac{z_j^T z_{j+1}}{\|z_j\|_2 \|z_{j+1}\|_2} \quad (3.7)$$

which ranges from 0 to 1 since $Z$ contains non-negative values, with 0 indicating no CNVs overlapping the two segments and 1 indicating CNV activity among the exact same individuals in each segment. We note that when AUC values are derived from categorical dosage (1 for deletions, 3 for duplications), each element of a column $z_j$ will be either 0 or $l_j$ where $l_j$ is the length of segment $j$. The constants $l_j$ and $l_{j+1}$ will cancel out of Eq. 3.7 since they can be factored out of both the numerator and denominator. For integer or continuous dosage values, we compute Eq. 3.7 based on indicator vectors $z_j^*$ and $z_{j+1}^*$ where $z_{ij}^* = I(z_{ij} \neq 0)$, understanding $I(\cdot)$ to equal 1 if the condition contained within is true and 0 otherwise.

The final cosine similarity (CS) weighting scheme for two consecutive segments in a fixed CNVR is given by

$$u_{(j,j+1)} = k \ast \{s(z_j, z_{j+1})\}^r \quad (3.8)$$

where $r$ again governs separation between high and low similarity values and $k$ appropriately scales the similarity measure. We refer to the weighted sparse fused lasso method using the CS weighting scheme with $k = k$ and $r = r$ as $D_{CS}(k, r)$.

To illustrate both weighting schemes, we consider the example of 3 individuals’ CNV activity shown in Fig. 3.1. The activity is captured by 4 CNV segments, indexed by $j = 1, \cdots, 4$. The thick black lines indicate CNVs and the vertical bars mark the endpoints of the CNV segments constructed on the CNV profiles. In this example we have $n_{1,2}^{CNV} = 2$, $n_{2,3}^{CNV} = 2$, and $n_{3,4}^{CNV} = 1$. For $k = 3$ and $r = 2$, this gives us the IF weights of $u_{(1,2)} = (3/2)^2$, $u_{(2,3)} = (3/2)^2$, and $u_{(3,4)} = (3/1)^2$. For the CS method, we use indicator vectors $z_j^*$ for simplicity, with $z_1^* = (0, 1, 1)^T$, $z_2^* = (1, 1, 1)^T$, $z_3^* = (1, 0, 1)^T$, and $z_4^* = (0, 0, 1)^T$. Again taking $k = 3$ and $r = 2$ yields the CS weights of $u_{(1,2)} = 3 \ast (2/\sqrt{3})^2$, $u_{(2,3)} = 3 \ast (2/\sqrt{3})^2$, and $u_{(3,4)} = 3 \ast (1/\sqrt{3})^2$. 

53
In the simulations and real data analysis, we explore two implementations of each weighting approach. Specifically, we set $k$ as the number of individuals with any CNV activity of the relevant type (duplication or deletion) in the analysis region. We evaluated $D_{IF}(k, 0)$, which corresponds to a block fused lasso, and $D_{IF}(k, 2)$ as well as $D_{CS}(k, 1)$ and $D_{CS}(k, 2)$. Finally, as comparator methods we implemented the standard lasso with no fusion and a sensitivity weighted scheme using $w_{(j,j+1)} = n_{CNV(j,j+1)}^c$, which we denote as the naive weighted sparse fused method.

### 3.2.4 Model selection with $\lambda$

To make inference, we evaluate the lasso fit at values of $\lambda$ which balance goodness of fit against parsimony and error control. The AIC and BIC functions are common choices for selecting an optimal model fit in lasso problems which we adopt in this analysis. In addition, we consider the false selection rate (FSR) function of Wu et al. [Wu07] which uses a data augmentation approach. An $n \times 2p$ augmented dataset is constructed in which the first $p$ variables are a duplicate of all $p$ variables in the original data, and the subsequent $p$ variables are “phony variables” generated to be independent of the response. Permuting the $n$ rows of the original data is one method for generating the phony variables which results in variables with the same marginal distributions as the original variables. Finally, we repeat this process to generate $B$ unique augmented datasets.

Based on the selection of phony variables in each augmented data model fitting, we can estimate the true false selection curve which represents the expected number of false positives out of all selected variables as $\lambda$ varies. The estimated FSR curve is given by

$$\hat{\gamma}_{RE}(\lambda) = \frac{\hat{k}_U(\lambda)\overline{U}_p(\lambda)/k_p}{\overline{S}_p(\lambda) - \overline{U}_p(\lambda)}, \quad (3.9)$$

where $\hat{k}_U(\lambda)$ is the number of unselected variables in the unaugmented data model fit at $\lambda$, $\overline{U}_p(\lambda)$ is the mean number of of phony variables selected over the $B$ augmented data model fittings at $\lambda$, and $\overline{S}_p(\lambda)$ is the total number of selected variables in the augmented data model at $\lambda$ averaged again.
over $B$ augmented data model fittings. Then an optimal $\lambda$ is the smallest value (i.e., the most liberal model) which controls the FSR at the predefined error rate $\gamma_0$:

$$\lambda_{FSR} = \inf_{\lambda} \{ \lambda : \hat{\gamma}_{RE}(\lambda) \leq \gamma_0 \}. \quad (3.10)$$

In summary, we implement the FSR-control procedure by first applying the lasso method to the original unaugmented data, from which we obtain $\hat{k}_U(\lambda)$. Then we perform the lasso on $B$ augmented datasets to calculate $\hat{U}_P$ and $\hat{S}_P(\lambda)$. Given the resulting FSR function $\hat{\gamma}_{RE}(\lambda)$, we extract the original model fit at $\lambda_{FSR}$.

### 3.2.5 Connection to CNV profile curves

Our presentation of the weighted sparse fused lasso up to this point has been methods-driven rather than problem-driven. Here we return to the CN profile curve representation of CNVs that we introduced in Chapter 2 which in fact motivates the weighted sparse fused lasso framework. To recapitulate, in [Bru20] we defined the duplication and deletion profile curves as functional representations of individuals’ CNV activity across the genome or subregions. Letting $q = 1, \cdots, P_{i,k}$ index the dosage, starting location, and ending location ($d_q, BP_{1,q}, BP_{2,q}$) features of the $P_{i,k}$ CNVs that individual $i$ has in chromosome $k$, we define the duplication and deletion profile curves as

$$f^{Dup}_{i,k}(s) = \sum_{q=1}^{P_{i,k}} I\{BP_{1,q} \leq s \leq BP_{2,q}\}|d_q - 2| \quad (3.11)$$

$$f^{Del}_{i,k}(s) = \sum_{q=1}^{P_{i,k}} I\{BP_{1,q} \leq s \leq BP_{2,q}\}|d_q - 2|, \quad (3.12)$$

where $s$ is a location on the genome on the same scale and space as $BP_{1,q}$ and $BP_{2,q}$.

In this work, we propose that the features of this continuous function may be associated with a phenotype through an effect curve $\beta(s)$. The curve operates on the same space as the profile curves, call it $S$, and its integrated product with the CN profile curve can flexibly capture CNV effects on a phenotype:

$$\int_S \beta(s)f^*_i(s) \, ds \quad (3.13)$$

for an individual $i$ with duplication or deletion profile curve $f^*_i(s)$. Then, assuming a Gaussian phenotype $y$, we have

$$E(y_i) = x_i^T \beta_X + \int_S \beta(s)f^*_i(s) \, ds \quad (3.14)$$

for individuals $i = 1, \cdots, n$, where the function $\beta(s)$ is the same across all individuals, and $x_i$ and $\beta_X$ are covariates and their effects. With a good estimate of $\beta(s)$, we accomplish the dual aims of our current work: identifying causal regions (i.e., $|\beta(s)| > 0$) and estimating effects in those regions.

We expect the true $\beta(s)$ to be non-zero in contiguous genomic regions and for the signal to be relatively constant over each region. If true, this suggests that $\beta(s)$ can be well approximated by a
piecewise constant function. Given a particular CNV dataset, we consider a particular approximation \( \tilde{\beta}(s) \) which is constant within CNV segments. Taking \( S_j \) as the genomic locations in segment \( j \) for \( j = 1, \cdots, M \), we have

\[
\int_S \beta(s) f_i^*(s) \, ds \approx \int_S \tilde{\beta}(s) f_i^*(s) \, ds \\
= \sum_{j=1}^M \int_{S_j} \tilde{\beta}(s) f_i^*(s) \, ds \\
= \sum_{j=1}^M \tilde{\beta}_j \int_{S_j} f_i^*(s) \, ds
\]

(3.15)

where \( \tilde{\beta}_j = \{ \tilde{\beta}(s) : s \in S_j \} \) for \( j = 1, \cdots, M \). The integral term in Eq. 3.15 is equivalent to the “area under the curve” as defined in [Bru20] which can be discretized into segment-wise AUCs. Thus, we can approximate the true model in Eq. 3.14 with

\[
E(y_i) = x_i^T \beta x + \sum_{j=1}^M \tilde{\beta}_j z_{ij}^*
\]

for \( i = 1, \cdots, n \) where \( z_{ij}^* \) is the duplication AUC of individual \( i \) in CNV segment \( j \). This matrix representation crucially allows us to estimate the unknown function \( \beta(s) \) using computationally efficient lasso methods. We apply a combination of lasso and weighted fused penalties to encourage the lasso solution to share key qualities with the true curve: sparseness and locally constant signals.

### 3.3 Simulations

To explore the performance of the proposed method in simulations we generated phenotypes using real CNV data from the Taiwan Biobank (TWB). The details of the dataset are given in [Bru20]. Briefly, the data includes the CNVs of 11,664 individuals whose CNVs have minor allele frequencies (MAF) ranging from < 0.001% to 20%. Dosage values were categorical with duplication, normal, and deletion CNVs coded as 3, 2, and 1 respectively. Using this real data and a simulated covariate, we generated phenotypes for \( n = 5000 \) individuals as detailed in the section below.

In the simulations, we generated phenotypes from duplications falling in the hsa01040 lipid metabolism pathway from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kan16]. Duplications in the data interrupted 23 genes in this pathway and were truncated at the boundaries of the genes. 437 individuals had at least one duplication in this pathway. CNV MAFs in the data ranged from 0.1-2%.

#### 3.3.1 Simulation model

We constructed the matrix \( Z \) capturing AUC data for \( n \) individuals with CNV activity in \( M = \sum_{g=1}^G M_g \) segments. The segments were constructed as in [Bru20]. The endpoints of the segments corre-
sponded to locations where a CNV in any one of the samples begins or ends, resulting in segments that all contain at least one CNV. The cross-section of CNV dosage values across individuals was constant within each segment, and the length of the segments varied depending on the nature of the TWB data. For the purpose of constructing the fusion penalty matrices, we noted the indices of “consecutive” segments separated by CNVR boundaries or regions of no CNV activity.

We generated phenotypes as a function of the CNV AUC data as well as a simulated binary sex covariate. In addition to the individuals with CNV data in the analysis regions, we added “control” individuals with no CNV activity to bring the total number of individuals to \( n = 5000 \). This mimicked the nature of real data CNV analysis, in which typically all individuals with phenotypes and covariates are included in the analysis regardless of whether they have CNV activity in the test region.

We simulated Gaussian phenotypes from the model

\[
Y_i = \beta_0 + \beta_X x_i + \sum_{j=1}^{M} \beta_{j}^{D_{up}} z_{ij}^{D_{up}} + e_i, \quad (3.17)
\]

where \( z_{ij} \) is the \((i, j)\) entry of matrix \( Z^{D_{up}} \), \( i = 1, \cdots, n \) indexes individuals, and \( j = 1, \cdots, M \) indexes CNV segments. \( e_i \) is a randomly generated residual \( \sim N(0, 1) \). We generated independent binary sex covariates \( x_i \) from a Bernoulli(0.5) distribution. In causal segments \( \beta^{D_{up}}_j \) was the effect of the AUC in segment \( j \) on the continuous response and was set to 0 in non-causal segments. The intercept \( \beta_0 \) was set to 20 and the covariate effect \( \beta_X \) was fixed at 1. We discuss the choice of CNV effects \( \beta = (\beta_1, \cdots, \beta_M)^T \) in the following section.

### 3.3.2 CNV effects \( \beta \)

We selected 3 genes to include causal effects: ELOVL6, SCP5, and HSD17B12. These genes were chosen since they contained at least 10 segments each and captured a wide range of MAFs (see Fig. ??). We set assigned 10 segments in each gene to have fixed (constant) causal effects, with the elements of \( \beta = (\beta_1, \cdots, \beta_M)^T \) corresponding to these segments being nonzero. We generated signals from each region according to \( c \ast |\log_{10}(M AF)| \) as in [IL13], where \( c \) is a constant and the MAF is the maximum MAF among the 10 segments in each causal region. In practice we set \( c = 0.5 \).

### 3.3.3 Implementation

In each simulation scenario, we implemented 6 methods total: the inverse frequency methods \( D_{IF}(k, 0) \) and \( D_{IF}(k, 2) \); the cosine similarity methods \( D_{CS}(k, 1) \) and \( D_{CS}(k, 2) \); and the standard lasso and naive weighting strategy based on CNV overlap. We implemented the fused lasso methods using the \texttt{genlasso} R package [Arn19] and the lasso using the \texttt{lars} package [Has13]. Both functions returned the complete solution paths. We selected optimal fits using the BIC, AIC, and the FSR-control approach of [Wu07] and we denote the minimizing \( \lambda \) values as \( \lambda_{AIC} \), \( \lambda_{BIC} \), and \( \lambda_{FSR} \). The degrees of freedom were estimated using the results from Tibshirani et al. [Tib12]. To reduce the scope of the \( \lambda \) optimization, we set \( \lambda_1 = \gamma \lambda \) where \( \gamma \geq 0 \) is a fixed penalty scaling parameter. Recall
that \( \lambda_1 \) governs the lasso penalty and \( \lambda_2 \) controls the fusion penalty. Thus, \( \gamma > 1 \) corresponds to relatively greater shrinkage and less fusion, and \( \gamma < 1 \) encourages a smoother but less sparse solution. In the simulations we varied \( \gamma \in \{0.5, 1, 2, 5, 10\} \).

In each simulation setting we generated phenotypes and fitted the lasso models in \( R = 100 \) independent replications. Phenotypes were simulated using Eq. 3.17 with random errors \( e \) varying across replications but with the covariate matrix \( X \) and CNV data \( Z_{D\text{up}} \) remaining fixed. This design was necessary since random or bootstrapped draws from the 5000 individuals would result in samples with no data in some segments, rendering their effects inestimable.

### 3.3.4 Performance metrics

From the 100 replications, we computed summary statistics for the coefficient estimates as well as performance metrics including true positive rate (TPR), false positive rate (FPR), false discovery rate (FDR), and mean squared error (MSE). We define \( \mathcal{N} \) as the set of indices corresponding to segments with true null effects, such that \( j \in \mathcal{N} \) indicates the true \( \beta_j = 0 \). The cardinality of this set \( |\mathcal{N}| \) is the number of null segments, and we also note the number of causal segments given by \( |\mathcal{N}^c| \). Then we have

\[
\text{TPR} = \frac{\sum_{r=1}^{R} \sum_{j=1}^{M} I(|\hat{\beta}_{rj}| > \epsilon \& |\beta_{rj}| > 0)}{|\mathcal{N}^c| * R}, \quad (3.18)
\]

\[
\text{FPR} = \frac{\sum_{r=1}^{R} \sum_{j=1}^{M} I(|\hat{\beta}_{rj}| > \epsilon \& |\beta_{rj}| = 0)}{|\mathcal{N}| * R}, \quad (3.19)
\]

\[
\text{FDR} = \frac{\sum_{r=1}^{R} \sum_{j=1}^{M} I(|\hat{\beta}_{rj}| > \epsilon \& |\beta_{rj}| = 0)}{\sum_{r=1}^{R} \sum_{j=1}^{M} I(|\hat{\beta}_{rj}| > 0)} \quad (3.20)
\]

\[
\text{MSE} = \frac{\sum_{r=1}^{R} \sum_{j=1}^{M} (\beta_{rj} - \hat{\beta}_{rj})^2}{M \times R}, \quad (3.21)
\]

where \( r \) indexes the replications, \( j \) indexes the coefficients in \( \beta \), and the denominators reflect the number of truly causal effects (TPR) and truly null effects (FPR).

In both the real data analysis and simulations we also computed the TPR and FPR for the fused lasso methods using a threshold \( \epsilon > 0 \) such that \( |\hat{\beta}_{rj}| > \epsilon \) was taken as a positive and \( |\hat{\beta}_{rj}| \leq \epsilon \) was interpreted as a negative. The thresholded TPR and FPR are given by

\[
\text{TPR}_\epsilon = \frac{\sum_{r=1}^{R} \sum_{j=1}^{M} I(|\hat{\beta}_{rj}| > \epsilon \& |\beta_{rj}| > 0)}{|\mathcal{N}^c| * R}, \quad (3.22)
\]

\[
\text{FPR}_\epsilon = \frac{\sum_{r=1}^{R} \sum_{j=1}^{M} I(|\hat{\beta}_{rj}| > \epsilon \& |\beta_{rj}| = 0)}{|\mathcal{N}| * R}, \quad (3.23)
\]

\[
\text{FDR}_\epsilon = \frac{\sum_{r=1}^{R} \sum_{j=1}^{M} I(|\hat{\beta}_{rj}| > \epsilon \& |\beta_{rj}| = 0)}{\sum_{r=1}^{R} \sum_{j=1}^{M} I(|\hat{\beta}_{rj}| > \epsilon)} \quad (3.24)
\]
3.4 Results

3.4.1 Simulations

In our first approach, we selected $\lambda$ using the BIC. The performance metrics for the 6 methods over 100 replications evaluated at $\lambda_{BIC}$ are given in Fig. 3.2 below. We vary $\gamma$ along the x-axis to illustrate how the fused lasso models approach the lasso as $\gamma$ increases and to examine trends in performance that might suggest an optimal value of $\gamma$ that preserves TPR and controls FPR. We observe that the fused lasso methods are well powered to detect the 3 causal genes, and they show substantial gains in power over the standard lasso due to their capture of rare variant effects through fusion. However, this sensitivity appears to come at the cost of large FPR and FDR. As $\gamma$ increases, the fused methods maintain an advantage in TPR while their FPR and FDR become comparable to that of the lasso.

The $D_{IF}(k,2)$, $D_{CS}(k,1)$, and $D_{CS}(k,2)$ decay towards the lasso more gradually as $\gamma$ due to the larger weights in their $D$ matrices. The smaller weights in the $D_{IF}(k,0)$ and naive method $D$ matrices lead these methods’ fusion and lasso forces to equalize at smaller values of $\gamma$.

![Fig. 3.2](image)

**Figure 3.2 Performance metrics for model selection using BIC across 100 replications for 6 lasso methods.** Metrics are reported at $\gamma = (0.5, 1, 2, 5, 10)$, where $\gamma$ is the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.

We also applied the AIC for model selection. Compared to the BIC, the AIC selected more satu-
rated models, which boosted the TPR among all methods. Not surprisingly, this was accompanied by an increase in the FPR and FDR. The ranking of the methods is similar to that in the BIC model selection results. The $D_{IF}(k, 2)$, $D_{CS}(k, 1)$, and $D_{CS}(k, 2)$ methods have the largest TPR as well as the highest false positive and false discovery rates over most $\gamma$.

![Graphs showing TPR, FPR, FDR, and MSE for different methods](image)

**Figure 3.3 Performance metrics for model selection using AIC across 100 replications for 6 lasso methods.** Metrics are reported at $\gamma = (0.5, 1, 2, 5, 10)$, where $\gamma$ is the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.

Under examination, we determined that the uncontrolled FPR and FDR of both selection approaches were due to very small non-zero estimates of effects in null regions. Fig. 3.4 illustrates this behavior under the $D_{IF}(k, 2)$ method evaluated at $\lambda_{BIC}$ and taking $\gamma = 1$. It appeared that the BIC and AIC were selecting saturated models with low bias which corresponded to small $\lambda$ values. At smaller $\lambda$, the weighted fusion forces seemingly overtook the shrinkage forces from the lasso penalty, resulting in fused sections with very small non-zero estimates in the non-causal regions. However, in most model fits we observed a degree of separation between the causal and null estimates, which suggested that a well chosen threshold for classifying estimates as negatives or positives could improve the FPR and FDR.
Before proceeding, we decided to adopt an ad-hoc thresholding approach which employed data augmentation in a similar fashion to the FSR method. Our approach included 2 main steps within each replication: 1) we obtained the unaugmented data model fit at $\lambda_{BIC}$ or $\lambda_{AIC}$; then 2) using a model fit on an augmented dataset, we set thresholds based on the characteristics of the phony variable estimates. To construct an augmented dataset with $n$ observations and $2p$ predictors, we let the first $p$ predictors be the $p$ predictors from the original data and generated $p$ phony variables by permuting the rows of the original data, stratified by presence/absence of CNV activity. Then, we fit a model on the augmented data and extracted the coefficient estimates at $\lambda_{BIC}$ and $\lambda_{AIC}$. Finally, we set a threshold as the minimum value $\epsilon$ that controlled the FPR (Eq. 3.24) among the phony variables at 5%. We derived unique method-specific thresholds in each replication, and calculated the TPR, FPR, and FDR according to Eq. 3.24.

The performance of the thresholding model fits is shown in Fig. 3.5 for the BIC and in Fig. 3.6 for the AIC. The FPR and FDR are much reduced from the rates without thresholding, while the TPR remains high for the $D_{IF}(k, 2)$, $D_{CS}(k, 1)$, and $D_{CS}(k, 2)$ methods.
Figure 3.5 Performance metrics for model selection using BIC with thresholding across 100 replications for 6 lasso methods. Metrics are reported at $\gamma = (0.5, 1, 2, 5, 10)$, where $\gamma$ is the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.
Figure 3.6 Performance metrics for model selection using AIC with thresholding across 100 replications for 6 lasso methods. Metrics are reported at $\gamma = (0.5, 1, 2, 5, 10)$, where $\gamma$ is the lasso penalty scaling factor. Note that $y$-axis range differs across the 4 panels.
While the thresholding approach was appealing in its simplicity and ease of implementation, it was an ad-hoc approach which could not guarantee error control in principle. In the following simulation study, we performed model selection using the FSR-control method of Wu et al. described in Section 3.2.4 above. We generated $B = 10$ augmented datasets which were constructed from the $p$ original predictors and $p$ phony variables generated as a random row-permutation of the original data. We observed that there was minimal variation in model fits across replications and decided to perform the procedure on just one replication, from which we obtained the curve estimate $\hat{\gamma}_{RE}(\lambda)$ and the $\lambda_{FSR}$ that controls the FSR at the desired level. In 100 new replications, we evaluated unaugmented data model fits at their method-specific $\lambda_{FSR}$. In preliminary implementations of this approach we noted that FSR control at 0.05 resulted extremely conservative model fits among the fused methods, in many settings resulting in completely null models, i.e., $\hat{\beta}(\lambda_{FSR}) = 0$. Moving forward we evaluated the methods’ performance under FSR control at 0.1, 0.15, 0.2, and 0.25.

In Fig. 3.7 below, we report the performance given $\gamma = 1$ and vary the FSR control level along x-axis. The performance across varying FSR control levels under $\gamma \in \{0.5, 2, 5, 10\}$ can be found in Appendix B.1. The FSR is controlled at close to the nominal level for all methods.

![Figure 3.7 Performance metrics for model selection using FSR control across 100 replications for 6 lasso methods, at $\gamma = 1$. Metrics are reported at FSR control levels of 0.1, 0.15, 0.2, and 0.25. $\gamma$ is the value of the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.](image)

Finally, we compared the model selection approaches by the features of their signal estimates.
We adopted a standard visualization of fused lasso fits which shows the estimated $\hat{\beta}(\lambda)$ on the y-axis with the x-axis indexing the ordered elements of $\hat{\beta}$, which we call a skyline plot. Within each replication, we evaluated the models at their respective optimal $\lambda$ values and stored the estimated coefficients. We took the average of these estimates over 100 replications and plot these means (red) against the true signal (black) in the skyline plots below.

We include the skyline plots for the BIC-selection and AIC-selection approaches in Fig. B.5 and Fig. B.6. Below we show the average estimates under FSR controlled at 0.25 and $\gamma = 1$. Under the criterion-based approaches, we observe that the fused methods’ estimates accurately capture the true signal with minimal bias. The lasso method’s variable selection is good but its estimates are quite noisy. Although not visible on this scale, many of the estimates in the null regions are not identically equal to zero (as shown above in Fig. 3.4).

In contrast, the FSR-controlling approach tends to select larger $\lambda$ values that result in sparser models. The model estimates are also biased towards zero. In these plots, estimates that appear to be zero are in fact zero, not slightly non-zero fused groups as in the more saturated BIC- and AIC-selected models. Among all approaches for selecting $\lambda$, the lasso methods that stand out as accurately capturing the signal are $D_{IF}(k, 2)$, $D_{CS}(k, 1)$, and $D_{CS}(k, 2)$.
3.5 Real Data Analysis

In the real data analysis we analyzed the triglyceride (TG) levels of 11,664 individuals from the TWB. In particular, we applied all 6 methods in two separate analyses of duplications and deletions in the region chr13q22 (chr13:73300000-101700000bp in the hg19 build). We adjusted for age, sex, BMI, and 10 principal components capturing population substructure as covariates. This data was previously analyzed in [Bru20], in which TG levels were found to be associated with CNV activity in all chromosomes through the CONCUR global association test. The analysis also identified the subregion of chromosome 13q22-32 as being associated with TG levels. In particular, the CONCUR p-value for an association between TG levels and deletions is 0.00083 and 0.00572 for duplications. The presence of these associations on the macro-level motivates the post-hoc analysis of the region that we perform here.

In chr13q22-32 there were 450 deletions among 299 individuals and 1206 duplications among
1115 individuals. The deletions were broken down into 175 segments and the duplications into 594 segments. We applied all 6 lasso methods setting $\gamma = 1$ and made inference using the thresholding and FSR-control approaches. Finally, based on instability observed in a preliminary analysis, we performed the following analyses on CNV data with endpoints rounded to the nearest 10kb and with ultra-rare CNVRs ($< 0.05\%$ MAF) removed. We fixed $\gamma = 1$ for all the fused lasso methods in the real data analysis, since this value correlates with high power and low false positive and selection rates in both the thresholding and FSR-control approaches.

### 3.5.1 Deletions in chr13q22-32

The results of the thresholded BIC and AIC fits are shown in Section B.3. Applied to the deletions, the BIC returns a sparse model so that the thresholding is not strictly needed. The $D_{IF}(k, 0)$, $D_{IF}(k, 2)$, and naive methods detect a signal in the region of 101,140-101,429kb on chromosome 13, and the naive method additionally detects a signal in 98,570-98,979kb. These region of 101,140-101,429kb includes the genes PCCA and TMTC4, and the region of 98,570-98,979kb contains the genes IPO5 and FARP1.

In the AIC model fits, we observe dramatically different behavior from the AIC simulations analyses. The estimates are extremely large in magnitude and, unlike in the simulations, the fused lasso methods display little to no fusion. This behavior results from the relatively small $\lambda$'s that the AIC selects in this analysis. The AIC model fits detect many more signals than the BIC analysis, but we suspect many of them are false positives and we demurred to mapping them back to their genomic locations. This analysis illustrates the risk of the AIC returning an overly saturated model, which we resolved using thresholding in the simulations; however, that ad-hoc approach fails here, and we do not make any inference based on these results.

For the FSR model selection approach, we analyzed deletions controlling the FSR at 0.2 and 0.25 with $B = 10$. We report the model fits under FSR controlled at 0.2 in Section B.3 and under FSR controlled at 0.25 in Fig. 3.9 below. The regions selected here are similar to those selected in the BIC model fits. The $D_{IF}(k, 0)$ method selects the region 101,140-101,349kb which contains the genes PCCA and TMTC4, which the naive method also captures (101,130-101,429kb) as well as the region 98,570-98,979kb containing genes IPO5 and FARP1. The $D_{IF}(k, 2)$ method flags a slightly larger region that covers the entire PCCA gene and TMTC4 (100,570-101,429kb). Finally, the lasso selects the region 99,980-100,079kb overlapping the UBAC2 gene and 101,140-101,299kb again including part of the PCCA gene.

### 3.5.2 Duplications in chr13q22-32

The results of the duplications analysis using thresholded BIC and AIC fits are shown in Section B.4. In the BIC model fits, we observe entirely null model fits that were not seen in the simulations but which are not altogether unexpected. We note that the large $n$ in the real data analysis leads to upweighting of the increasing degrees of freedom term in the BIC. This term grows as $\lambda$ decreases and in this case must overpower the log-likelihood term which shrinks as $\lambda$ approaches 0. Indeed,
when we inspected the $BIC(\lambda)$ curve, its minimizer was at $\lambda = \infty$, the null model, since the curve was roughly monotonically increasing as $\lambda$ shrunk towards 0.

In contrast to the deletions analysis, the thresholded model fits selected by the AIC are quite interpretable and can be mapped to genomic regions that plausibly contain a signal. The $D_{IF}(k,2)$, $D_{CS}(k,1)$, and $D_{CS}(k,2)$ methods all detect the region 96,120-98,609kb which contains the genes UGGT2 and MBNL2. The $D_{IF}(k,2)$ includes the tail end of the ABCC4 gene and the $D_{CS}(k,2)$ method selects the entire ABCC4 gene. This gene is well studied and the literature suggests an association with triglycerides [Hal13; Sch17]. The signals detected by the $D_{IF}(k,0)$, naive, and lasso methods are more difficult to interpret. The $D_{IF}(k,0)$ model selects the regions 97,180-97,419kb and 101,800-103,409kb which contain several genes but have very little CNV activity. The naive method selects the region 88,000-89,109kb which includes a single long CNV which was not discarded during our data cleaning due to its small overlap with a CNV region with a small MAF that met our threshold. It also selects the region 97,110-98,249kb which contains around a dozen CNVs and covers the gene MBNL2. Finally, the lasso detects a signal in the region 92,440-92,699kb, containing the GPC5 gene;
97,180-97,419kb, which covers the HS6ST3 gene; and 98,060-98,199kb which includes no genes but is spanned by several long CNVs.

The results of the FSR-controlled model fits are shown in Section B.4 (controlled at 0.2) and below in Fig. 3.10 (controlled at 0.25). The $D_{IF}(k, 2)$ method detect a signal from the same region as its AIC selected analogue. The $D_{CS}(k, 1)$ method estimates a less sparse signal, selecting the regions 78,650-80,809kb containing the gene RBM26; 83,320-84,319kb which includes no genes and is overlapping by 6 CNVs; 93,100-94,239kb covering the GPC5 gene and is spanned by only 5 CNVs; and 95,650-98,609kb which is the same region identified by the AIC-selected model. The $D_{CS}(k, 2)$ method selects the region 96,360-98,609kb. The naive method selects predictors covering the region 97,110-98,249kb which includes many long (>1Mb) CNVs and covers the MBNL2 gene, and the region 101,800-103,959kb which has no gene activity but has 6 CNVs that are roughly 1.5Mb long. The lasso selects roughly the same region: 101,800-103,409kb.

![Figure 3.10 Model fits for chr13q duplications analysis at $\lambda_{FSR}$ for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.25. The estimates $\hat{\beta}(\lambda_{FSR})$ are plotted in red on the y-axis. The coefficients are ordered according as shown on the x-axis.](image_url)
3.6 Discussion

In this work we developed a method for identifying candidate causal sites within pathways, chromosomes, and other genomic subregions. We proposed the weighted sparse fused lasso and a data-driven weighting scheme to return smooth signal estimates and give power to rare variant detection. The method does not require any cross validation or hold-out data to fit the model, instead relying on an augmented data approach that can accommodate rare features. In simulations we demonstrated the variable selection and estimation abilities of the method, with the weighted fused lasso outperforming the lasso in a variety of settings. We applied the method to post-hoc analysis of chr13q22-32 to characterize the known association between CNVs in these regions and TG levels, and the genomic regions selected by the method were supported by gene annotations.

We chose to control the FSR at 0.2 and 0.25 in the real data analysis due to the combination of good power and low FPR and FDR we observed at these levels in the simulations. The simulations evaluating the FSR-control approach suggested that at levels 0.2 and 0.25, the choices of $\gamma \in \{0.5, 1, 2, 5\}$ should maintain good power and protection against false positives in most of the fused lasso methods. Indeed, the FPR and FDR remained relatively moderate among all methods for $\gamma \in \{0.5, 1, 2, 5\}$ and FSR control at 0.2 and 0.25, but the TPR of the $D_{IF}(k,0)$, then the naive, and finally the $D_{CS}(k,1)$ methods decayed.

Making inference from the real data analysis, we have the most confidence in results under the FSR-control approach which perform stably in both the deletions and duplications analysis. The AIC produced a noisy model fit in the deletions analysis which likely contains many spurious signals, and we believe the BIC model fits in the duplications analysis are overly conservative since the CONCUR test results indicate that there is a signal. Under the FSR model selection, the $D_{IF}(k,2)$, $D_{CS}(k,1)$, and $D_{CS}(k,2)$ methods identify signals that reflect the CNV activity and are interpretable and plausible given the gene annotations.

All three methods highlight the ABCC4 gene region in the duplications analysis, which in our data is marked by a high concentration (roughly 1% MAF) of very short duplications. As previously mentioned, the relationship between this gene and triglycerides is well supported by the literature and in this analysis is a likely true positive. In the analysis of deletions, we draw conclusions from the $D_{IF}(k,2)$ fit selected under FSR control at 0.25. The method identifies a region encompassing all of the PCCA gene and the TMTC4 gene. The region is populated by two main clusters of activity, with long CNVs (roughly 500Mb) spanning the two. While the FSR-control approach is powerful in variable selection, we expect that the estimates it returned in this analysis and the duplications analysis are biased. Based on the CNV activity, there is reason to suspect possible effect heterogeneity in the identified causal regions. However, the FSR-control approach is not designed to return unbiased estimates of these effects. If less biased estimates are desired, we would recommend a refitting approach, such as applying ordinary least squares to the predictors selected by the fused lasso methods.

We remark that the proposed method can be used to construct a risk profile curve, which may be
of interest in its functional form or as means of calculating a collapsing risk score. After fitting and model selection, we can translate the estimated coefficients to a curve \( \hat{\beta}(s) \) that takes in a location \( s \) in the target region and outputs the estimated effect in that location. \( \hat{\beta}(s) \) may take on non-zero values for locations inside segments with non-zero \( \hat{\beta} \), or with \( |\hat{\beta}| > \epsilon \) under a thresholding approach, and value 0 otherwise. Then, as discussed in Section 3.2.5, we can consider an individual’s risk profile as \( \int f_i^*(s) \hat{\beta}(s) \), where \( f_i^*(s) \) is the duplication or deletion profile curve. In future work we hope to explore this downstream use of the weighted fused lasso fits.
4.1 Thesis conclusion

In this thesis we develop novel methods for the global and local association analysis of CNVs. In Chapter 2, we propose the CONCUR method which leverages kernel machine regression to model the aggregated effects of rare CNVs on the genome, chromosome, or pathway level. The method integrates dosage and length information in the novel CNV profile curve and derives a new similarity measure based on this concept. It additionally addresses the limitations of existing methods through its treatment of fine-resolution genomic location as locus, circumventing the need for artificially constructed CNVR or laborious off-locus alignment. In simulations and real data analysis it shows enhanced performance over comparator methods. An R package implementing the method is available and we continue to update it for additional functionality.

The weighted sparse fused method introduced in Chapter 3 novelly applies the lasso to the local analysis of both rare and common CNVs. We borrow concepts from Chapter 2 to represent CNV activity functionally across the genome, which through discretization maps the CNV data to design matrices that are compatible with a plethora of statistical modeling techniques. We show the advantage of weighted fused lasso methods over the standard lasso and unweighted fused lasso, with weights based on patterns in CNV activity in the sample data. Moreover, the method harnesses existing powerful optimization tools for penalized regression and model selection. Finally, the method is highly customizable through the $D$ penalty design matrix.
4.2 Future work

For the CONCUR method, there is the potential to combine it with a burden kernel as in SKAT-O [Lee12] for enhanced power under burden signals. In addition, we are currently adding functionality to the CONCUR R package to allow users to incorporate alternative dosage transform functions besides $|d - 2|$. Currently we are advising users to explore CNV-level transformations that modify $d$. For example, CNVs of low quality may have scaled down $d$ values so that their contribution to similarity metrics is downweighted.

We note the representation of the CNV profile curve as a point process, [Dal07], specifically a marked temporal point process. The AUC measure of similarity was inspired by distance measures for temporal point processes, in particular that proposed by Xiao et al. in [Xia17]. They map point processes to discrete distributions and derive a distance between the distributions that translates to the volume by which two counting measures differ. This representation opens the door to other distance metrics which may be applied to point process representations of CNV activity. We highlight this connection and recommend this field for further inspiration.

For the weighted sparse fused lasso method, we are extending the method to the generalized linear model setting, as many studies of CNVs involve binary phenotypes. This work relies on the constrained lasso discussed in Section 1.2.3. We further intend to research additional model selection approaches for the fused lasso setting, rare feature setting, and their intersection. We also note the alternative of applying the $\ell_2$ penalty rather than $\ell_1$ penalty to the generalized lasso term $D\beta$. A given graph-based $D$ may be mapped to an equivalent Laplacian matrix for a network-constrained lasso approach with the same smoothing structure. We may consider the $\ell_2$-penalty analogues of our methods in future study.

Finally, although not explored here, we remark that trace plots showing the complete solution path are a useful tool in their own right and may lead to inference without the need for model selection. For instance, as explored in [Zho10], the order of entry of fused groups as $\lambda$ decreases may indicate the relative strength of the signals the groups contain.


APPENDICES
A.1 Proof of symmetry and positive semi-definiteness of cAUC kernel

In order to show that $k_{cAUC}(\cdot, \cdot)$ is a valid kernel function, we need to show the associated kernel matrix is symmetric and positive semi-definite (PSD). The cAUC kernel function is given by

$$k_{cAUC}(Z_i, Z_j) = \sum_{k=1}^{22} \int_N \left[ \min \left( f_{ik}^{Dup}(x), f_{jk}^{Dup}(x) \right) + \min \left( f_{ik}^{Del}(x), f_{jk}^{Del}(x) \right) \right] d\mu(x)$$

(A.1)

where $Z_i$ and $Z_j$ represent the input data matrices containing CNV chromosome, starting and end locations, and dosage information for all CNVs in the profiles of individuals $i$ and $j$, for $i, j = 1, \cdots, n$ individuals; where $f_{ik}^{Dup}(x)$ and $f_{ik}^{Del}(x)$ are the duplication profile curve and deletion profile curve of individual $i$ on chromosome $k$; and $\mu(x)$ is the counting measure. The derivation of the duplication and deletion profile curves as a function of the input data $Z_i$ is detailed in the manuscript Methods section. We define the $n \times n$ kernel matrix $K^{Dup}$ such that its $(i, j)$th entry is given by

$$K_{ij}^{Dup} = \sum_{k=1}^{22} \int_N \min \left( f_{ik}^{Dup}(x), f_{jk}^{Dup}(x) \right) d\mu(x)$$

(A.2)

for $i, j = 1, \cdots, n$, with $K^{Del}$ defined similarly so that $K = K^{Dup} + K^{Del}$.

It is sufficient to show that $K^{Dup}$ is symmetric and PSD. Define $f_i$ to be $f_{ik}^{Dup}(x)$ at a fixed value
of \( x \) on a fixed chromosome \( k \) and likewise for \( f_j \). Then define the function

\[
    k(f_i, f_j) = \min(f_i, f_j)
\]  

(A.3)

and a corresponding kernel matrix \( K \) with its \((i, j)\)th element \( K_{ij} = k(f_i, f_j) \). Note that \( K_{ij}^{Dup} \) is simply the sum of \( k(f_i, f_j) \) across all \( x \) in a chromosome across all chromosomes \( k = 1, \cdots, 22 \). Therefore, if \( K \) is symmetric and PSD, so is the \( K^{Dup} \).

First, it is evident that \( K \) is symmetric kernel, since the minimum operator is symmetric. Now, to show that \( K \) is PSD, we must show that

\[
    \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j k(f_i, f_j) \geq 0 \tag{A.4}
\]

\( \forall \alpha = (a_1, \cdots, a_n) \) and for all possible \( f_i, f_j \geq 0 \). Call the domain of the duplication profile curves \( F \). Then we have

\[
    \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \min(f_i, f_j) = \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{F} \alpha_i \alpha_j 1_{t \leq f_i}(t) 1_{t \leq f_j}(t) \, dt \tag{A.5}
\]

\[
    = \int_{F} \left( \sum_{i=1}^{N} \alpha_i 1_{t \leq f_i}(t) \right) \left( \sum_{j=1}^{N} \alpha_j 1_{t \leq f_j}(t) \right) \, dt \tag{A.6}
\]

\[
    = \int_{F} G(t)^2 \, dt \geq 0 \tag{A.7}
\]

where \( 1_{t \leq f}(t) \) is a function that takes value 1 if \( t \leq f_i \) and value 0 otherwise, and \( G(t) = \sum_{i=1}^{N} \alpha_i 1_{t \leq f_i}(t) \). Hence \( K \) is positive semi-definite and therefore so is \( K^{Dup} \), and \( K^{Del} \) by the same logic.

### A.2 Details of TwinGene pseudo CNV data whole genome (TGP-WG) simulation design

In the TGP-WG simulations, we examined the methods' performance in two main scenarios: under a dosage×length signal and under a dosage-only signal. Within each scenario, we considered three sub-scenarios with different causal effects: causal effects from both duplications and deletions, causal effects from duplications only, and causal effects from deletions only. In each sub-scenario, we designated varying percentages of the causal segments to be deleterious (D) or protective (P). When both duplication and deletion were causal, the settings included \( (D_{Dup}, P_{Dup}, D_{Del}, P_{Del}) = (100, 0, 100, 0), (50, 50, 50, 50), (90, 10, 10, 90) \) and \( (10, 90, 90, 10) \). In scenarios where duplications (or deletions) alone are causal, possible settings included \( (D_\ast, P_\ast) = (100, 0), (50, 50), \) and \( (10, 90) \).

Among the CNV segments across the genome, we selected 200 segments to be causal, which consist of 100 causal “dup-segments” with at least one duplication and another 100 causal “del-segments” with at least one deletion. A causal dup-segment cannot be a causal del-segment. These
causal segments were chosen as a random draw of 50 pairs of adjacent segments which both contained duplications, and another 50 pairs of adjacent segments which both contained deletions. This adjacent causal segment approach was designed to ensure that causal regions had more realistic lengths, since some segments were very short by chance.

The implementation of the CONCUR_cat, CCRET, and CKAT methods is given in the main text. The power was evaluated in the range of odds ratios \( (e^{\beta}) \) 1.02-1.10 in the scenarios with dosage×length effects and 1.1-1.9 in the scenarios with dosage-only effects. Power estimates are reported for a range of effect sizes such that the power ranges roughly from 0.2 to 0.8.

### A.3 TGP-WG simulation results
Figure A.1 Results of TGP-WG simulation with causal dosage×length effects. This figure shows the power comparison between CONCUR, CCRET, and CKAT in the TGP-WG simulations (causal dosage×length effects across the genome of TGP data). The top panel shows power under combined duplication and deletion effects, the middle panel shows power under effects from duplications only, and the bottom panel shows power under effects from deletions only. Different proportions of deleterious vs. protective effects are considered as indicated by $(D_{Dup}, P_{Dup}, D_{Del}, P_{Del})$ with $D_{Dup}$ and $P_{Dup}$ reflecting the proportions of deleterious and protective segments among causal duplication segments, and with $D_{Del}$ and $P_{Del}$ defined similarly for causal deletion segments.
Figure A.2 Results of TGP-WG simulation with causal dosage effects. This figure shows the power comparison between CONCUR, CCRET, and CKAT in the TGP-WG simulations (causal dosage effects across the genome of TGP data). The top panel shows power under combined duplication and deletion effects, the middle panel shows power under effects from duplications only, and the bottom panel shows power under effects from deletions only. Different proportions of deleterious vs. protective effects are considered as indicated by $(D_{Dup}, P_{Dup}, D_{Del}, P_{Del})$ with $D_{Dup}$ and $P_{Dup}$ reflecting the proportions of deleterious and protective segments among causal duplication segments, and with $D_{Del}$ and $P_{Del}$ defined similarly for causal deletion segments.
A.4 Post-hoc pathway analysis of Taiwan Biobank CNV data in lipid metabolism pathway hsa01040

To illustrate possible CONCUR post hoc analyses to probe the potential sources of the pathway-level signals found in the TWB analysis, we looked more closely at one pathway, hsa01040 (biosynthesis of unsaturated fatty acids), for which both CONCUR and CKAT were significant while CCRET was borderline significant. Previous studies have reported that monounsaturated fat acids or polyunsaturated fatty acids can effect TG levels [Gru89; Ooi15]. Given the major function of the genes in hsa01040 (i.e., the biosynthesis of unsaturated fatty acids), it is not unexpected that CNVs in these genes were significantly associated with TG levels. In Table A.1, we reported summary statistics describing CNV length and dosage in hsa01040 for individuals with different levels of TG. Based on the TG quantiles from the sample data, we classified individuals as having high TG (>75th percentile [>140 mg/dL]), medium TG (25th–75th percentile [68-140 mg/dL]) and low TG (<25th percentile [<68 mg/dL]). We applied ANOVA to detect differences in CNV length and in dosage characteristics, and applied chi-squared tests to assess differences in the proportion of individuals with CNVs across TG levels. In addition, we examined CNV features in all CNVs together and in duplications and deletions separately.
**Table A.1 Descriptive statistics for hsa01040 pathway.** TG values are classified as Low (<the 25th percentile [<68 mg/dL]; n=2,931), Medium (the middle 50% [68 - 140 mg/dL]; n=5,844), and High (>the 75th percentile [>140 mg/dL]; n=2,889). The percent of individuals with CNVs is with respect to the total number of individuals in each TG category. The mean number of CNVs per individual and mean total length of CNVs (bp) per individual are reported, as well as the mean lengths (bp) and mean dosage per CNV. “Promising” associations with TG are marked with ⋆⋆ to indicate p-value < 0.01 and with ⋆ to indicate p-value < 0.05.

<table>
<thead>
<tr>
<th>CNV Type</th>
<th>TG Level</th>
<th>Pct Individuals with CNV</th>
<th>Mean # CNVs per Individual</th>
<th># Genes Interrupted</th>
<th>Mean Total CNV Length per Individual (bp)</th>
<th>Mean CNV Length (bp)</th>
<th>Mean CNV Dosage</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>Low</td>
<td>6.18%</td>
<td>3.33</td>
<td>23</td>
<td>25143.71</td>
<td>2433.58</td>
<td>1.63</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>6.07%</td>
<td>3.52</td>
<td>23</td>
<td>24447.30</td>
<td>2473.48</td>
<td>1.63</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>7.17%</td>
<td>3.48</td>
<td>23</td>
<td>31091.65</td>
<td>2471.43</td>
<td>1.64</td>
</tr>
<tr>
<td>Deletion</td>
<td>Low</td>
<td>2.8%</td>
<td>5.84</td>
<td>16</td>
<td>29630.62</td>
<td>2590.55*</td>
<td>1.41</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>2.74%</td>
<td>6.24</td>
<td>17</td>
<td>28107.36</td>
<td>2593.05*</td>
<td>1.40</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>3.32%</td>
<td>5.79</td>
<td>16</td>
<td>32039.63</td>
<td>2067.96*</td>
<td>1.39</td>
</tr>
<tr>
<td>Duplication</td>
<td>Low</td>
<td>3.62%</td>
<td>1.17</td>
<td>20</td>
<td>7811.20</td>
<td>1827.24**</td>
<td>2.50*</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>3.54%</td>
<td>1.22</td>
<td>23</td>
<td>10009.61</td>
<td>2001.81**</td>
<td>2.52*</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>4.15%</td>
<td>1.38</td>
<td>22</td>
<td>27897.23</td>
<td>3831.02**</td>
<td>2.49*</td>
</tr>
</tbody>
</table>
Taking p-values $< 0.05$ as suggestive of a "promising" association with TG, we did not observe any CNV associations when all CNVs were analyzed together; but for duplications only, there were promising differences in CNV length (p-value=0.0063) and weaker differences in dosage (p-value=0.0255) across TG levels. There were also some weak significance in CNV length for deletions (p-value=0.0423). We were cautious to not over-interpret these “promising” associations since this stratified analysis reflected only marginal associations of a CNV feature, and the tests did not account for the effect heterogeneity that motivates the application of kernel-based methods. We also applied CONCUR to duplications and deletions separately, and found a very significant association with TG in duplications (p-value $< 1 \times 10^{-8}$) and a weaker signal in deletions (p-value=0.0313).

To further explore the signal from duplications, we visualized CNVs in the 23 genes in hsa01040. Figure A.3 displays duplications and deletions in the CNV profiles of individuals categorized by their TG level (low, medium, and high), with profiles clustered using the Weighted Pair Group Method with Arithmetic Mean (WPGMA) hierarchical clustering method.
Figure A.3 Visualization of CNV activity in pathway hsa01040 by level of triglycerides (TG). CNV activity in genes in hsa01040 is shown by level of TG (Low, Medium, and High), with duplications in red and deletions in darkolive. Columns represent individuals, and genes shown here are the 23 genes in the pathway that contain CNVs, ordered by the number of CNVs contained therein.
For further exploration, we applied CONCUR to duplications in each gene and found that several genes had strong association p-values (i.e., $< 10^{-4}$), BAAT, ELOVL4, ELOVL6, ELOVL5, HSD17B4, and SCD5. Notably, BAAT is an amino acid N-acyltransferase for bile acid. Previous studies have demonstrated that bile acids are important regulators for TG level through crosstalk with farnesoid X receptor (FXR) [Lie14; Wat04]. Since conversion of cholesterol to bile acid is an essential step in preventing the accumulation of TG, copy number duplications in BAAT may directly affect TG levels in the blood. Three ELO genes had significant CNV associations. Since the major functions of these genes focus on the elongation of fatty acids, CNV events in these genes are likely to affect the production and metabolism of TG. For example, one study showed that hepatic steatosis was observed in ELOVL5-knockout mice due to the activation of SREBP-1c and its target genes [Sas14]. HSD17B4 is a dehydrogenase, which is able to inhibit the production of DHEA [DL99]. A previous study showed that TG levels were inversely correlated to DHEA levels in men with type 2 diabetes [Bou01], suggesting a potential link between CNVs in HSD17B4 and TG levels. SCD5 serves as a critical enzyme providing a double bond to construct complex lipid molecules such as TG [Cas11; Flo08], and thus dysregulation of SCD5 expression may impact TG levels. Further analyses are required to formally localize the sources of the CNV association signal in this pathway and others, but this exploratory analysis nonetheless serves to enrich our understanding of the association in pathway hsa01040 through examination of CNV-level and gene-level features.
A.5 TWB negative control analyses in lipid metabolism pathways and chromosome 13

Figure A.4 QQ-plot for CONCUR p-values in TWB lipid metabolism pathways negative control analysis. P-values from pathway analyses with permuted TG levels, with 20 replications per pathway (20*15 p-values). Expected p-values come from a Uniform(0,1) distribution.

Figure A.5 QQ-plot for CONCUR p-values in TWB chromosome 13 negative control analysis. P-values from 1000 replications of the chromosome 13 analysis with permuted TG levels. Expected p-values come from a Uniform(0,1) distribution.
APPENDIX

B

SUPPLEMENTAL INFORMATION FOR CHAPTER 3
B.1 Simulation analysis: performance of the FSR-control model selection approach at varying $\gamma$

Figure B.1 Performance metrics for model selection using FSR control across 100 replications for 6 lasso methods, at $\gamma = 0.5$. Metrics are reported at FSR control levels of 0.1, 0.15, 0.2, and 0.25. $\gamma$ is the value of the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.
Figure B.2 Performance metrics for model selection using FSR control across 100 replications for 6 lasso methods, at $\gamma = 2$. Metrics are reported at FSR control levels of 0.1, 0.15, 0.2, and 0.25. $\gamma$ is the value of the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.
Figure B.3 Performance metrics for model selection using FSR control across 100 replications for 6 lasso methods, at $\gamma = 5$. Metrics are reported at FSR control levels of 0.1, 0.15, 0.2, and 0.25. $\gamma$ is the value of the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.
Figure B.4 Performance metrics for model selection using FSR control across 100 replications for 6 lasso methods, at $\gamma = 10$. Metrics are reported at FSR control levels of 0.1, 0.15, 0.2, and 0.25. $\gamma$ is the value of the lasso penalty scaling factor. Note that y-axis range differs across the 4 panels.
B.2 Simulation analysis: Skyline plots for the AIC and BIC model selection approaches

![Skyline plots for different models](image-url)

Figure B.5 Model fits at $\lambda_{BIC}$ across 100 replications for 6 lasso methods, at $\gamma = 1$. The mean $\hat{\beta}(\lambda_{BIC})$ is plotted in red on the y-axis and the true $\beta$ is shown in black. The coefficients are ordered according as shown on the x-axis.
Figure B.6 Model fits at $\lambda_{AIC}$ across 100 replications for 6 lasso methods, at $\gamma = 1$. The mean $\hat{\beta}(\lambda_{AIC})$ is plotted in red on the y-axis and the true $\beta$ is shown in black. The coefficients are ordered according as shown on the x-axis.
B.3 Analysis of deletions in chr13q22-32 in the TWB data

Figure B.7 Model fits for chr13q22-32 deletions analysis at $\lambda_{BIC}$ with thresholding for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.2. The estimates $\hat{\beta}(\lambda_{BIC})$ are plotted in red on the y-axis, and the black line is the threshold. The coefficients are ordered according as shown on the x-axis.
Figure B.8 Model fits for chr13q22-32 deletions analysis at $\lambda_{AIC}$ with thresholding for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.2. The estimates $\hat{\beta}(\lambda_{AIC})$ are plotted in red on the y-axis, and the black line is the threshold. The coefficients are ordered according as shown on the x-axis.
Figure B.9 Model fits for chr13q22-32 deletions analysis at $\lambda_{FSR}$ for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.2. The estimates $\hat{\beta}(\lambda_{FSR})$ are plotted in red on the y-axis. The coefficients are ordered according as shown on the x-axis.
### B.4 Analysis of duplications in chr13q22-32 in the TWB data

![Graphical representation of duplications analysis](image)

**Figure B.10** Model fits for chr13q22-32 duplications analysis at $\hat{\lambda}_{\text{BIC}}$ with thresholding for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.2. The estimates $\hat{\beta}(\hat{\lambda}_{\text{BIC}})$ are plotted in red on the y-axis, and the black line is the threshold. The coefficients are ordered according as shown on the x-axis.
Figure B.11 Model fits for chr13q22-32 duplications analysis at $\lambda_{AIC}$ with thresholding for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.2. The estimates $\hat{\beta}(\lambda_{AIC})$ are plotted in red on the y-axis, and the black line is the threshold. The coefficients are ordered according as shown on the x-axis.
Chr13q duplications: FSR control at 0.2

Figure B.12 Model fits for chr13q22-32 duplications analysis at $\lambda_{FSR}$ for 6 lasso methods, at $\gamma = 1$ and FSR control level 0.2. The estimates $\hat{\beta}(\lambda_{FSR})$ are plotted in red on the y-axis. The coefficients are ordered according as shown on the x-axis.