

ABSTRACT

MARCEAU WEST, RACHEL ELIZABETH. Flexible Kernel Machine Methods for Complex Genomic Data. (Under the direction of Jung-Ying Tzeng and Wenbin Lu.)

Rare variant associations are essential to fully understanding and trying to effectively treat complex diseases. Rare variants are often hypothesized to jointly explain much of the missing heritability from common variant genome-wide association studies, and are believed to be more likely than their common counterparts to have direct functional effect on disease etiology. However, rare variant associations are still quite difficult to detect with existing statistical methodology, especially when the question of interest goes beyond simple genetic main effect testing. Kernel machine models provide a flexible framework to facilitate complex rare variant studies. We consider three such interesting problems: (1) multi-factor analysis using multiple kernels to jointly model different variable sets, testing for the effect of one set while adjusting for nuisance effects; (2) cross-disorder gene variable selection, using kernels to leverage information on the co-occurrences of and genetic correlations between diseases; and (3) rare variant prioritization using protein structure, forming variant-level tests from local kernels that borrow information from variants that are nearby on the protein tertiary space.

In project 1, we demonstrate use of a low-rank decomposition of the nuisance effect terms to improve computational efficiency over traditional estimation (i.e., using an Expectation Maximization algorithm or penalization). We extend this idea in project 2, showing how this decomposition can be used to facilitate gene-level variable selection using group lasso when sample size and number of variants is large. Further, in project 2 we demonstrate a significant power gain from borrowing information from correlated diseases. In project 3, we continue looking at variable selection but try to prioritize rare variants at the variant level and demonstrate how protein folding structure can be incorporated into variant-level association testing to improve power and generate more biologically interesting hypotheses.

We establish the validity and power of our methods using simulation studies with continuous,

binary, and survival traits. We further apply our methods to data from the Vitamin Intervention for Stroke Prevention (VISP), CoLaus, and Action to Control Cardiovascular Risk in Diabetes (ACCORD) clinical trials, finding promising associations for follow-up analyses.

© Copyright 2017 by Rachel Elizabeth Marceau West

All Rights Reserved

Flexible Kernel Machine Methods for Complex Genomic Data

by
Rachel Elizabeth Marceau West

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

2017

APPROVED BY:

Arnab Maity

Eric Stone

Jung-Ying Tzeng
Co-chair of Advisory Committee

Wenbin Lu
Co-chair of Advisory Committee

DEDICATION

To Clover.

BIOGRAPHY

Rachel Elizabeth Marceau was raised equally in the Triad of North Carolina and in Nova Scotia, Canada. She gained a love of statistics during her brief time at Southwest Guilford High School, playing with M&M distributions in A.P. Statistics. After a short year studying astrophysics at the University of Arizona, Rachel transferred to North Carolina State University to pursue her Bachelor's Degree in Statistics. During this time, she had the opportunity to take two genetics classes from Dr. Ted Emigh, who piqued her curiosity in how statistics can be applied to help answer interesting questions in the worlds of human genetics and population biology. She was able to further build a passion for statistical genetics problems in the Computation for Undergraduates in Statistics Program (CUSP) under the direction of Dr. David Reif and Dr. Alison Motsinger-Reif. In 2011, Rachel graduated summa cum laude with a minor in mathematics. After a summer working with Dr. Spencer Muse trying to quantify nonsynonymous mutation rates of mitochondrial DNA, she began her graduate studies at NC State University, earning a Masters of Statistics in 2013 with a concentration in statistical genetics. During this time, Rachel got married to her undergrad sweetheart Charlie and gave birth to her wonderful daughter Clover. Rachel is set to obtain her Ph.D. in 2017 under the direction of Dr. Jung-Ying Tzeng and Dr. Wenbin Lu, after which she hopes to continue working towards improving our understanding of complex diseases through statistical analyses for many years to come.

ACKNOWLEDGEMENTS

I would like to start by thanking my incredible advisors Dr. Jung-Ying Tzeng and Dr. Wenbin Lu for all of their support and guidance, as well as the rest of my committee, Dr. Eric Stone, Dr. Arnab Maity, and Dr. Denis Fourches for all of their time, patience, knowledge, and thoughtful suggestions.

I would also like to thank my collaborators who provided access to the clinical trial data studied in this paper, and who helped me to explore and better understand the complex genetic consequences of our findings: Dr. Fang-Chi Hsu, Dr. Michèle Sale, Dr. Bradford Worrall, and Dr. Stephen Williams, co-authors of the fastKM paper, and collaborators with the VISPC clinical trial, and the ever-incredible Dr. Méline Kuenemann, and Dr. Daniel Rotroff, from the ACCORD clinical trial project. I would also like to thank Song et al. (2012) for access to the *PLA2G7* data, Dr. Fourches and Dr. Kuenemann for use of their protein structure figures for *PLA2G7* and *PCSK9*, respectively, and Dr. Shannon Holloway for her help in creating the fastKM R package.

I have been blessed to have many other wonderful mentors on my academic journey, including Dr. Spencer Muse and Dr. Alison Motsinger-Reif, who have inspired and guided me since I was an undergraduate student. I would also like to thank Dr. Joy Smith and Dr. Emily Griffith for guiding me as a statistical consultant and as a person. I would also like to thank the countless professors from NCSU who have helped my love of statistics grow.

Finally, I would like to thank my wonderful family and friends who have always supported me, and my loving and patient husband, Charlie. Thank you for sticking with me through it all.

This dissertation was supported by the NIH training grant T32GM081057: Biostatistics Training in the Omics Era.

TABLE OF CONTENTS

LIST OF TABLES	vii
LIST OF FIGURES	ix
Chapter 1 Introduction	1
1.1 Multi-Kernel Models	3
1.2 Cross Disorder Analysis	4
1.3 Rare Variant Prioritization via Protein Structure	6
1.4 Summary of Dissertation	7
Chapter 2 A Fast Multiple-Kernel Method with Applications to Detect Gene-Environment Interaction	9
2.1 Abstract	9
2.2 Key Words	10
2.3 Introduction	10
2.4 Materials and Methods	13
2.4.1 The KM Model for G x E Interactions	13
2.4.2 FastKM Test for G x E Interactions	15
2.4.3 Extension to Survival Traits	16
2.4.4 Implementation Detail	17
2.4.5 Simulation Study	18
2.4.6 Application to VISP Study	19
2.5 Results	21
2.5.1 Quantitative Traits	21
2.5.2 Binary and Survival Traits	26
2.5.3 VISP Study	30
2.6 Discussion	33
2.7 Acknowledgments	35
2.8 Supplementary Materials	35
2.8.1 Common Variants based Simulations	35
Chapter 3 Cross Disorder Kernel Machine Modeling	37
3.1 Abstract	37
3.2 Introduction	38
3.2.1 Motivation for Cross Disorder Analysis	38
3.2.2 Current Methods	39
3.2.3 Introduction to fastLasso	47
3.3 Methods	48
3.3.1 Kernel Evaluation	49
3.3.2 Dimension Reduction	50
3.3.3 fastLasso	50
3.3.4 Computational Efficiency	51

3.4	Simulation Study	52
3.4.1	Data Generation	52
3.4.2	fastLasso Simulation	53
3.5	Results	54
3.6	Discussion	57
Chapter 4 Rare Variant Prioritization Using Structure-Supervised Kernel Association		
	Tests	59
4.1	Abstract	59
4.2	Introduction	60
4.3	Methods	65
4.3.1	Structure-Supervised Kernel Machine Association Testing	65
4.4	Simulation Study	70
4.4.1	Simulation Set Up	70
4.4.2	Simulation Study Results	72
4.5	Application to ACCORD Study	83
4.5.1	ACCORD Trial Background	83
4.5.2	ACCORD Analysis	84
4.5.3	ACCORD Results	85
4.6	Discussion	89
REFERENCES		91
APPENDICES		104
Appendix A	Additional Information for Cross Disorder Kernel Machine Simulation	105
A.0.1	Case Control Sampling of Genotype Matrix for Binary Traits	105
A.0.2	Cross Disorder and Single Disorder Tuning Parameter Summaries . . .	107
Appendix B	Local Score Test Limiting Distribution	108
Appendix C	Additional Local Kernel Simulation Results	113
C.0.1	Additional Simulation Results: Quantitative Traits Local Burden Kernel Test	114
C.0.2	Additional Simulation Results: Binary Traits Local Burden Kernel Test .	116
C.0.3	Additional Simulation Results: Quantitative Traits Local Linear Kernel Test	118
C.0.4	Additional Simulation Results: Binary Traits Local Linear Kernel Test .	120
Appendix D	ACCORD Rare Variants Summary	122

LIST OF TABLES

Table 2.1	Average run time in minutes (and corresponding standard error) for quantitative traits when the G×E effect is zero with a sample size of $n = 5,000$ individuals.	23
Table 2.2	Average run time in minutes (and corresponding standard error) for quantitative traits when the G×E effect is nonzero with a sample size of $n = 5,000$ individuals.	24
Table 2.3	P -values for the fastKM analyses of VISP study data, including (a) testing gene × age interaction on post-methionine change in total Hcy, treating change as continuous, with a IBS kernel or polynomial kernel ($d=2$); (b) testing gene × age interaction on post-methionine change in total Hcy, treating change as binary using the 90 th sample percentile as a cut off; (c) testing gene × intervention interaction on time to recurrent stroke * P -values that are < 0.05	32
Table 3.1	Average true positive and false positive rates (and corresponding standard deviation) for cross disorder (CD), the union of single disorder (SD-U), and the intersection of single disorder (SD-I) continuous trait kernel machine model analyses over 100 simulations. Largest values within each category are in bold font.	56
Table 4.1	Type I error and power for single-variant and local burden kernel tests for quantitative traits over various simulation scenarios using <i>PLA2G7</i> protein tertiary structure	74
Table 4.2	Type I error and power for single-variant and local burden kernel tests for binary traits over various simulation scenarios using <i>PLA2G7</i> protein tertiary structure	75
Table 4.3	Type I error and power for single-variant and local linear kernel tests for quantitative traits over various simulation scenarios using <i>PLA2G7</i> protein tertiary structure	78
Table 4.4	Type I error and power for single-variant and local linear kernel tests for binary traits over various simulation scenarios using <i>PLA2G7</i> protein tertiary structure	79
Table 4.5	Scan statistic results for binary traits, summarized over 500 replications	82
Table 4.6	<i>PCSK9</i> rare variant single SNP and localized kernel test summary using PDB entry 4K8R. Significant results are given in bold font. Q-values are calculated assuming $\pi_0 = 0$ (e.g., Benjamini-Hochberg corrected values) due to low variant pool size.	86
Table A.1	Average optimal tuning parameter (and corresponding standard deviation) for the cross disorder and single disorder continuous trait kernel machine models over 100 simulations	107

Table C.1	Type I error and power for continuous trait local burden kernel test, weighted by minor allele frequency, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using <i>PLA2G7</i> protein tertiary structure	114
Table C.2	Type I error and power for continuous trait local burden kernel test, unweighted, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using <i>PLA2G7</i> protein tertiary structure	115
Table C.3	Type I error and power for continuous trait local burden kernel test, unweighted, with $c = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$ over various causal variant scenarios using <i>PLA2G7</i> protein tertiary structure	115
Table C.4	Type I error and power for binary trait local burden test, weighted by minor allele frequency, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using <i>PLA2G7</i> protein tertiary structure	116
Table C.5	Type I error and power for binary trait local burden kernel test, unweighted, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using <i>PLA2G7</i> protein tertiary structure	117
Table C.6	Type I error and power for binary trait local burden kernel test, unweighted, with $c = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$ over various causal variant scenarios using <i>PLA2G7</i> protein tertiary structure	117
Table C.7	Type I error and power for continuous trait local linear kernel test, weighted by minor allele frequency, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using <i>PLA2G7</i> protein tertiary structure	118
Table C.8	Type I error and power for continuous trait local linear kernel test, unweighted, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using <i>PLA2G7</i> protein tertiary structure	119
Table C.9	Type I error and power for continuous trait local linear kernel test, unweighted, with $c = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$ over various causal variant scenarios using <i>PLA2G7</i> protein tertiary structure	119
Table C.10	Type I error and power for binary trait local linear kernel test, weighted by minor allele frequency, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using <i>PLA2G7</i> protein tertiary structure	120
Table C.11	Type I error and power for binary trait local linear kernel test, unweighted, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using <i>PLA2G7</i> protein tertiary structure	121
Table C.12	Type I error and power for binary trait local linear kernel test, unweighted, with $c = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$ over various causal variant scenarios using <i>PLA2G7</i> protein tertiary structure	121
Table D.1	<i>PCSK9</i> rare variants summary using PDB entry 4K8R	122

LIST OF FIGURES

Figure 2.1	Type I error for fastKM, originalKM, and the weighted counting burden-based G×E test for quantitative traits with $M = 100$ loci, $n = 5,000$ individuals, and varying main effect parameter γ_G . Models with a continuous environmental E covariate are on the left and those with a binary environmental E covariate are on the right. The KM tests are based on the IBS kernel.	22
Figure 2.2	Power for fastKM, originalKM, and the weighted counting burden-based G×E test for quantitative traits with $M = 100$ loci and $n = 5,000$ individuals over varying interaction effect sizes γ_{GE} . The left panel shows the results of no genetic main effect (i.e., $\gamma_G = 0$) and the right panels shows the results of nonzero main effect (i.e., $\gamma_G = 1$). For each plot, continuous E covariates are on the left and binary E covariates are on the right. The KM tests are based on the IBS kernel.	23
Figure 2.3	Type I error for fastKM, originalKM, and the weighted counting burden-based G×E test for quantitative traits with continuous environmental E covariate and varying main effect parameter γ_G . The left panel shows the results of $M = 100$ loci and $n = 5,000$ individuals. The right panel shows the results of $M = 200$ loci and $n = 1,000$ individuals. The KM tests are based on the IBS kernel and the polynomial kernels with $d = 2$ and 3.	25
Figure 2.4	Power for fastKM, originalKM, and the weighted counting burden-based G×E test for quantitative traits with continuous E covariate over varying main effect size γ_G ($\gamma_G = 0$ for zero main effect, and $\gamma_G = 1$ for nonzero main effect) and interaction effect size γ_{GE} . The left panel shows the results of $M = 100$ loci and $n = 5,000$ individuals. The right panel shows the results of $M = 200$ loci and $n = 1,000$ individuals. The KM tests are based on the IBS kernel and the polynomial kernels with $d = 2$ and 3.	26
Figure 2.5	Type I error for fastKM and the weighted counting burden-based G×E test for binary traits for $M = 100$ loci, $n = 5,000$ individuals, and varying main effect parameter γ_G . Models where E is generated from a Gaussian distribution are displayed on the left, and those where E is from a Bernoulli distribution are on the right. The KM tests are based on the IBS kernel.	27
Figure 2.6	Type I error for survival traits for fastKM and the weighted counting burden-based G×E test with $M = 100$ loci, $n = 5,000$ individuals, $c=15\%$ and 40% censoring proportions, and varying main effect parameter γ_G . Models where E is generated from a Gaussian distribution are displayed on the left, and those where E is from a Bernoulli distribution are on the right. The KM tests are based on the IBS kernel.	28

Figure 2.7	Power for fastKM and the weighted counting burden-based G×E test for binary traits with $M = 100$ loci and $n = 5,000$ individuals over varying interaction effect sizes γ_{GE} . The left panel shows the results of no genetic main effect (i.e., $\gamma_G = 0$) and the right panels shows the results of nonzero main effect (i.e., $\gamma_G = 1$). For each plot, continuous E covariates are on the left and binary E covariates are on the right. The KM tests are based on the IBS kernel.	29
Figure 2.8	Power for fastKM and the weighted counting burden-based G×E test for survival traits with $M = 100$ loci and $n = 5,000$ individuals for varying interaction parameter γ_{GE} over two censoring proportions ($c=15\%$ and 40%). The left panel shows the results of no genetic main effect (i.e., $\gamma_G = 0$) and the right panels shows the results of nonzero main effect (i.e., $\gamma_G = 1$). For each plot, continuous E covariates are on the left and binary E covariates are on the right. The KM tests are based on the IBS kernel.	29
Figure 2.9	Type I error (left panel) and power (right panel) for fastKM, originalKM, and the weighted counting burden-based G×E test for common variants with continuous E covariate over varying main effect size γ_G and interaction effect size γ_{GE} . The KM tests are based on the IBS kernel and the polynomial kernels with $d = 2$ and $d = 3$, all with 95% kPCA.	36
Figure 3.1	Histogram of the non-zero coefficients from the null model fastLasso fit	54
Figure 3.2	True positive and false positive rates for cross disorder (CD), the union of single disorder (SD-U), and the intersection of single disorder (SD-I) continuous trait kernel machine model analyses over 100 simulations	57
Figure 4.1	<i>PLA2G7</i> (a) rare variants location on the protein tertiary structure, and (b) corresponding Euclidean distance-based clustering of variants. Protein tertiary structure published with permission from Dr. Fourches.	71
Figure 4.2	Amount of borrowing from neighboring variants for <i>PLA2G7</i> variant 110 for different values of c .	73
Figure 4.3	Variant correlation for <i>PLA2G7</i> SNVs	76
Figure 4.4	Summary of proportion of times a given c value was chosen as optimal for the local burden kernel test for causal and noncausal variants.	76
Figure 4.5	Variant selection probabilities for single variant and local burden kernel tests, separated by noncausal variants (effective type I error), and causal variants (effective power)	77
Figure 4.6	Summary of proportion of times a given c value was chosen as optimal for the local linear kernel test for causal and noncausal variants	80
Figure 4.7	Variant selection probabilities for single variant and local linear kernel tests, separated by noncausal variants (effective type I error), and causal variants (effective power)	80
Figure 4.8	Variant borrowing for the best scale c for local kernel tests of association between rare variants in <i>PCSK9</i> and LDL	87

Figure 4.9 *PCSK9* (a) distance-based clustering of *PCSK9* variants, and (b) rare variants location on the protein tertiary structure of *PCSK9* and binding site of *LDLR* (in yellow); Significant associations from the single variant approach are blue and new associations found using the local kernel test are pink. Protein tertiary structure published with permission from Dr. Kuenemann. 89

Chapter 1

Introduction

A major goal in statistical genetics is to facilitate genetic association studies involving complex genomic data - finding links between genes and phenotypes to help us better understand the etiology of complex diseases and thus how to better classify and treat them. Genome-wide association studies, or GWAS, have been used quite extensively in the past to look for associations between common genetic variants and disease phenotypes. However, these simple analyses leave a high proportion of heritability unexplained (Eichler et al., 2010, Manolio et al., 2009). Increasing attention is being placed on rare variants which occur infrequently (e.g., with a minor allele frequency, or MAF of less than 1%) in the human population, with researchers hypothesizing that combined effects from many of these rare variants within different genes may be able to explain a lot of this missing heritability (Bodmer and Bonilla, 2008, Eichler et al., 2010, Manolio et al., 2009, Morris and Zeggini, 2010). Improving genotyping technology, such as the introduction of Next Generation Sequencing (NGS) technology, which facilitates sequencing of billions of short reads relatively inexpensively (Lee et al., 2014), has improved our access to rare variant data, further motivating the shift in focus to studying less frequent variants.

Rare variants, however, tend to be quite difficult to detect individually, requiring much larger sample and effect sizes to detect associations (Lee et al., 2014, Li and Leal, 2008), and often pose a

challenge, especially when considering more complex genetic data. Two main classes of methods have been proposed to increase our power to detect rare variant associations. The first are burden-based methods (e.g., Li and Leal (2008), Madsen and Browning (2009), Morris and Zeggini (2010), Price et al. (2010)), which aggregate information across a set of variants using a weighted sum of effects. These models are quite simple to fit, but lose power from collapsing over noise variants, and possibly over variants with opposite effects. The second are similarity-based approaches, using the (statistically and intuitively) related mixed effects models (e.g., Goeman et al. (2004), Lin et al. (2013)), similarity regressions (e.g., Tzeng et al. (2014, 2009, 2011), Wang et al. (2014), Zhao et al. (2015)), and kernel machine models (e.g., Kwee et al. (2008), Lin et al. (2011), Liu et al. (2008, 2007), Wu et al. (2010, 2011, 2013)). These approaches aggregate information from multiple variants in a flexible manner, modeling the genetic similarity between pairs of individuals in variance/covariance of genetic random effect terms, similarity matrices, or kernel matrices, respectively.

We focus our attention on the robust kernel machine (KM) models. Kernel machine models allow for incorporation of complex genomic relationships, easily modeling nonlinear/non-additive epistatic effects. The framework can be applied to multiple trait types, such as continuous, binary, and survival traits, and can provide efficient dimension reduction to facilitate analysis of large sample size data required for rare variant analysis.

In this dissertation, we show how the kernel machine framework can be used to approach three important topics within rare variant genetic association testing:

1. Multi-kernel analyses, testing for association while accounting for nuisance effects, (e.g. accounting for genetic main effect terms when testing for gene-environment interaction, or testing for genetic main effect when accounting for population substructure),
2. Cross-disorder analyses, using unified information to better understand pleiotropy and discover genes associated with at least one trait, and
3. Prioritizing rare variants within a gene using protein structure information.

1.1 Multi-Kernel Models

Kernel machine models with multiple kernels can be useful for incorporating multiple random effects terms into a rare variant genetic association analysis, e.g., testing for gene-environment (GxE) (Lin et al., 2013, Tzeng et al., 2011, Wang et al., 2015d, Zhao et al., 2015) or gene-gene (GxG) (Larson and Schaid, 2013, Wang et al., 2014) interaction while accounting for genetic main effect, testing for genetic main effect while accounting for population substructure, or even testing for copy number variant (CNV) dosage while accounting for CNV length (Tzeng et al., 2015). With multiple effect terms, model misspecification is even more of a problem than with single variant-set genetic main effect models, so the robustness of KM models (e.g., over burden models) becomes even more important (Wang et al., 2015d).

The multi-kernel KM model is powerful, robust, and flexible but requires estimation of nuisance effects in order to perform valid score tests. Estimation of these nuisance effects can be very demanding due to their high dimensional nature, with dimension equal to the sample size. Existing approaches to estimating nuisance effects typically use penalization or an Expectation-Maximization (EM) algorithm. Penalization approaches (e.g., Lin et al. (2013)) use regularization to impose more sparse solutions, leading to easier estimation. EM algorithm-based approaches (e.g. Tzeng et al. (2011), Wang et al. (2014, 2015d), Zhao et al. (2015)) treat the nuisance variable as a random effect and aim to estimate the corresponding variance component. Both types of approaches are quite computationally demanding due to tuning or to multiple large matrix inversions, respectively, and do not straightforwardly extend to other trait types. They are especially more difficult to use for rare variant analyses, where in general larger sample sizes are required to obtain sufficient signal.

Kernel matrices in rare variant genetic association studies, however, typically have the property of being low rank, with rank often much less than the minimum of the sample size and number of genetic variants studied. This is intuitive as the genetic similarity between individuals over all

variants within a set may be quite low for rare mutations. Therefore, a low rank approximation of the nuisance kernel can be used in place of the kernel itself, allowing for the nuisance effects to be estimated as a low dimensional fixed effects term for which penalization is not necessary. The “fastKM” method does this low rank approximation to help make multi-kernel analyses more manageable, bending the problem down to an effective single kernel model which can be fit using existing software (e.g., SKAT (Wu et al., 2011)), while maintaining proper type I error and power as can be found using the much slower EM-based and penalization-based approaches.

1.2 Cross Disorder Analysis

KM models can also be useful for rare variant cross-disorder analysis. This is an important topic as we can use information on correlated traits to increase our power to detect genetic associations and yield more biologically accurate understanding of the disorders considered, helping us to understand pleiotropy, or the effect a gene has on a set of traits (Casale et al., 2015, Korte et al., 2012, Li et al., 2014, Zhou and Stephens, 2014). With rare variants having lower minor allele frequency, appearing infrequently in smaller populations, with weaker effects, having effectively increased our sample size can help increase signal (Li et al., 2014, Maier et al., 2015).

Currently, three main approaches exist to analyze multi-trait data: meta-analysis and combined tests, dimension reduction, and multi-trait regression (Galesloot et al., 2014, Yang and Wang, 2012).

Meta analyses, including the work of Andreassen et al. (2013), Bolormaa et al. (2014), Yang et al. (2010) and Van der Sluis et al. (2013), combine summary statistics (e.g., test statistics or p-values) from single trait genome-wide association studies (GWAS). Because they work with summary statistic data, they can combine data from non-overlapping individuals – even from published results – and can analyze traits that do not follow the same distribution, e.g. continuous and dichotomous traits simultaneously (Bolormaa et al., 2014, Van der Sluis et al., 2013, Yang et al., 2010). They easily scale up to many traits and make less assumptions on genetic effects (with no canceling of opposing effects)

(Bolormaa et al., 2014, Van der Sluis et al., 2013, Yang et al., 2010). However, they lose power due to high multiple testing burden and by not incorporating coheritability and comorbidity information (Wang et al., 2015b).

Dimension reduction methods use principal component analysis (PCA) (e.g., Aschard et al. (2014), Klei et al. (2008)) and canonical correlation analysis (e.g., Ferreira and Purcell (2008)) to create a low-rank summary of the multiple phenotypes, trying to create a new response that explains the highest proportion of variability in phenotype or heritability (or covariability between phenotype and genotype). These approaches directly incorporate the correlations between phenotypes (Aschard et al., 2014), but are not easily interpretable, looking at linear combinations of traits rather than the actual traits. They also tend to focus on Gaussian distributed traits.

The last class, multi-trait regression often uses a random effects approach to directly model variance/covariance of genotypes between and within traits. Many multivariate linear mixed effects models (mLMMs) have been proposed to estimate coheritability and genetic correlation as a surrogate for pleiotropy (e.g., Korte et al. (2012), Lee et al. (2012b), Loh et al. (2015), Vattikuti et al. (2012)), or to predict genetic risk (e.g., Maier et al. (2015)). They have also been used to test for genetic association (e.g., Casale et al. (2015), Zhou and Stephens (2014)). These models are interpretable and successfully borrow information between traits, but assume normally distributed, or linearized, phenotype data. Other multi-trait regression models have also been proposed, e.g., looking at a functional linear regression for continuous gene-location incorporation (Wang et al., 2015b), using bivariate ridge regression for whole-genome genetic risk prediction (Li et al., 2014), or, finally, using similarity or kernel frameworks (Broadaway et al., 2016, Maity et al., 2012, Wei and Lu, 2015). These approaches provide additional flexibility, looking at how trait similarity relates to genotype similarity, but doesn't allow us to pinpoint genes with rare variants that are significantly associated with at least one disorder.

As in Maity et al. (2012), the kernel machine framework can be used to model complex associations between the rare variants from multiple genes and multiple disorders. Used in conjunction

with group lasso, kernels can be very useful for performing variable selection to detect which variants are likely associated with at least one disorder, combining information in a direct but flexible manner. Using a fastKM approximation makes this approach feasible and efficient.

1.3 Rare Variant Prioritization via Protein Structure

Since rare variants can be so difficult to identify individually, rare variant association tests usually require borrowing of signal - from other rare variants, or from additional genetic annotation information. When causal variant prioritization is of interest, however, straight aggregation of information over all rare variants within a variant set is not helpful, as it does not allow you to detect where the signal is likely coming from. Two main classes of genetic variant localization (prioritization) exist in the literature: penalization and most promising subset approaches.

Penalization methods use regularization to incorporate information from multiple genetic variants without forcing collapsing, shrinking estimates of noncausal loci or groups of loci. Examples include the work of Larson and Schaid (2014), Xu et al. (2012), and Zhou et al. (2011, 2010), examining the effects of group and combined penalties on power to detect likely causal rare variants. These methods are powerful but do not perform explicit association testing, and require sufficient sequencing information.

Most promising subset methodology, on the other hand, looks at clustering variants and trying to identify which clusters are most likely to be causal. They are based on the assumption that causal variants are likely to be close together in some dimension, be it on the 2 dimensional sequence, 3 dimensional structure, or functional domain (Fier et al., 2012, Ionita-Laza et al., 2012, Larson and Schaid, 2014, Yue et al., 2010). Some (e.g., Ionita-Laza et al. (2012), Kulldorff (1997)) create windows of variants along DNA sequence, assuming causal variants have the same effect on the trait of interest, with decreasing power for decreasing window and gene size, and decreasing variability within a gene (Ionita-Laza et al., 2012). Others (e.g., Fier et al. (2012), Lin (2014), Schaid et al. (2013), Tango

(2010)) have used kernel approaches to perform spatial clustering of variants, but focus on DNA sequence location to cluster variants. They also do not prioritize variants or calculate variant-level significance.

The kernel machine framework can be used here as well to incorporate biostructural information, e.g. location in the protein tertiary folding space, to perform local variant-level tests while still borrowing information from neighboring variants. Local kernels can be defined for each variant to determine the genetic similarity between individuals for a variant and those nearby on the 3-dimensional protein space, weighting on Euclidean distance from each variant, providing a local kernel machine framework. This is motivated by the idea that variants nearby on the protein tertiary space are more likely to behave similarly, e.g. due to binding domains or hydrophobic/hydrophilic nature of a region (Song et al., 2012).

1.4 Summary of Dissertation

The rest of the dissertation proceeds as follows: in chapter 2, we propose the fast kernel machine “fastKM” method for approximating nuisance effect kernels with low-dimensional fixed effects terms, showing how it can be used to facilitate analysis of multiple kernel models (e.g., GxE analysis) in an efficient manner. We demonstrate its utility in a simulation study, showing it is as powerful as an EM-based approach for estimating nuisance variance components, but much faster for continuous and binary traits, and makes the analysis of survival traits feasible. We further apply the fastKM method to the Vitamin Intervention for Stroke Prevention clinical trial to understand the effects of gene-by-vitamin regimen and gene-by-patient age on recurrent stroke risk and change in homocysteine level, respectively. Chapter 2 has been previously published in *Genetic Epidemiology* (Marceau et al., 2015).

In chapter 3, we propose the related “fastLasso” method for performing gene selection in cross-disorder analysis. We show how the fastKM framework can be used in conjunction with group lasso

to provide an efficient way to analyze genetic effects on multiple traits simultaneously, improving our power to detect genes associated with at least one disorder and flexibility to include different subjects for binary and continuous traits. We perform a simulation study to compare this cross-disorder analysis with the union of single disorder analyses, showing it increases the true positive rate while keeping a similar average false positive rate.

Finally, in chapter 4 we further examine the notion of prioritization of causal variants. We define a framework for incorporating protein tertiary structure into a kernel machine modeling approach to create powerful local (variant-level) kernel tests for rare variant association. We perform a simulation study, showing how our method is more powerful than single variant score tests and the sequence-based scan statistic, while maintaining a reasonable type I error for binary and continuous traits. We further apply our method to the Action to Control Cardiovascular Risk in Diabetes (ACCORD) clinical trial, prioritizing rare variants associated with lowering low-density lipoprotein (LDL), finding three new promising variants which appear to fall within or near to the *PCSK9-LDLR* protein-binding domain.

Chapter 2

A Fast Multiple-Kernel Method with Applications to Detect Gene-Environment Interaction

Rachel Marceau, Wenbin Lu, Shannon Holloway, Michèle M. Sale, Bradford B. Worrall, Stephen R. Williams, Fang-Chi Hsu, Jung-Ying Tzeng

*previously published in *Genetic Epidemiology* (Marceau et al., 2015)

2.1 Abstract

Kernel machine (KM) models are a powerful tool for exploring associations between sets of genetic variants and complex traits. Although most KM methods use a single kernel function to assess the marginal effect of a variable set, KM analyses involving multiple kernels have become increasingly popular. Multikernel analysis allows researchers to study more complex problems, such as assessing gene-gene or gene-environment interactions, incorporating variance-component based methods for population substructure into rare-variant association testing, and assessing the conditional

effects of a variable set adjusting for other variable sets. The KM framework is robust, powerful, and provides efficient dimension reduction for multifactor analyses, but requires the estimation of high dimensional nuisance parameters. Traditional estimation techniques, including regularization and the “expectation maximization (EM)” algorithm, have a large computational cost and are not scalable to large sample sizes needed for rare variant analysis. Therefore, under the context of gene-environment interaction, we propose a computationally efficient and statistically rigorous “fastKM” algorithm for multikernel analysis that is based on a low-rank approximation to the nuisance effect kernel matrices. Our algorithm is applicable to various trait types (e.g., continuous, binary, and survival traits) and can be implemented using any existing single-kernel analysis software. Through extensive simulation studies, we show that our algorithm has similar performance to an EM-based KM approach for quantitative traits while running much faster. We also apply our method to the Vitamin Intervention for Stroke Prevention (VISP) clinical trial, examining gene-by-vitamin effects on recurrent stroke risk and gene-by-age effects on change in homocysteine level.

2.2 Key Words

multiple-kernel analysis; kernel machine regression; exon level association test; gene-environment interaction; gene-gene interactions

2.3 Introduction

Kernel machine (KM) based approaches (Kwee et al., 2008, Lin et al., 2011, Liu et al., 2008, 2007, Wu et al., 2010, 2011) provide a powerful and popular strategy for evaluating associations between a set of genetic variants and complex traits of various types. The KM method uses a kernel function to quantify the pairwise genetic similarity for individuals based on multiple genetic variants; it then assesses the gene-trait association by examining if the genetic similarity of a pair of individuals is associated with their trait similarity (Lin et al., 2011, Wu et al., 2013). Many other marker-set

methods, for example, variance component tests (Goeman et al., 2004, Lin et al., 2013) and similarity regressions (Tzeng et al., 2014, 2009, 2011, Wang et al., 2014, Zhao et al., 2015), are closely related to KM tests under a random effects model. Therefore, although this paper primarily focuses on KM methods, the proposed approach and discussions are applicable to other variance component and similarity-based tests.

While the popular KM methods, for example, the Sequence Kernel Association Test (SKAT) (Wu et al., 2011), focus on single-kernel analysis (i.e., using one kernel function to model a single variable set), analyses involving multiple kernels (i.e., using separate kernels to simultaneously model multiple variable sets) are also frequently encountered in genomic research. Multikernel approaches include tests for gene-environment (G×E) interactions (Lin et al., 2013, Tzeng et al., 2011, Wang et al., 2015d, Zhao et al., 2015), tests for gene-gene interactions (Larson and Schaid, 2013, Wang et al., 2014), conditional tests for evaluating the effect of a single variable set adjusting for other variable sets (Pang et al., 2014, Wang et al., 2015c,d), and SKAT analysis coupled with a variance component method (Kang et al., 2010) to account for population substructure. The number of explanatory variables in these analyses is much higher than in a single marker-set analysis. In these situations, KM offers efficient dimension reduction and yield higher power to evaluate the effects of interest compared to other alternatives, such as burden-based methods (e.g., Madsen and Browning (2009), Price et al. (2010)). In addition, KM methods can model nonlinear/non-additive effects, accommodate variables with different direction and magnitude of effects, and are more robust than burden-based methods because they impose fewer assumptions on the underlying effects. The latter is particularly important in multikernel analysis — for example, a KM G×E test is more robust against misspecification of the main effects of G and E than the burden-based G×E test (Wang et al., 2015d).

However, the merits of KM approaches come with high computational cost for multikernel analyses, which substantially limits their practical utility. In a multikernel model, each effect is modeled via a $n \times n$ kernel matrix and an n -dimensional parameter vector (where n is the number

of individuals). Computing the test statistic to evaluate the effect of interest in a multikernel analysis requires estimation of at least one set of n -dimensional nuisance parameters. For example, performing a KM G×E test, even under the null hypothesis of no G×E effects, requires the estimation of nuisance genetic main effects. Current attempts to overcome the dimensionality challenges include treating the n -dimensional parameters as random and using the EM algorithm to estimate its variance component (Tzeng et al., 2011, Wang et al., 2014, 2015d, Zhao et al., 2015), or imposing penalization on these parameters (Lin et al., 2013). While both techniques have proven to be valid, the estimation procedures are usually phenotype-specific (e.g., the algorithms developed for quantitative traits cannot be applied to binary or survival traits) and computationally intensive (e.g., requiring the inversion of a n -dimensional matrix at each iteration of the EM algorithm or the tuning of a regularization parameter), making them not scalable to the large samples considered in rare variant studies.

Using KM tests for G×E interactions as an example, we illustrate our solution to resolve these computational challenges: a computationally efficient and statistically rigorous algorithm for performing KM tests in multikernel analyses. Our algorithm is motivated by the fact that the $n \times n$ kernel matrix is often not full rank — its rank is generally much less than the minimum of the number of individuals and the number of variables (e.g., SNPs) in a variable set. Thus, by decomposing the kernel matrices of nuisance effects, we can reduce the dimensionality to a manageable size so that a random-effect treatment or penalization is not necessary, and consequently a fixed effect null model can be fit to estimate nuisance parameters. The proposed method is fast, scalable to larger n , and applicable to a variety of trait types; most importantly, it can be implemented using any existing software for single KM analysis. For example, our algorithm would allow one to perform a KM G×E test using the existing software for main effect KM tests, such as SKAT.

We explore the performance of our method through an in-depth simulation study for quantitative, binary, and survival traits. We also apply our method to the Vitamin Intervention for Stroke Prevention (VISP) clinical trial (Toole et al., 2004). Focusing on the nine candidate genes within

the homocysteine pathway, we examine the gene-by-vitamin effects on recurrent stroke risk (Hsu et al., 2011, Tzeng et al., 2014) and gene-by-age effects on change in homocysteine level (Tzeng et al., 2011). In such, we are able to see the true flexibility and unifying features of our method, in its ability to perform multikernel analysis for different trait types in a computationally efficient manner.

2.4 Materials and Methods

2.4.1 The KM Model for G x E Interactions

Consider a study with n individuals. For individual i , $i = 1, \dots, n$, let Y_i denote the phenotype of interest and $G_i = (g_{i1}, \dots, g_{iM})^T$ denote the genetic markers. For now we assume an additive genetic effect where g_{im} is the count of minor alleles that individual i has at marker m , $m = 1, \dots, M$, though it is straightforward to extend this to recessive or dominant modes of inheritance. In addition, let $X_i = (1, X_{i1}, \dots, X_{iq})^T$ denote the baseline covariates that have no interaction with genetic markers and E_i the covariate interacting with genetic markers. For simplicity, we assume that E_i is a scalar.

We consider a generalized linear model for Y_i

$$g(\mu_i) = X_i^T \beta_X + E_i \beta_E + h_G(G_i) + h_{GE}(G_i, E_i), \quad (2.1)$$

where $\mu_i = E(Y_i | X_i, E_i, G_i)$, $g(\mu_i)$ is a canonical link function, for example, $g(\mu_i) = \mu_i$ for quantitative traits and $g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$ for binary traits, $h_G(\cdot)$ and $h_{GE}(\cdot)$ are two nonparametric smooth functions representing the main effect of genetic markers (i.e., G effect) and interaction effect between the E covariate and genetic markers (i.e., G×E effect). Model (1) can also be expressed in matrix notation as

$$g(\mu) = X \beta_X + E \beta_E + h_G(G) + h_{GE}(G, E), \quad (2.2)$$

where, for example, $X = (X_1, \dots, X_n)^T$, $G = (G_1, \dots, G_n)^T$, $E = (E_1, \dots, E_n)^T$, $g(\mu) = (g(\mu_1), \dots, g(\mu_n))^T$,

and $h_G(G) = (h_G(G_1), \dots, h_G(G_n))^T$.

Under this model, we can test for G×E interaction using the null hypothesis $H_0 : h_{GE}(\cdot) = 0$. Since $h_G(\cdot)$ and $h_{GE}(\cdot)$ are smooth functions that lie in a Hilbert space, by the representer theorem we can write $h_G(G)$ and $h_{GE}(G, E)$ in dual form expressions (Kimeldorf and Wahba, 1971): $h_G(G) = K_G \alpha_G$ and $h_{GE}(G, E) = K_{GE} \alpha_{GE}$, where $K_G = \{K_G(G_i, G_j) : 1 \leq i, j \leq n\}$ is an $n \times n$ kernel matrix for the G effect, $K_{GE} = \{K_{GE}(\{G_i, E_i\}, \{G_j, E_j\}) : 1 \leq i, j \leq n\}$ is an $n \times n$ kernel matrix for the G×E effect, and α_G and α_{GE} are $n \times 1$ vectors of unknown parameters. One commonly adopted kernel function for the G main effect is the weighted identity by state (IBS) kernel (Kwee et al., 2008, Wu et al., 2010), that is, $K_G^{IBS}(G_i, G_j) = \frac{\sum_{m=1}^M w_m \{2I(g_{im}=g_{jm}) + I(|g_{im}-g_{jm}|=1)\}}{\sum_{m=1}^M w_m}$. The IBS kernel quantifies genetic similarity using the weighted average number of alleles for which two individuals have in common in the marker set. The weights w_m 's are prespecified to upweight or downweight a variant based on certain features. For example, one can weight against the minor allele frequency of marker m so as to upweight similarities that are contributed by rare variants. Given the genetic main effect kernel K_G , one possible way to construct the interaction kernel K_{GE} is to take the element-wise product of the genetic main effect kernel and the environmental kernel K_E , as described in Larson and Schaid (2013) and Wang et al. (2015d). When the environmental covariate is a scalar, this simplifies to $K_{GE} = D_E K_G D_E$, where D_E is a diagonal matrix with elements E (Tzeng et al., 2011). However, caution must be taken when using this direct product kernel to avoid duplicating the main effect terms in the G×E kernel (Wang et al., 2015d).

The smooth functions $h_{GE}(G, E)$ can be viewed as random effects (Liu et al., 2008, 2007) and modeled through a multivariate normal distribution with mean zero and variance-covariance $\tau_{GE} K_{GE}$, that is, $h_{GE}(G, E) \sim N(0, \tau_{GE} K_{GE})$. This is equivalent to $\alpha_{GE} \sim N(0, \tau_{GE} K_{GE}^{-1})$ because $h_{GE}(G, E) = K_{GE} \alpha_{GE}$. Using this representation, testing $H_0 : h_{GE}(\cdot) = 0$ is equivalent to testing the null hypothesis $H_0 : \tau_{GE} = 0$ via a variance component score test (Liu et al., 2008, 2007).

Lin (1997) showed that the variance component score test is locally most powerful for testing the genetic main effect and only requires fitting the model under the null hypothesis, i.e., a standard

generalized linear model. For interaction tests however, fitting the null model requires estimation of the nonparametric function $h_G(\cdot)$. There are two main approaches for fitting the null model: the first strategy treats $h_G(G)$ as random effects following a multivariate normal distribution with mean zero and variance-covariance $\tau_G K_G$, and uses an EM algorithm to estimate the nuisance variance component τ_G (e.g., Tzeng et al. (2011), Wang et al. (2014, 2015d), Zhao et al. (2015)). The second strategy uses penalization techniques, such as ridge regression, to estimate the $n \times 1$ vector of parameters α_G (e.g., Lin et al. (2013)). Both strategies, however, are difficult computationally. The random effect approach is time consuming due to inverting an $n \times n$ matrix at each iteration of the EM algorithm, and has difficulties with estimation on the boundary of the parameter space. On the other hand, penalization methods require selection of a proper tuning parameter.

2.4.2 FastKM Test for G x E Interactions

We propose to take advantage of the low-rank structure of the kernel matrix K_G to enhance computational efficiency. Clearly, the weighted IBS kernel matrix is almost never full rank—typically, $rank(K_G) \ll \min(n, M)$ (i.e., less than the number of individuals and the number of markers of interest). Since K_G is a symmetric matrix, it can be decomposed using eigendecomposition as $K_G = Q\Lambda Q^T$, where Q is the matrix of eigenvectors of K_G and Λ is a diagonal matrix of eigenvalues of K_G . Removing the near-zero eigenvalues or taking only the leading eigenvalues that capture a high percentage of the total variation results in a low-rank decomposition $K_G = Z_{n \times r} Z_{r \times n}^T$, where $r \ll n$ is the number of positive eigenvalues kept. The null model, then, reduces to the form

$$g(\mu) = X\beta_X + E\beta_E + ZZ^T\alpha_G \triangleq X\beta_X + E\beta_E + Z\gamma. \quad (2.3)$$

where $\gamma = Z^T\alpha_G$. Model (2.3), referred to as the fastKM null model, is a standard GLM with low dimensional parameters. The fastKM null model can be rewritten in terms of an augmented design matrix $A = (X, E, Z) \equiv (A_1, \dots, A_n)^T$ and corresponding parameter vector $\theta = (\beta_X^T, \beta_E^T, \gamma^T)^T$ as $g(\mu) =$

$A\theta$; the parameter θ can be directly estimated by the maximum likelihood estimation using standard software. In the same spirit, we can rewrite Model (2.2) and obtain the following fastKM model: $g(\mu) = X\beta_X + E\beta_E + Z\gamma + h_{GE}(G, E)$. We then construct the score test statistic for $H_0: \tau_{GE} = 0$ as $U_n = n^{-1}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)K_{GE}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)^T$, where $\hat{\theta}$ is the maximum likelihood estimator of θ under the null and $\hat{\epsilon}_i = Y_i - g^{-1}(A_i^T \hat{\theta})$ ($i = 1, \dots, n$) are fitted residuals.

Note that our score test statistic shares the same form of the KM score test statistic for genetic main effects (except that K_{GE} is involved instead of K_G). Therefore, the KM G×E test can be conducted using any existing testing software for genetic main effects, such as SKAT (Wu et al., 2011), providing the augmented design matrix A and the G×E kernel K_{GE} as input. Moreover, as with main effect KM tests, the limiting distribution of the fastKM test statistic under the null can also be represented as $\sum_{i=1}^d \lambda_i \chi_{1,i}^2$, where $\chi_{1,i}^2$ ($i = 1, \dots, d$) are independent Chi-squared random variables with one degree of freedom and weights λ_i 's are the positive eigenvalues of a nonnegative definite matrix Σ . Here matrix Σ is the variance-covariance matrix of the limiting distribution of $n^{-1/2} \sum_{i=1}^n \hat{\epsilon}_i Z_{GE,i}$, where $Z_{GE,i}$ is the i -th row of matrix Z_{GE} with $Z_{GE}Z_{GE}^T = K_{GE}$ (Lin, 1997, Zhang and Lin, 2003). The associated P -value can be calculated using moment matching method (Duchesne and Lafaye De Micheaux, 2010), the Davies method (Davies, 1980), or empirically using resampling techniques. For a resampling method, one can generate many sets of independent Chi-squared random variables with one degree of freedom and calculate $U_{n,b} = \sum_{i=1}^d \hat{\lambda}_i \chi_{1,i}^2$, $b = 1, \dots, B$, where $\hat{\lambda}_i$ is a consistent estimator of λ_i . Then the estimated P -value is $B^{-1} \sum_{b=1}^B I(U_n < U_{n,b})$.

2.4.3 Extension to Survival Traits

As of yet, no methods have been developed for testing marker-set G×E effects for survival traits. Our fastKM method, however, can be naturally extended to include survival traits. For individual i , let T_i denote the event time of interest and C_i the censoring time. Further, define $\tilde{T}_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. As usual, we assume $T_i \perp C_i$ given X_i , E_i and G_i . For simplicity, we consider the proportional hazards model, though other survival models can also be used following the derivations

in Tzeng et al. (2014). Under the null, we fit the proportional hazards model with the augmented covariates A_i as defined previously. Let $\hat{\theta}$ denote the maximum partial likelihood estimator of θ and $\hat{\Lambda}(\cdot)$ denote the Breslow estimator of the baseline cumulative hazard function. Our proposed test statistic for the null hypothesis $H_0 : h_{GE}(\cdot) = 0$ is given by $U_n = n^{-1}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)K_{GE}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)^T$, where $\hat{\epsilon}_i = \delta_i - \hat{\Lambda}(\tilde{T}_i)\exp(A_i^T \hat{\theta})$ ($i = 1, \dots, n$) is a martingale residual. The fastKM test statistic again shares the same form as the score test statistic for the genetic main effect tests discussed in Lin et al. (2011) and Tzeng et al. (2014), with our augmented design matrix A as the covariate design matrix and $G \times E$ kernel K_{GE} as the kernel matrix. Consequently, the fastKM test statistic and its P -value can be obtained by using the existing software for KM main-effect tests with survival traits (e.g., Lin et al. (2011), Tzeng et al. (2014)).

2.4.4 Implementation Detail

In practice, the low rank decomposition may contain several near-zero eigenvalues, which would result in the Z matrix (and hence also the augmented design matrix A) containing multiple near-zero columns and lead to unstable parameter estimation. We suggest performing kernel principal component analysis (kPCA) to further reduce dimensionality and improve stability (Cai et al., 2011, Schölkopf et al., 1998). In practice, choosing to keep the top eigenvalues that collectively explain $p = 95\%$ or 99% of the total variability can give good empirical results, especially for continuous traits. However, when the variants of interest are rare and there are many near-zero eigenvalues, a smaller p may be necessary for achieving estimation stability with binary and survival traits. In our numerical studies (simulation and real data analysis), we identify a suitable $p\%$ by starting with a high percentage (e.g., 99%) and then gradually reducing p until the null model fits reasonably well (e.g., no warning messages in the GLM fits or no extremely large coefficients). This is also the rule we implement in the fastKM R function.

2.4.5 Simulation Study

2.4.5.1 Data Generation

We generated a set of 10,000 haplotypes using the COSI software of Schaffner et al. (2005) with a coalescent model mimicking the linkage disequilibrium and population history of the European population. We then formed a marker-set of M rare (minor allele frequency <0.05) loci, of which the first 40 were considered causal. We generated a sample of n individual genotypes by randomly sampling two haplotypes with replacement. We set $M = 100$ loci and $n = 5000$ individuals. We also considered $M = 200$ and $n = 1000$ in a subset of scenarios to investigate the impact of M and n on the performance of fastKM. We considered a single environmental covariate E that is either continuous (generated from a Normal(0,1) distribution) or binary (generated from a Bernoulli(0.5) distribution). We assumed no confounding covariates in the simulations.

We evaluate the performance of the fastKM G×E tests using data generated from a fixed effect model, where the genetic main and interaction effects depend on mutational burden. Specifically, given each individual's environmental covariate E and genotypes, define $\eta(E_i, G_i) = \gamma_0 + \gamma_E E_i + \gamma_G \sum_{m=1}^{M_c} g_{im} + \gamma_{GE} E_i (\sum_{m=1}^{M_c} g_{im})$, where $M_c = 40$ represents the number of causal variants. We considered three types of traits: quantitative, binary, and survival. Quantitative responses were generated from the model $Y_i = \eta(E_i, G_i) + \epsilon$ where $\epsilon \sim N(0, 1)$. Binary responses were generated in a case-control framework from a Bernoulli distribution with $P(Y_i = 1 | E_i, G_i) = \frac{e^{\eta(E_i, G_i)}}{1 + e^{\eta(E_i, G_i)}}$. Finally, survival traits were generated from a Cox proportional hazards model: $\log(T_i) = -\eta(E_i, G_i) + \epsilon_i^*$, where $\gamma_0 = 0$ and ϵ_i^* follows a standard extreme value distribution. For survival traits, censoring times were generated from a uniform distribution on $[0, c]$, where c was chosen to yield censoring proportions of 15% and 40%.

2.4.5.2 Examining Power and Type I Error of FastKM

In each simulation scenario, we perform 2000 replicates for type I error analysis and 1000 replicates for power analysis. We compared our fastKM G×E test to the burden-based G×E test. For quantitative traits, we also compare with the traditional EM-based KM G×E test (referred to as “originalKM”). We calculated the IBS kernel for K_G as described in Tzeng et al. (2011) and set the variant-specific weight as $w_m = (1 - q_m)^{24}$ (Wu et al., 2011). For certain settings under quantitative traits, we also calculated the polynomial kernel (i.e., $K_G^{poly} = (1 + G_i^T G_j)^d$) with $d = 2$ and $d = 3$.

After performing the eigenvalue decomposition of K_G , we rounded those eigenvalues with magnitude $< 10^{-10}$ to zero, and kept the top eigenvalues to explain $p\%$ of the variability ($p\%$ kPCA); the value of p is chosen to retain the maximum amount of variation while still yielding stable GLM estimates.

For the fastKM algorithm, we fit the fastKM null model for each trait with linear predictor $\eta(A) = \gamma_0 + \gamma_A A$ where A is the augmented design matrix composed of a standardized covariate and the reduced Z matrix. P -values were calculated using the Davies method (Davies, 1980) as implemented in Duchesne and Lafaye De Micheaux (2010). For the burden-based test, we fit the models with the covariate effects $\eta(G_i, E_i) = \gamma_0 + \gamma_E E_i + \gamma_G \sum_{m=1}^M g_{m,i} w_m + \gamma_{GE} E_i \sum_{m=1}^M g_{m,i} w_m$, and conducted a Wald test of $H_0 : \gamma_{GE} = 0$. Finally, for the originalKM method for quantitative traits, we used the G×E test of Tzeng et al. (2011), which estimates the variance component τ_G via an EM algorithm.

2.4.6 Application to VISP Study

We apply our method to the VISP clinical trial data (Toole et al., 2004). The VISP trial was a multi-center study in which 3680 ischemic stroke patients, with informed consent, were randomly assigned to one of two vitamin dosage arms and were followed up until they experienced a subsequent stroke, or alternatively suffered from a myocardial infarction or death. Two vitamin dosages were

administered: the low-dose arm consisted of 200 μg B_6 , 6 μg B_{12} , and 20 μg folic acid; the high-dose arm consisted of 25 mg B_6 , 0.4 mg B_{12} , and 2.5 mg folic acid. The genetic substudy enrolled and consented 2,164 individuals (Hsu et al., 2011, Tzeng et al., 2014). Following the studies of Hsu et al. (2011) and Tzeng et al. (2014), we focus on the nine candidate genes within the homocysteine (Hcy) pathway: *BHMT*, *BHMT2*, *CBS*, *CTH*, *MTHFR*, *MTR*, *MTRR*, *TCN1*, and *TCN2*, treating each as a recessive gene.

Our primary interest was to test whether dosage level significantly interacts with any of the genes in the Hcy pathway to determine the time until subsequent stroke. Individuals were considered censored if they dropped out of the study or did not have another stroke before the end of the study. Approximately 91% of the patients were censored. Hsu et al. (2011) and Tzeng et al. (2014) previously examined this problem, respectively using single SNP and gene-level genetic main effect tests stratified by treatment dosage. We extend the analysis by conducting a G×E aggregation test with intervention group as the environmental variable.

To perform the G×E test, we fit a Cox proportional hazards model looking at the effect of gene × intervention interaction on time until stroke, adjusting for age, sex, and race. We assume any missingness is at random and exclude all individuals with missing covariate values or genotype and all loci with too much missingness. The final data set consisted of 1914 total subjects and 74 loci. This data consisted of a mixture of common and rare variants, so the weight $w_m = q_m^{-3/4}$ was used in calculating the weighted IBS kernel (Pongpanich et al., 2011, Tzeng et al., 2014). For quantitative traits, we also considered the polynomial kernel with $d = 2$. To improve stability and efficiency, we consider dimension reduction via kernel PCA. We find 85% kPCA to be sufficient reduction.

As a secondary research question, we consider whether a patient's age interacts with their genotype in affecting the change in homocysteine during a 2 hour fasting methionine load test performed at baseline (prerandomization), similar to Tzeng et al. (2011). The change in total Hcy was analyzed as both a continuous trait and a binary trait using the sample 90th percentile as a cut-off to dichotomize patients (i.e., the phenotype is 1 for individuals whose change in total Hcy

is in the top 10%, and 0 for all others). A 90% cut off has been used in the past as an indicator for possible hyperhomocysteinaemia (e.g., van der Griend et al. (2002)). Sex and race were included as covariates in both models. As with the primary outcome, we used weight $w_m = q_m^{-3/4}$, and used 95% kPCA for dimension reduction. P -values were calculated via Davies' method (Davies, 1980) as implemented in Duchesne and Lafaye De Micheaux (2010). The analysis of continuous outcomes was also compared to the originalKM method.

2.5 Results

2.5.1 Quantitative Traits

Performance Evaluation with IBS Kernel, $M = 100$ Loci and $n = 5000$ Individuals.

Figure 2.1 shows how fastKM and the originalKM method had very similar type I error levels for the G×E test. The type I error rates of both tests were close to the nominal level of 0.05 when the E variable was Gaussian distributed, but were slightly conservative when E was binary. The burden-based test appeared to only be valid for the null models with no genetic main effect (i.e., $\gamma_G = 0$); it had inflated type I error rates for all simulated scenarios with $\gamma_G > 0$, and the magnitude of inflation increased with γ_G . This is consistent with previous findings that fixed effects G×E tests, including burden-based tests, may have inflated type I error when the genetic main effect is modeled incorrectly (Voorman et al., 2011, Wang et al., 2015d).

Figure 2.2 shows that fastKM had almost identical power to originalKM for quantitative traits, quickly increasing with larger interaction sizes γ_{GE} . The burden-based tests had the lowest power among the three tests, despite having inflated type I error rates for non-zero main effects. The power of all three tests is reduced when the E variable was generated from a Bernoulli distribution.

Tables 2.1 and 2.2 show the computational time required for the different approaches' G×E tests for type I error analysis and power analysis, respectively. Computations were carried out on one processor of the IBM dual-Xeon (HS21 blade) computer nodes (2.66 GHz) with 4 GB RAM.

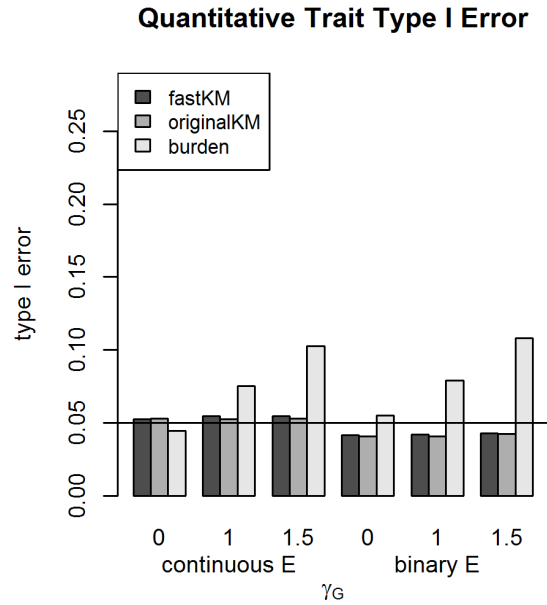


Figure 2.1 Type I error for fastKM, originalKM, and the weighted counting burden-based G×E test for quantitative traits with $M = 100$ loci, $n = 5,000$ individuals, and varying main effect parameter γ_G . Models with a continuous environmental E covariate are on the left and those with a binary environmental E covariate are on the right. The KM tests are based on the IBS kernel.

We found that the burden-based test was the quickest method, but this speed came at the cost of poor overall performance (i.e., inflated type I error and lower power). On the other hand, while fastKM and originalKM had very similar empirical performance, it is evident that fastKM was much more efficient, taking just over 2 min per run for a sample size of $n = 5000$ individuals regardless of simulation scenario. The originalKM method was slower, ranging from taking just over three times as long to over 10 times as long per run. It took the longest when there was no genetic main effect, because the EM algorithm has extremely slow convergence when $\tau_G = 0$, i.e., when the nuisance variance component is at the boundary of the parameter space.

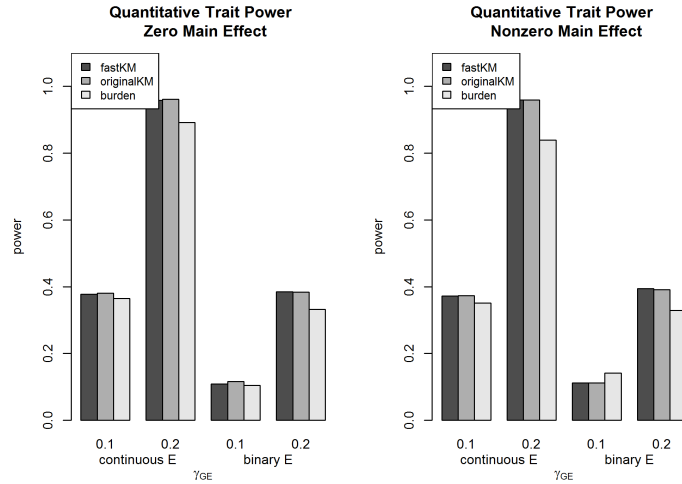


Figure 2.2 Power for fastKM, originalKM, and the weighted counting burden-based $G \times E$ test for quantitative traits with $M = 100$ loci and $n = 5,000$ individuals over varying interaction effect sizes γ_{GE} . The left panel shows the results of no genetic main effect (i.e., $\gamma_G = 0$) and the right panels shows the results of nonzero main effect (i.e., $\gamma_G = 1$). For each plot, continuous E covariates are on the left and binary E covariates are on the right. The KM tests are based on the IBS kernel.

Table 2.1 Average run time in minutes (and corresponding standard error) for quantitative traits when the $G \times E$ effect is zero with a sample size of $n = 5,000$ individuals.

Covariate	Main Effect	FastKM	OriginalKM	Burden
Continuous	0	2.30 (0.011)	31.2 (0.348)	0.0037 (0.001)
	1	2.31 (0.009)	7.9 (0.023)	0.0022 (5×10^{-4})
	1.5	2.38 (0.018)	7.8 (0.023)	0.0028 (7×10^{-4})
Binary	0	2.10 (0.006)	31.5 (0.354)	0.0013 (2×10^{-5})
	1	2.09 (0.005)	7.5 (0.016)	0.0013 (2×10^{-5})
	1.5	2.14 (0.006)	7.4 (0.016)	0.0014 (6×10^{-5})

Table 2.2 Average run time in minutes (and corresponding standard error) for quantitative traits when the G×E effect is nonzero with a sample size of $n = 5,000$ individuals.

Covariate	Main Effect	Interaction	FastKM	OriginalKM	Burden
Continuous	0	0.1	2.32 (0.012)	29.6 (0.476)	0.0021 (5×10^{-5})
		0.2	2.33 (0.011)	29.1 (0.479)	0.0022 (5×10^{-5})
	1	0.1	2.62 (0.014)	7.8 (0.046)	0.0015 (1×10^{-5})
		0.2	2.52 (0.014)	7.8 (0.046)	0.0015 (1×10^{-5})
Binary	0	0.1	2.42 (0.011)	26.5 (0.488)	0.0021 (5×10^{-5})
		0.2	2.45 (0.010)	17.1 (0.386)	0.0021 (5×10^{-5})
	1	0.1	2.62 (0.013)	7.5 (0.046)	0.0015 (1×10^{-5})
		0.2	2.63 (0.013)	8.1 (0.049)	0.0015 (1×10^{-5})

Impact of Kernel Choices and (M, n) on Method Performance

Figures 2.3 (type I error rates) and 2.4 (power) show the impact of marker-set size and sample size, (M, n) , and of kernel selection on the performance of fastKM. In particular, we compare the performance of $(M = 200, n = 1000)$ to $(M = 100, n = 5000)$ under the IBS kernel and polynomial kernels with $d = 2$ and 3. By comparing the left panel of $(M = 100, n = 5,000)$ and right panel of $(M = 200, n = 1,000)$ in Figures 2.3 and 2.4, we see that, when sample size n is large and marker set size is relatively smaller (e.g., $M = 100, n = 5000$), fastKM performed similarly to the originalKM for all kernel types (i.e., IBS and polynomial with $d = 2$ and $d = 3$). When $M = 200$ and $n = 1,000$, while fastKM with the IBS kernel still has similar performance to originalKM, fastKM with the more complex polynomial kernels is slightly more conservative than originalKM and has reduced power. We observed that the IBS kernel tends to have slightly higher power than the polynomial kernel even with originalKM, which is not unexpected given the effect mechanism considered in the simulation.

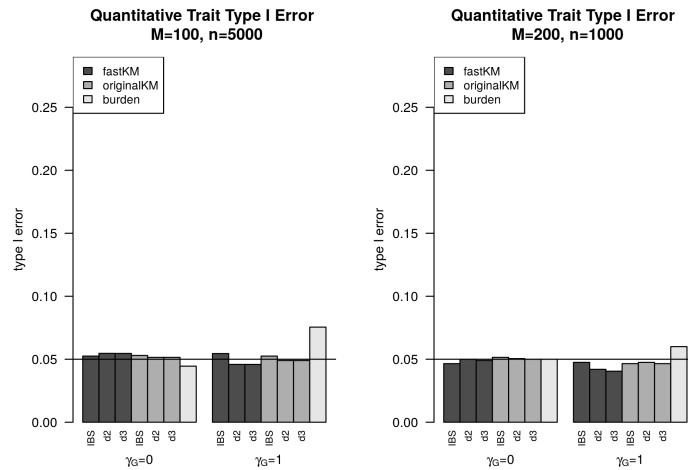


Figure 2.3 Type I error for fastKM, originalKM, and the weighted counting burden-based $G \times E$ test for quantitative traits with continuous environmental E covariate and varying main effect parameter γ_G . The left panel shows the results of $M = 100$ loci and $n = 5,000$ individuals. The right panel shows the results of $M = 200$ loci and $n = 1,000$ individuals. The KM tests are based on the IBS kernel and the polynomial kernels with $d = 2$ and 3.

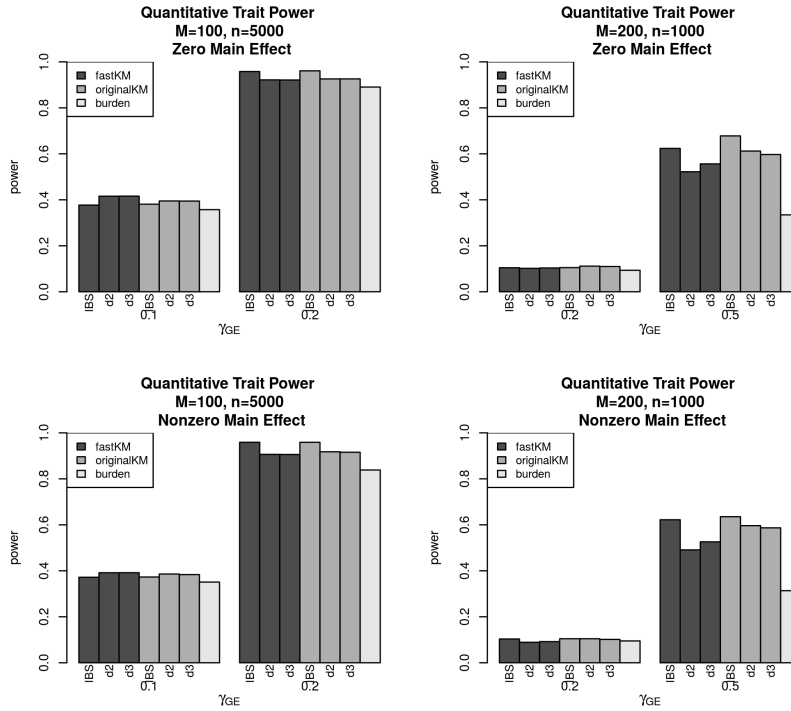


Figure 2.4 Power for fastKM, originalKM, and the weighted counting burden-based $G \times E$ test for quantitative traits with continuous E covariate over varying main effect size γ_G ($\gamma_G = 0$ for zero main effect, and $\gamma_G = 1$ for nonzero main effect) and interaction effect size γ_{GE} . The left panel shows the results of $M = 100$ loci and $n = 5,000$ individuals. The right panel shows the results of $M = 200$ loci and $n = 1,000$ individuals. The KM tests are based on the IBS kernel and the polynomial kernels with $d = 2$ and 3 .

2.5.2 Binary and Survival Traits

For binary traits and survival traits, we compared the performance of the fastKM $G \times E$ test with burden-based $G \times E$ tests. The relative performance of the two approaches was similar to what was observed in quantitative traits. The results of type I error analyses are shown in Figure 2.5 (for binary traits) and Figure 2.6 (for survival traits). The fastKM $G \times E$ tests had type I error rates around the nominal level, except in the case of binary traits with a continuous E variable, where the test was slightly conservative. The burden-based $G \times E$ test was valid for small genetic main effect sizes, but

its type I error rate increased with larger γ_G .

The binary trait power analysis (Figure 2.7) shows that fastKM was more powerful than the burden-based test, which is similar to the quantitative traits results. For survival traits (Figure 2.8), fastKM had similar or higher power compared to the burden-based test when the censoring proportion was low (i.e., 15%). The power difference between fastKM and burden-based G×E test became more obvious when the censoring proportion was 40%. Overall, when the trait is continuous, binary, or survival, fastKM is a valid test and has pretty good power that scales quickly with increasing interaction effect γ_{GE} .

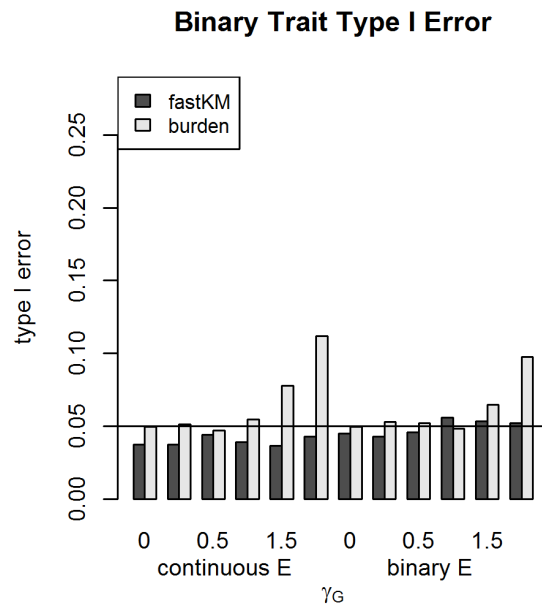


Figure 2.5 Type I error for fastKM and the weighted counting burden-based G×E test for binary traits for $M = 100$ loci, $n = 5,000$ individuals, and varying main effect parameter γ_G . Models where E is generated from a Gaussian distribution are displayed on the left, and those where E is from a Bernoulli distribution are on the right. The KM tests are based on the IBS kernel.

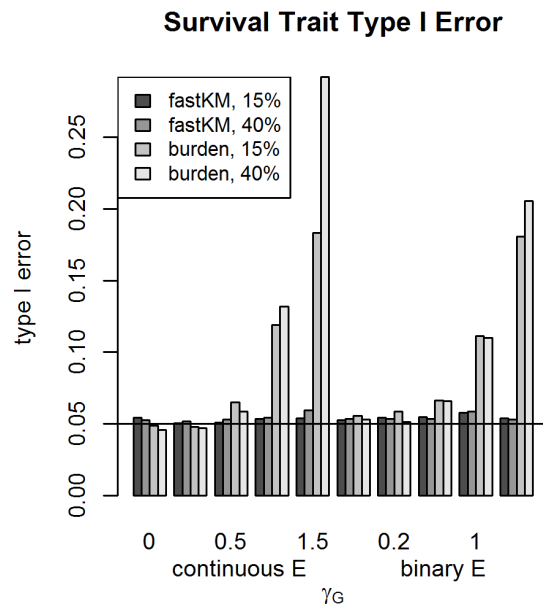


Figure 2.6 Type I error for survival traits for fastKM and the weighted counting burden-based $G \times E$ test with $M = 100$ loci, $n = 5,000$ individuals, $c=15\%$ and 40% censoring proportions, and varying main effect parameter γ_G . Models where E is generated from a Gaussian distribution are displayed on the left, and those where E is from a Bernoulli distribution are on the right. The KM tests are based on the IBS kernel.

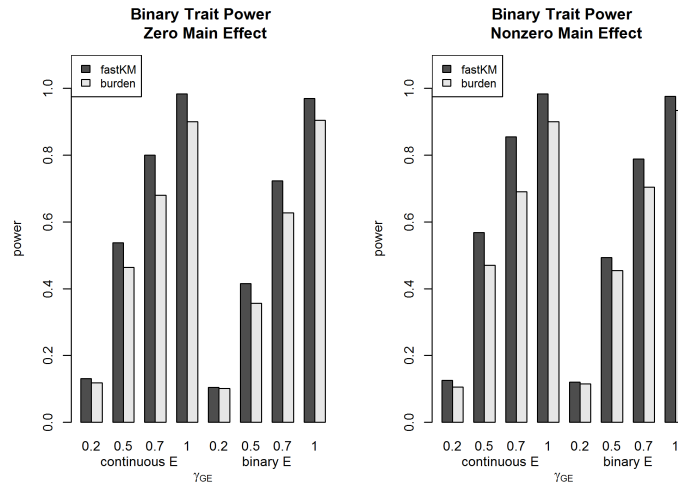


Figure 2.7 Power for fastKM and the weighted counting burden-based $G \times E$ test for binary traits with $M = 100$ loci and $n = 5,000$ individuals over varying interaction effect sizes γ_{GE} . The left panel shows the results of no genetic main effect (i.e., $\gamma_G = 0$) and the right panels shows the results of nonzero main effect (i.e., $\gamma_G = 1$). For each plot, continuous E covariates are on the left and binary E covariates are on the right. The KM tests are based on the IBS kernel.

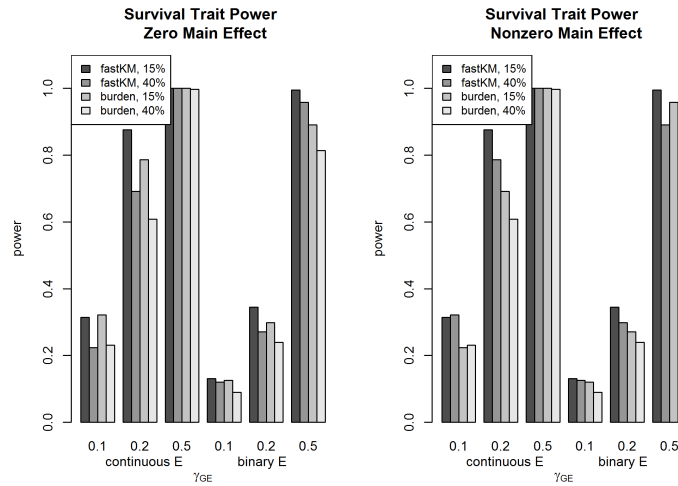


Figure 2.8 Power for fastKM and the weighted counting burden-based $G \times E$ test for survival traits with $M = 100$ loci and $n = 5,000$ individuals for varying interaction parameter γ_{GE} over two censoring proportions ($c=15\%$ and 40%). The left panel shows the results of no genetic main effect (i.e., $\gamma_G = 0$) and the right panels shows the results of nonzero main effect (i.e., $\gamma_G = 1$). For each plot, continuous E covariates are on the left and binary E covariates are on the right. The KM tests are based on the IBS kernel.

2.5.3 VISP Study

We first performed a survival analysis of gene-by-intervention interaction on time until subsequent stroke (Table 2.3) using fastKM. We found no significance at the Bonferroni correction threshold of $0.05/9=0.0056$. However, the two genes with the lowest *P*-values, *TCN2* (*P*-value 0.0408) and *CTH* (*P*-value 0.0171), were also the most significant found in a previous study (Tzeng et al., 2014). Tzeng et al. (2014) performed a stratified analysis by vitamin intervention and found these two genes had the smallest *P*-values in the low-dose intervention group, using a KM genetic main effect test. Both *TCN2* and *CTH* are members of the folate one-carbon metabolism (FOCM) pathway. The FOCM mediates many key biological processes in the cell, including methionine metabolism, Hcy synthesis, B-vitamin utilization, and provision of de novo cellular methyl group availability through conversion of S-adenosyl-methionine (SAM) to S-adenosyl-homocysteine (SAH). *TCN2* is the primary plasma facilitator of cellular uptake of B12 (Seetharam et al., 1999), while *CTH* is responsible for converting cystathionine into cysteine. Although they did not meet or exceed our Bonferroni corrected *P*-value, previous work shows that *TCN2* was found to be associated with recurrent stroke in VISP participants randomized to the low-dose B-vitamin arm of the trial (Hsu et al., 2011). Related to these findings, *CTH* has been found to be associated with Hcy levels (Wang et al., 2004), a well-recognized risk factor for stroke, and *TCN2* has been associated with Hcy levels in healthy individuals (Lievers et al., 2002) and among subjects with low B12 (Stanislawska-Sachadyn et al., 2010).

In the analyses of gene \times age effects on Hcy change treated as continuous trait, fastKM and originalKM had similar *P*-values. Both tests found *MTR* to significantly interact with age to influence change in homocysteine when treated as a continuous trait, even after correcting for multiple testing. A similar result was obtained when treating the change in Hcy as binary. *MTR* is a member of the FOCM and is responsible for the methylation of Hcy in the resynthesis of methionine. As an essential component of the FOCM, the *MTR* enzyme is activated by *MTRR*, utilizing B12 as a necessary

component of the Hcy methylation reaction. Mutations in *MTR* have previously been identified as the underlying cause of multiple metabolic disorders including cases of hyperhomocysteinemia, as a dysfunctional *MTR* enzyme would lead to the inability to convert Hcy in to methionine (Mellman et al., 1979). Also, mutations in *MTRR* can lead to hyperhomocysteinemia through the inability to activate *MTR* (Leclerc et al., 1998, Rosenblatt et al., 1985, Schuh et al., 1984, Zavadáková et al., 2005). Through genome-wide association studies, a common *MTR* functional variant, A2756G (D919G) has been found to be associated with a modest, but significant, increase in plasma Hcy levels (Dekou et al., 2001, Harmon et al., 1999, Tsai et al., 2000, Wang et al., 1999). However, no consistent effect on risk of developing vascular disease has been found (Dekou et al., 2001, Hyndman et al., 2000).

Given the fact that Hcy levels rise with age (Nygård et al., 1995) and that *MTR* is essential in conversion of Hcy to methionine, it is plausible that the effect between change in Hcy and an inherited heterozygous mutation of *MTR* in combination with other FOCM mutations, including *MTR*, may change over time. Additionally, it must be noted that there are other clinical factors that should be taken into account, including kidney and liver function which are also known to diminish with age and are both related to B-vitamin utilization and FOCM function (Spence et al., 1999).

Table 2.3 *P*-values for the fastKM analyses of VISP study data, including (a) testing gene \times age interaction on post-methionine change in total Hcy, treating change as continuous, with a IBS kernel or polynomial kernel (d=2); (b) testing gene \times age interaction on post-methionine change in total Hcy, treating change as binary using the 90th sample percentile as a cut off; (c) testing gene \times intervention interaction on time to recurrent stroke

* *P*-values that are < 0.05.

Gene Name		<i>BHMT</i>	<i>BHMT2</i>	<i>CBS</i>	<i>CTH</i>	<i>MTHFR</i>	<i>MTR</i>	<i>MTRR</i>	<i>TCN1</i>	<i>TCN2</i>
Number of Loci		5	3	6	10	7	20	5	3	15
<u>Trait</u>	<u>Method</u>									
Hcy change (quantitative)	OriginalKM, IBS	0.9997	0.7762	0.8572	0.7882	0.2932	0.0009*	0.3904	0.3970	0.8396
	OriginalKM, d2	0.9089	0.9117	0.5616	0.4553	0.2410	0.0008*	0.5364	0.2447	0.8739
	FastKM, IBS	0.9998	0.7953	0.8373	0.8046	0.3163	0.0008*	0.4000	0.3940	0.8504
	FastKM, d2	0.8856	0.9351	0.5570	0.4700	0.2322	0.0007*	0.5648	0.2569	0.8673
Hcy change (binary)	FastKM, IBS	0.9675	0.5212	0.3893	0.9362	0.4229	0.0007*	0.3617	0.2703	0.2967
Time to stroke (survival)	FastKM, IBS	0.8436	0.6934	0.9106	0.0176*	0.2230	0.4918	0.0545	0.9973	0.0430*

2.6 Discussion

Although multikernel analyses are frequently encountered, the practical utility of existing approaches may be limited due to the computational cost and method complexity. The intensive computation makes the KM analysis unscalable to large samples. The method complexity, mainly arising from the estimation of nuisance variance components in the null model, has led the majority of multi-kernel approaches to focus on the analysis of quantitative traits. Although a few methods currently exist for binary traits (e.g. Lin et al. (2013), Zhao et al. (2015)), there is still a lack of multikernel methods for survival traits to the best of our knowledge.

In this work, we use the G×E interaction test to illustrate our solution to address these issues based on a low-rank approximation to the kernel matrix for genetic main effects. We demonstrate that the proposed low-rank fastKM framework, when coupled with the IBS kernel, can retain the power and validity of the robust G×E KM test based on random-effects models for quantitative traits. The fastKM method greatly enhances the feasibility of the robust G×E KM test in several aspects. First, fastKM speeds up computation — in some cases our algorithm is up to ten times faster than the robust G×E KM test. The reduction in computation time of the fastKM is expected to be even larger for binary and survival traits. This decrease in computation increases the scalability of multikernel approaches to larger sample sizes, as are required for rare variant analyses, and to a larger number of SNP sets, making whole genome interactive analysis much more feasible. Second, fastKM is applicable to general trait types, from continuous, binary, to survival traits. By creating an augmented covariate-genotype matrix, fastKM transforms the interaction test into a single kernel analysis framework, allowing one to perform an interaction test using any existing main effect testing software. Specifically, single kernel analysis softwares, for example, SKAT (Wu et al., 2011), require the input of a covariate design matrix and a kernel matrix, so one can perform fastKM by providing the augmented matrix as the required covariate design matrix and the kernel matrix for the effect to be tested (e.g., G×E) as the required kernel matrix. We provide the R functions that carry

out these fastKM steps on the authors' website <http://www4.stat.ncsu.edu/~sthollow/JYT/fastKM/>.

We note that a fundamental presumption of fastKM is that the kernel matrix of the genetic main effect can be well approximated by a low rank matrix decomposition; therefore, the null model can be fit using augmented covariates by including the leading components of the low rank matrix decomposition. The low-rank structure of the kernel matrix may depend on the choice of kernels, sample size (n) and the number of variants that are jointly analyzed (M). In our simulation studies, we found that fastKM with the IBS kernel performed appropriately with varying M and n . When fastKM is coupled with a more complex kernel (e.g., higher order of polynomial kernels), it would have appropriate performance when n is large relative to M (e.g., $M = 100$ and $n = 5000$). However, caution is needed when a complex kernel is applied with a moderate sample size relative to M (e.g., $M = 200$ and $n = 1,000$). Based on these findings, we would recommend to use fastKM with IBS kernel. If a complex kernel is needed and the sample size n is moderate relative to M , one may consider to perform the original KM method with EM algorithm or penalization, which tends to be feasible with moderate sample sizes.

In this paper, we focused our investigation of fastKM on rare variants. It would be of interest to examine if the findings can be extended to common variants. In our limited investigation (described in the supplementary information), we considered a randomly selected 50-SNP region from HapMap3 CEU data and generated 1,000 individuals. The results suggest fastKM with IBS kernel and polynomial kernel has similar performance to their EM counterpart. Nevertheless, further studies would be needed to fully understand the performance of fastKM when apply to common variant analyses.

The proposed low-rank KM framework has broad impact on KM modeling and beyond. It greatly enhances the computational efficiency of KM tests that contain multiple kernel components and involve high-dimensional nuisance parameters, e.g., the $G \times G$ kernel tests (Larson and Schaid, 2013) and the conditional kernel tests (Wang et al., 2015c,d). It can be generalized to study copy-number variants (CNVs), for example, to extend burden-based CNV tests (Raychaudhuri et al., 2010), which

simultaneously model multiple CNV features, to the framework of KM tests (e.g., testing for a CNV dosage effect while adjusting for length and gene interruption status) (Tzeng et al., 2015). In addition, it can also be extended to perform KM interaction tests for multivariate-phenotype analysis (e.g., Davenport et al. (2017), Maity et al. (2012)). Lastly, because KM has a generalized linear mixed model (GLMM) representation, our low-rank framework can also benefit other GLMM-equivalent methods, for example the GLMM-based $G \times E$ test (Lin et al., 2013) and the SimReg $G \times G$ and $G \times E$ tests (Tzeng et al., 2011, Wang et al., 2014, Zhao et al., 2015).

2.7 Acknowledgments

This work was partially supported by NIH grants 5T32GM081057 (to R.M.; PI Muse), R01 CA140632 (to W.L.), U01 HG005160 (to M.M.S., B.B.W., S.R.W., and F.C.H.; PIs M.M.S. and B.B.W.), R01 NS34447 (to M.M.S. and F.C.H.; PI Toole), R01 MH084022 (to J.Y.T.), and P01 CA142538 (to W.L., S.H., and J.Y.T.; PIs Kosorok, Davidian, Owzar).

2.8 Supplementary Materials

2.8.1 Common Variants based Simulations

In the common variant simulation, we obtained a haplotype population consisting of 234 phased haplotypes from chromosome 21 of the CEU (Utah residents with ancestry from northern and western Europe) samples in HapMap 3. We randomly select a 50-SNP region, where the minor allele frequencies (MAFs) range from 0.06 to 0.47 with median 0.24 and mean 0.27. We generated haplotypes for 1000 individuals by randomly sampling 1000 pairs of haplotypes with replacement from the 234 haplotypes under a Hardy-Weinberg equilibrium assumption. Then we select the 25th and the 30th SNPs as causal (MAF 0.23 and 0.42) and simulated the continuous trait values of each individual from $\text{Normal}(\mu, 1)$ with $\mu = 0.5 + 2 \times E + \gamma_{G_1} G_1 + \gamma_{G_2} G_2 + \gamma_{GE_1} G_1 E + \gamma_{GE_2} G_2 E$,

where G_1 and G_2 are the minor allele counts of the causal loci. We assume the same effect of both loci (i.e., $\gamma_{G1} = \gamma_{G2} = 0, 1$; $\gamma_{GE1} = \gamma_{GE2} = 0, 0.05, 0.1$). We evaluated the performance of IBS and polynomial kernels ($d=2$ and 3). The results are presented in Figure 2.9 suggest that fastKM and the EM algorithm have very similar performance.

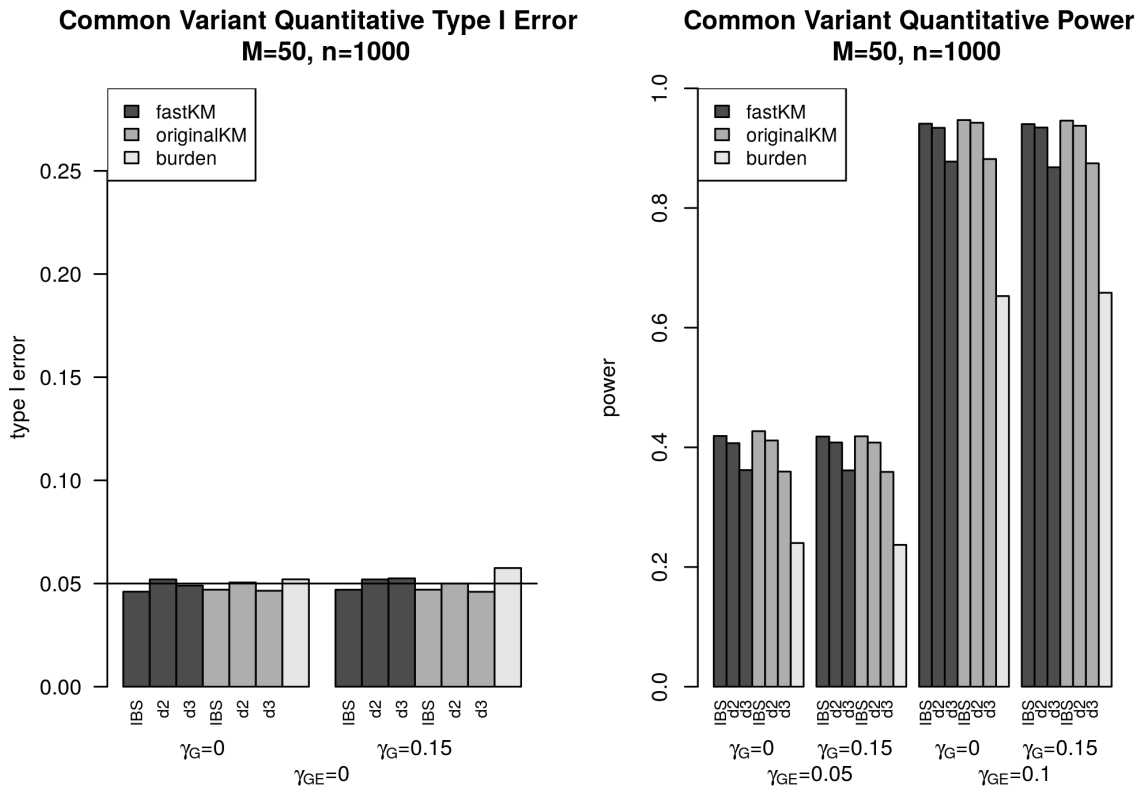


Figure 2.9 Type I error (left panel) and power (right panel) for fastKM, originalKM, and the weighted counting burden-based $G \times E$ test for common variants with continuous E covariate over varying main effect size γ_G and interaction effect size γ_{GE} . The KM tests are based on the IBS kernel and the polynomial kernels with $d = 2$ and $d = 3$, all with 95% kPCA.

Chapter 3

Cross Disorder Kernel Machine Modeling

3.1 Abstract

When trying to find genetic associations, traditional analyses follow a “bottom-up” approach, examining one gene (or variant) and one disorder at a time, using meta analysis to combine results for multiple genes/disorders. These approaches may be underpowered by ignoring comorbidities of disorders and coheritability of variants and due to high multiple testing burden of individual tests. We propose a “fastLasso” method to simultaneously analyze the effects of multiple genes along a pathway on multiple diseases. In particular, we use a fast kernel machine approach in conjunction with gene-level group lasso to pinpoint probable causal genes within a pathway for a group of related phenotypes. Our approach takes advantage of shared genetic risk between phenotypes, leading to increased power and better understanding of the biological mechanism of shared disorders. Further, it is computationally efficient and flexible, with support for both binary and continuous phenotypes, as well as for incorporation of data from different individuals for the different disorders considered. We demonstrate the utility and performance of our method over pathway-based single disorder analysis via simulation study.

3.2 Introduction

3.2.1 Motivation for Cross Disorder Analysis

Pleiotropy, or the effect of one gene on multiple traits, is an important topic in statistical genetics. Increasing evidence of comorbidity of diseases and of coheritability of variants that are associated with given disorders suggests that we can gain a better understanding of the genetic architecture of related disorders by considering them together within an analysis. This increased understanding of gene multifunctionality can be used to improve detection, diagnosis, classification, and treatment of correlated disorders (Hu et al., 2016, Insel et al., 2010, Lee et al., 2012b, Morris and Cuthbert, 2012, Sanislow et al., 2010). Data for multi-disorder studies is also more widely available with electronic health data available to help quantify co-occurring disorders (Hu et al., 2016), and multi-institution initiatives being created to better understand shared disease pathology (e.g., NIMH's Research Domain Criteria RDoC for studying related psychological disorders (Insel et al., 2010, Morris and Cuthbert, 2012, Sanislow et al., 2010)).

A motivating example of related phenotypes includes the psychological disorders anorexia nervosa (AN), schizophrenia (SCZ), and obsessive compulsive disorder (OCD), which evidence suggests have a large proportion of shared heritability (between 40-60%) (Anttila et al., 2016) and pairwise comorbidity much larger than the population prevalence of the individual disorders (Achim et al., 2009, Buckley et al., 2008, Fawzi and Fawzi, 2012, Foulon, 2003, Godart et al., 2000, Götestam et al., 1995, Hoff, 2012, Hudson et al., 2007, Kaye et al., 2004, Khalil et al., 2011, Kouidrat et al., 2014, Lysaker and Whitney, 2009, Mukhopadhaya et al., 2009, Poyurovsky et al., 2005, 2012, Rubenstein et al., 1992, Ruscio et al., 2010, Schirmbeck and Zink, 2013, Seeman, 2014, Swinbourne et al., 2012, Yum et al., 2009). By studying all three together, we can improve diagnosis and classification, making them more biologically-based and the disorders themselves easier to detect (Insel et al., 2010, Morris and Cuthbert, 2012). Li et al. (2014) discusses many other studies where leveraging pleiotropy can

increase power, e.g., for psychiatric disorders (Andreassen et al., 2013, Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013), cancer (Sakoda et al., 2013), and metabolic traits (Lee et al., 2012b, Vattikuti et al., 2012).

In addition to increasing our functional knowledge of pleiotropic effects, utilizing information from multiple disorders simultaneously, as well as from multiple genes along a pathway, can increase our signal to detect important genetic associations, as single-trait analyses ignore shared information from correlated traits (Kiezun et al., 2012, Manolio et al., 2009) and can have high multiple testing burden (Wang et al., 2015b). The gain in power from multi-trait analyses is especially important when dealing with variants with low effects and low minor allele frequencies, which may be more difficult to detect, as is the case with rare variant analysis. Not only does incorporation of correlation phenotypes effectively increase the sample size (Li et al., 2014, Maier et al., 2015), but leveraging coheritability information also enables additional borrowing of signal.

3.2.2 Current Methods

Three main classes of methods currently exist to model multiple disorders simultaneously: meta analysis and combined statistics, dimension reduction (e.g., principal component analysis, canonical correlation analysis, similarity based), and multi-response regression (Galesloot et al., 2014, Yang and Wang, 2012).

3.2.2.1 Meta Analysis and Combined Tests

Meta analyses and combined univariate tests are examples of “bottom-up” procedures, looking at single genes and disorders individually, e.g., through univariate genome-wide association studies (GWAS), and then combining results to detect pleiotropic effects and obtain tests of association at a multi-trait level.

Examples of meta analyses/combined tests include the work of Andreassen et al. (2013), Bolormaa et al. (2014), Yang et al. (2010), and Van der Sluis et al. (2013). The approaches of both Bolormaa

et al. (2014) and Yang et al. (2010) create a vector of test statistics from univariate association tests of a variant on a single trait and calculate a multivariate test statistic as a function of these test statistics. They both aim to test the null hypothesis of no genetic effect on any of the traits against the alternative that at least one trait has significant genetic effect from the variant of interest.

Yang et al. (2010) assumes the vector of test statistics follows a multivariate normal distribution and uses a modified O'Brien method (O'Brien, 1984, Wei and Johnson, 1985) to test whether or not the mean of this distribution is equal to zero, i.e. whether there is any association of the variant (or group of variants) with at least one of the traits. They create a test statistic that is the linear combination of the multivariate normal means (and corresponding estimated or known covariance of these means), estimating weights using sample splitting/cross validation and obtain significance via resampling and permutation.

Bolormaa et al. (2014) creates a quadratic test statistic from the signed t-values from GWAS (one for each variant of interest, if multiple variants are to be considered) and the correlation between each pair of traits over all variants, which approximately follows a chi-square distribution with degrees of freedom equal to the number of traits being considered.

Andreassen et al. (2013) and Van der Sluis et al. (2013) also look at summary statistic data from univariate GWAS tests, but instead of test statistics focus on combining the p-values. Andreassen et al. (2013) focuses on determining multivariate significance through the conditional false discovery rate (FDR) of two traits. In particular, they used the p-values from univariate GWAS tests to calculate the conditional cumulative distribution function (CDF) of the (corrected) p-values for each trait, conditional on the nominal p-value from the other trait, which they used to calculate the conditional FDR for each trait, creating a 2 dimensional "look up" table, looking at the maximum of the two FDRs for each variant, which they compared against a mixture model-based estimated distribution of SNPs (unconditional analysis). They note their approach has merit because you would expect a higher likelihood of a true positive variant association if it deemed significant in two associated phenotypes (Andreassen et al., 2013). It is also nonparametric with few assumptions on the traits

or genetic variants. However, it does not extend to more than two traits like the aforementioned approaches.

Van der Sluis et al. (2013) combines univariate p-values for each trait into a “trait-based” p-value in their TATES (Trait-Based Association Test that uses Extended Simes procedure) approach, calculating the minimum p-value over all traits for a given variant, weighted by the effective number of independent p-values. The effective number of p-values is calculated using an eigendecomposition of the correlation matrix between the p-values, thus taking into account correlations between the traits. Again this is testing the null hypotheses of no association between a particular variant and any of the traits. They note that follow-up is required to test more specific hypotheses of the genotype-phenotype model.

Combined tests and meta analyses have the benefit that, because they only use summary statistics (namely, the test statistic) from each GWAS test, they can analyze data with different subjects (even using published data where subject-level data is not available) and with different types of traits together (e.g., quantitative, binary, and survival), without making many assumptions on the distribution of the traits (Bolormaa et al., 2014, Van der Sluis et al., 2013, Yang et al., 2010). Further, opposing effects of variants on different traits will not cancel each other out to reduce power (Van der Sluis et al., 2013). In addition, the approaches of Yang et al. (2010), Bolormaa et al. (2014), and Van der Sluis et al. (2013) can analyze an arbitrary number of traits. However, these “bottom-up” methods may lose power by not taking into account unified information, such as the comorbidity and coheritability of traits, that can be incorporated by using the raw subject-level rather than summary data. These approaches also lose power due to high multiple testing burden from performing separate tests for each genetic variant (Wang et al., 2015b). Finally, some, like Fisher’s method of combining test statistics, can have inflated type I errors when traits are correlated (Aschard et al., 2014).

3.2.2.2 Dimension Reduction

Dimension reduction methods of multivariate analysis include principal component analysis (PCA) and canonical correlation analysis (CCA). Rather than combining summary statistics, dimension reduction approaches combine raw information, directly accounting for the correlation between traits (Aschard et al., 2014). As such, they, along with multi-trait regression methods, are examples of “top-down” approaches. These approaches take advantage of combined information from multiple genes and/or phenotypes, effectively performing meta analysis at the start, then refining to localize significant associations.

Aschard et al. (2014) and Klei et al. (2008) use PCA to perform multi-trait analysis. Aschard et al. (2014) suggests loss of power by only considering the top principal components (PCs) (e.g. the orthogonal linear combinations of data that explain the highest proportion of variability in the phenotypes), and therefore proposes a global multistep combined PC (mCPC) score. The CPC test statistic is a function of the cumulative distribution function of the aggregate of tests of association between the leading PCs and genotype, and of the aggregate of tests of association between the remaining PCs and genotype, and follows a chi-square distribution under the null hypothesis. They note their method easily generalizes to many traits, and can be used as part of a multivariate linear model to account for population or family structure.

Klei et al. (2008) considers principal components of phenotype to not be biologically accurate enough and proposes instead to look at tests for association between genotype and the principal components of heritability (PCH). They create a new phenotype that is the linear combination of the trait phenotypes that has the highest heritability (Klei et al., 2008). They use sample splitting/bagging to estimate these optimal linear weights and note that they can use this approach on residuals from PCs rather than the PCs themselves. They perform a test of association between genotype and their PCH using a t-test.

Ferreira and Purcell (2008) uses CCA to calculate a linear combination of traits that explains the highest proportion of covariability between genotype and phenotype, as is implemented in PLINK. MultiPhen (O'Reilly et al., 2012) is a similar method that is somewhat between dimension reduction and multi-trait regression models. MultiPhen performs an ordinal (proportional odds logistic) regression, modeling the probability of the genetic variants being less than or equal to a value (0,1,2) on a linear combination of the phenotypes, then using likelihood ratio tests for each variant to test whether that variant is significantly associated with at least one of the traits.

Dimension reduction techniques, again, can easily incorporate multiple (more than two) traits and often have lower multiple testing burden than meta-analysis techniques. In addition, they directly include correlations between traits, unlike meta analyses. However, they tend to be applicable mostly to normal traits only, and are not able to combine traits. In addition, they do not provide as interpretable results, as they relate linear combinations of traits with genotype, rather than the traits themselves.

3.2.2.3 Multi-trait Regression

Multi-trait regression approaches, like dimension reduction approaches, are “top-down” approaches, leveraging information about correlations between traits, comorbidity, and coheritability directly. Most existing multi-trait regression models fit into the category of multivariate linear mixed effects models.

3.2.2.3.1 Multivariate Linear Mixed Effects Models

Multivariate linear mixed effects models (multivariate LMMs, or mLMMs) have been commonly used for genetic analyses involving multiple traits and multiple variants. mLMMs use a random effects framework to explicitly model genetic sharing through the variance/covariance of a genetic random effect term. Many mLMM methods focus on different aspects of multi-trait analysis, such as estimating heritability and pleiotropy through the genetic correlation between a set of traits (e.g.,

Korte et al. (2012), Lee et al. (2012b), Loh et al. (2015), Vattikuti et al. (2012)) and multivariate genetic risk prediction (e.g., Maier et al. (2015)).

Vattikuti et al. (2012) and Lee et al. (2012b) proposed a bivariate LMM to estimate the genetic correlation between a set of traits as a surrogate predictor of genome-wide pleiotropy. Vattikuti et al. (2012) used an EM algorithm for restricted maximum likelihood (REML) estimation for continuous traits, while Lee et al. (2012b) proposed using an efficient average information restricted maximum likelihood (AIREML) approach, approximating the Hessian with the average information (Gilmour et al., 1995, Loh et al., 2015) to estimate on a continuous scale, and showed how a liability threshold model could be used to obtain genetic correlation when working with case/control data. Li et al. (2014), however, notes that the AIREML algorithm occasionally fails to converge and is not ideal for binary traits as it uses normality assumptions.

Loh et al. (2015) proposed “BOLT-REML” to increase efficiency and scalability (up to 50,000 subjects) of AIREML to estimate variance components and thus heritability and genetic correlations, using Monte Carlo sampling to approximate the gradient for the mixed models. They focus on common variants, however, and use liability scale to convert from case control data.

Korte et al. (2012) proposed a multitrait mixed model (MTMM) to estimate genome-wide heritability and genetic correlation of a pair of traits as functions of estimated variance components of the model, taking into account relatedness/kinship of individuals and environmental effects. They set up their model with two random effects terms to separately model within-trait and between-trait effects (as an interaction between the trait an observation is for and the genotype), allowing them to perform three marker-level tests for GWAS data, testing for: (1) common and differing effect loci between traits, (2) common genetic effects between traits, and (3) differing effects between traits. This model is more flexible but less efficient than that proposed by Maier et al. (2015) for estimating pleiotropy, and only discusses testing for one marker at a time for GWAS testing, which can have a high multiple testing burden.

Maier et al. (2015) proposed a mLMM for genetic risk prediction. Making use of the AIREML approach, they calculate multi-trait genomic best linear unbiased predictors (MTGBLUPs) for individual risk prediction of sampled individuals and use these to calculate snp-level BLUPs which can be projected to predict risk for individuals not in the sample. Their approach allows for individuals to come from different samples, but has lower accuracy for polygenic traits when not also incorporating additional gene annotation information.

While these approaches have been successful, their focus is not on association testing of genotype with the traits, but on understanding and quantifying how the traits are related, or on predicting phenotype for new individuals. Two approaches that do aim to perform association testing are those of Zhou and Stephens (2014) and Casale et al. (2015).

Zhou and Stephens (2014) proposed a mLMM for GWAS, accounting for external covariates such as population substructure and kinship, using the EM algorithm with Newton-Raphson to combine stability and fast convergence. Their method does not allow for missingness in phenotype data, however, and requires all phenotypes be measured on the same subject. Further, they require a separate likelihood ratio test (LRT) for each variant of interest, leading to a higher multiple testing burden.

Casale et al. (2015) proposed the multi-trait set test “mSet” model, using two variance components to model the relatedness of individuals and population substructure (“relatedness” random effect) along with the combined genetic effect over a variant set (“set” random effect). Their model allows for testing of no genetic effect (no “set” component) for genome-wide data on up to 500,000 subjects using efficient linear algebra to make it take a similar amount of time as fitting variance component models with a single variance component (Casale et al., 2015). However, they do not pinpoint which variants within the set are more likely to be associated with at least one of the phenotypes.

As mixed models, mLMMs are flexible and efficient, and are more robust and higher power than fixed effects models for polygenic traits, as they can aggregate information over sets of variants with

weak individual effect (Korte et al., 2012, Wu et al., 2010, 2011). Like dimension reduction approaches, they take advantage of shared information, coheritability and comorbidity, but yield much more interpretable results and may allow for phenotype data to come from different individuals/studies. However, they assume normality of the phenotype data, or simply perform a linearization of binary case/control data, which may work well for heritability estimates (Lee et al., 2012b) but is not valid for association testing because of poor modeling of confounding effects.

3.2.2.3.2 Other Multi-trait Regression Models

Others have looked at non-LMM multi-trait regression models. Wang et al. (2015b) proposed a multivariate functional linear regression, which, rather than looking at the genetic loci as discrete variables, includes their effects as a smooth function of genetic position. Approximate F-tests, adjusting for covariates, then can be used to test for no genetic effect on any of the traits of interest (Wang et al., 2015b). This has the benefit of taking into account covariates and genomic position and incorporating information on linkage disequilibrium in a natural manner, but does not differentiate where the genetic signal, if any, is coming from.

Li et al. (2014) suggested the related bivariate ridge regression to predict multiple phenotypes, using the correlation between the diseases to increase prediction accuracy (the area under the receiver operator curve) over single-trait models. They suggest that by effectively increasing sample size, they can overcome one of the main bottlenecks in genetic risk prediction (Makowsky et al., 2011, Wray et al., 2013). Their model includes three regularization parameters for two disorders - one for each of the genetic effect of each disorder, and one for the correlation between them, which they tune using a grid search and cross-validation to choose the optimal values. Their use of ridge regression is due to the belief (De Los Campos et al., 2010) that prediction models are more powerful with the inclusion of more traits with weaker effects (even when including noise and opposite effect terms), e.g. a whole-genome model, than a sparse model with only a few strong effects, as would be selected with a lasso model (Li et al., 2014). This is good for risk prediction, but less ideal for

pinpointing variants most likely to be causal within a variant set.

Other approaches are similarity-based. Wei and Lu (2015) proposes a generalized similarity U test for sequencing data that can be applied to multiple traits. Maity et al. (2012) and Broadaway et al. (2016) propose kernel-based similarity methods. Maity et al. (2012) suggested a multivariate kernel machine regression model, using a kernel term to express complex epistatic effects of different variants. They use a score test statistic to test for no genetic effect of a set of variants. This is similar to a mLMM, which can be seen through equivalence of norm functions from the penalized log likelihood for a fixed covariance matrix, but can be generalized to other exponential family distributions and allows more flexible modeling of relatedness between traits. Broadaway et al. (2016) proposed the Gene Association with Multiple Traits (GAMuT), which uses a “machine learning kernel distance-covariance” approach to test for association between multiple traits and a set of genetic variants (Broadaway et al., 2016). Their approach is nonparametric, relating the similarity between traits to the similarity between genotypes on a pairwise level. It does not assume normality of phenotype, and is easy to include any arbitrary number of genetic variants. However, neither of these methods focus on variant selection of genetic variants that are associated with at least one trait.

3.2.3 Introduction to fastLasso

Following the work of Maity et al. (2012), we propose a kernel machine approach to look for associations between genetic variants and a group of traits. Rather than testing for overall association between a variant set and the traits, however, we wish to perform gene refining to identify which genes within a pathway are more likely to be the causal genes. We propose the “fastLasso” method for performing cross-disorder variable selection on genes within a pathway. Our method performs group lasso (Yuan and Lin, 2006) on an efficient decomposition of a cross-disorder kernel matrix in order to identify which single nucleotide variants (SNVs) inside of genes within a pathway are associated with at least one of multiple traits. We choose the lasso rather than ridge regression, as in Li et al. (2014), for regularization because we wish to pinpoint causal genes and generate hypotheses

for further biological follow up, requiring sparser and more defined models than are required for genetic risk prediction.

The fastLasso approach

1. takes advantage of the ability of kernel methods to capture complex epistatic relationships between genetic variants,
2. is able to simultaneously perform effect estimation and variable selection on the SNVs along a pathway for continuous or binary traits, and
3. can combine information from different studies, not requiring overlapping subjects for the different traits considered.

By combining information from multiple disorders we have increased signal to detect rare variants. We are able to do this in an efficient, scalable manner by borrowing the low-rank fastKM decomposition of Marceau et al. (2015).

We perform a simulation study based off of the CoLaus genome wide association study (GWAS) and exon-sequencing of single nucleotide variants (SNVs) to examine the performance of our method compared with traditional approaches, only studying one disorder at a time, then combining results.

3.3 Methods

We consider a study with D disorders of interest with some expected genetic or diagnostic commonality. We let Y_d denote a $n_d \times 1$ vector of responses for all patients whose disease status (continuous or binary phenotype) is known for disorder $d = 1, \dots, D$. Further, we define X_d to be a $n_d \times p_d$ matrix of non-genetic covariates (e.g., age, sex, population substructure) for disorder d , and $G_{d,\ell}$ to be a $n_d \times m_\ell$ genotype design matrix for gene $\ell = 1, \dots, L$ within a pathway of interest, where m_ℓ is the total number of markers (single nucleotide variants, snvs) genotyped for gene ℓ .

For simplicity, we consider the case where we are interested in $D = 3$ coheritable disorders, and let $Y_{n \times 1} = (Y_1, Y_2, Y_3)^T$ and $X_{n \times p} = \begin{bmatrix} X_1 & 0 & 0 \\ 0 & X_2 & 0 \\ 0 & 0 & X_3 \end{bmatrix}$ be the combined phenotype and covariate design matrices, respectively. Here $n = \sum_{d=1}^3 n_d$, $p = \sum_{d=1}^3 p_d$

Our goal is to determine which genes within a pathway of interest are significantly associated with at least one of the disorders, simultaneously performing variable selection and estimating effect size from each variant/gene. To do this, we wish to perform group lasso based on the cross-disorder kernel machine regression model

$$g(\mu_Y) = g \begin{pmatrix} \mu_{Y_1} \\ \mu_{Y_2} \\ \mu_{Y_3} \end{pmatrix} = \beta_0 + X\beta + \sum_{\ell=1}^L K_\ell \alpha_\ell \quad (3.1)$$

where $\beta_{p \times 1} = (\beta_1, \beta_2, \beta_3)^T$, $\mu_Y = E(Y|X, G)$ is the phenotypic mean given all genetic and non-genetic covariates, and $g(\mu_Y)$ is the canonical link function. Further, K_ℓ is a $n \times n$ kernel similarity matrix for gene ℓ and α_ℓ is a $n \times 1$ random effect of gene ℓ , $\alpha_\ell \sim N(0, \tau_\ell K_\ell^{-1})$ for invertible K_ℓ , or more generally $h_\ell = K_\ell \alpha_\ell \sim N(0, \tau_\ell K_\ell)$.

In order to make this computationally feasible, we follow three main steps: (1) form a kernel matrix to evaluate the genetic similarity between individuals within and between disorder studies, (2) perform dimension reduction on the similarity kernel and form a low-rank fixed effect term to summarize genetic effects over all studies/disorders, in the fastKM manner, and (3) fit a fastLasso group lasso model using the low-rank fastKM term.

3.3.1 Kernel Evaluation

We form a kernel matrix to evaluate genetic similarity between all individuals using column-standardized $n \times m_\ell$ genotype design matrices, $\tilde{G}_\ell = (\tilde{G}_{1,\ell}, \tilde{G}_{2,\ell}, \tilde{G}_{3,\ell})^T$, using the identity by state (IBS) kernel or linear kernel $\tilde{G}_\ell \tilde{G}_\ell^T$. We note that for rare variants these are nearly equivalent. We

can alternatively express K_ℓ in terms of its subcomponents: $K_\ell = \begin{bmatrix} K_{11,\ell} & K_{12,\ell} & K_{13,\ell} \\ K_{12,\ell}^T & K_{22,\ell} & K_{23,\ell} \\ K_{13,\ell}^T & K_{23,\ell}^T & K_{33,\ell} \end{bmatrix}$. Here $K_{d,d',\ell}$ is $n_d \times n_{d'}$ matrix representing the genetic similarity between the individuals from disorder/study d and disorder/study d' . This emphasizes the explicit incorporation of covariance of variants between individuals from different studies (focusing on different disorders) since $K_{d,d',\ell}$ is not required to be zero.

3.3.2 Dimension Reduction

We wish to make group lasso computationally efficient for the large sample size and number of variants. In order to do so, we perform kernel principal component analysis (kPCA) on our L kernel matrices. We perform an eigendecomposition of each kernel matrix as $K_\ell = Q_\ell \Lambda_\ell Q_\ell^T$, where $Q_\ell, n \times m_\ell$ is a matrix of eigenvectors, and $\Lambda_\ell, m_\ell \times m_\ell$ is a diagonal matrix of eigenvalues. We then take the top k eigenvalues which collectively explain $e\%$ (e.g., 95%) of the variability in the kernel matrix to form a rank- k decomposition, where $k < m_\ell < n$. Following the fastKM methodology (Marceau et al., 2015), we can form a low rank approximation for the gene effect as: $(K_\ell \alpha_\ell)_{n \times 1} \approx Z_\ell Z_\ell^T \alpha_\ell \equiv (Z_\ell \gamma_\ell)_{k \times 1}$. We can thus form a new cross-disorder fastKM model of the form

$$g(\mu_Y) = g \begin{pmatrix} \mu_{Y_1} \\ \mu_{Y_2} \\ \mu_{Y_3} \end{pmatrix} = \beta_0 + X\beta + \sum_{\ell=1}^L Z_\ell \gamma_\ell \quad (3.2)$$

where γ_ℓ is a $k_\ell \times 1$ vector, and $k_\ell \ll n$, improving the computational efficiency, scalability, and stability of a group lasso model fit.

3.3.3 fastLasso

We can fit a group lasso model based on the cross-disorder fastKM model using existing software, e.g. the `grpreg` package in R (Breheny and Huang, 2015), using the fastKM design matrix $Z_{n \times (1+p+k)} =$

$(1, X, Z_1, Z_2, \dots, Z_L)^T$, $k = \sum_{\ell=1}^L k_\ell$ and cross-disorder phenotype vector Y as input.

As a group lasso model, fastLasso solution is the γ that minimizes (Breheny and Huang, 2009)

$$Q(\gamma) = \frac{1}{2n} \|Y - Z\gamma\|^2 + \lambda \sum_{\ell=1}^L \sqrt{k_\ell} \|\gamma_\ell\| \quad (3.3)$$

imposing sparsity on a pathway level, but borrowing signal from all variants within chosen genes (Breheny and Huang, 2009).

We use Bayesian Information Criterion (BIC), $BIC(\lambda) = 2L_\lambda + \log(n)df_\lambda$ (Breheny and Huang, 2009), to tune the regularization parameter λ , as BIC is known to be consistent and computationally efficient (Yang, 2005). Here df_λ is the effective number of model parameters, which can be estimated as a function of the fitted coefficients $\hat{\gamma}$ and unpenalized fitted coefficients (Breheny and Huang, 2009).

We obtain as output a list of the genes which are likely to be associated with at least one of the traits, as well as relative effect sizes for the variants within those genes.

3.3.4 Computational Efficiency

The computational burden of the fastLasso approach is dominated by three operations: (1) calculating the genetic similarity kernel matrices, (2) subsequently performing eigenvalue decomposition on said kernel matrices, and (3) tuning and fitting a group lasso model. The first two can be straightforwardly parallelized, as the separate gene kernel matrices are independent from one another. We can further improve the efficiency of (2) by noting that the rank of these similarity kernel matrices are always $\leq \min(m_\ell, n)$, so we can either compute just the top m_ℓ eigenvalues for each kernel matrix (using efficient numerical linear algebra, as in Qiu and Mei (2016)), or equivalently perform eigendecomposition on the $m_\ell \times m_\ell$ matrix $\tilde{G}^T \tilde{G}$ (we assume $m_\ell \ll n$ since each kernel matrix is gene-level). (3) is relatively efficient using a fast coordinate descent algorithm in combination with the efficient BIC criterion (Breheny and Huang, 2009), and we improve upon this further with use of

the fastKM design matrix.

3.4 Simulation Study

3.4.1 Data Generation

We perform a simulation study to examine the type I error and power of our method for $D = 3$ traits, using the CoLaus clinical trial study data of Firmann et al. (2008) as a basis to generate simulated genotypes, using real data to take advantage of the natural correlations between SNVs. The CoLaus study was a population-based trial examining cardiovascular, psychological, and related metabolic risk factors in Caucasians in Lausanne, Switzerland (Firmann et al., 2008, Preisig et al., 2009). From the initial $n = 1769$ individuals for which we have full genotype information (GWAS with imputations for missing genotype information), we first form a gene pool from which to base our simulations.

To do so, we extract information from genes within chromosomes 1-9 in the CoLaus study. We are interested in how leveraging information from multiple disorders can help in the identification of rare variant associations, so we only include rare variants, which we here define as having a minor allele frequency (MAF) of less than or equal to 1%, in our gene pool. Further, we consider only those genes with at least 5 rare variants, leaving us with 5421 variants from 102 genes in our analysis, with between 5 and 230 SNVs per gene considered. The median number of rare variants/gene was 42. We perform sampling from this variant pool to form sampled individuals and genotypes using random sampling for continuous traits, as described below. An approach to perform case control sampling for binary traits can be found in Appendix A.

3.4.1.1 Random Sampling of Genotype Matrix

For continuous trait simulations, we perform random sampling of the variant pool. We first create a 6000×5421 sample genotype matrix G^* , creating each individual genotype by individually sampling each gene with replacement from the genotypes from the original subjects, then repeating this

process 6000 times to get 6000 sampled genotypes. The first 2000 individuals in G^* were assigned to disorder 1, the next 2000 to disorder 2, and the last 2000 to disorder 3.

We randomly sample 20% of genes to be causal for one or more of the disorders. Of these, we consider $s = 40\%$ or $s = 60\%$ of the causal variants to be common between all three disorders and therefore 60% or 40% to be unique to only one of the disorders, spread evenly amongst all three disorders. For simplicity, we consider all variants within causal genes to be causal.

Continuous phenotypes for subjects $j = 1, \dots, 2000$ within disorder $d = 1, 2, 3$ were randomly generated from a normal distribution $y_{j,d} \sim N(\mu_{j,d}, 1)$ with mean $\mu_{j,d} = \beta_0 + X\beta_X + G_{j,d}^*\beta_d$. Here $G_{j,d}^*$ denotes the $d \times j^{th}$ row of the random genotype matrix G^* , i.e. the genotype for the j^{th} individual within disorder d .

For our simulations, we set $\beta_0 = 1$ to approximate a 50% disease rate, and set

$$\beta_d = \begin{cases} \gamma_G & \text{if gene } \ell \text{ is causal for disorder } d \\ 0 & \text{if gene } \ell \text{ is noncausal for disorder } d \end{cases}$$

For simplicity, we do not consider any non-genetic covariates, so $X\beta_X = 0$. Further, we consider the same effect size γ_G for all causal genes within all disorders, rather than basing on minor allele frequency. We consider $\gamma_G = 1, 2$ for continuous traits, leading to models where approximately 70% and 90% of the variability in the model is explained by the causal variants.

3.4.2 fastLasso Simulation

We use the `grpreg` package in R (Breheny and Huang, 2015) to perform group lasso on the fastKM genetic design matrix, defining a group to be a gene. We find the optimal model over a grid of λ tuning parameters using BIC, but further perform hard thresholding of the model to obtain better separation of normed coefficients between causal and noncausal variants. This is due to the properties of the null model fit, which still includes many nonzero coefficient terms, likely due

to the fact that the genotype design matrix is very rare. We use the null model to determine an appropriate threshold, examining the distribution of the optimal fastLasso coefficients. A histogram of the non-zero coefficients from this fit can be found in Figure 3.1 below. We see that the largest absolute value coefficient is just over 0.03, indicating $T = 0.02$ and $T = 0.03$ are good choices for threshold values. To perform the hard thresholding, variants whose β coefficients were less than threshold T in the BIC-chosen optimal model were set to zero.

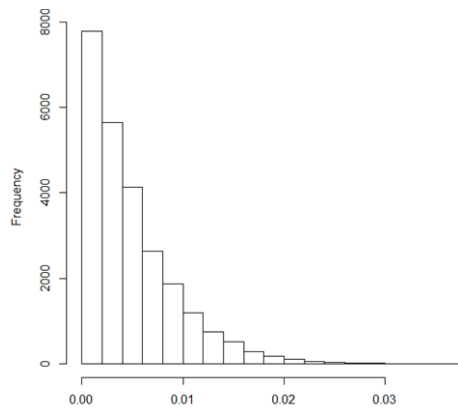


Figure 3.1 Histogram of the non-zero coefficients from the null model fastLasso fit

We compare the cross-disorder model fit to that of fitting a group lasso separately within each disorder. We summarize results from the single disorder analyses by determining the union and intersection of the genes found to have non-zero coefficients over the three single disorder model fits. These provide positive and negative controls, respectively.

3.5 Results

From Table 3.1 below we see that in all simulated scenarios the cross disorder (CD) model is able to find on average around double the causal genes that can be found using the union of the single disorder (SD-U) approach. This is even more evident among the “unshared” causal variants, i.e.

those that are causal for a single disorder. While both methods do better in finding the variants that are shared (i.e. causal for all three disorders), the magnitude of improvement of the cross disorder model over the union of single disorder models is much larger for unshared than shared variants. We see that the cross disorder model also outperforms the union of single disorder models in terms of false positive rate, picking on average fewer non-causal genes in the optimal model in all scenarios considered except for when $\gamma_G = 2$ and $T = 0.02$. Though we see this trend, we note that the median false positive rate is actually on average much lower for the single disorder models, indicating they may just choose fewer genes as significant overall but have less stability in model fit than the cross disorder model. This can be seen in Figure 3.2 below. We note that with approximately 5000 genetic variants, a model with sample size $n = 6000$ is much more likely to be stable than one with $n = 2000$. The intersect of the single disorder models performs poorly (has close to zero true positive rate) in all scenarios, but also does not pick out any non-causal genes (i.e., a zero false positive rate) and is overall of little interest statistically.

Table 3.1 Average true positive and false positive rates (and corresponding standard deviation) for cross disorder (CD), the union of single disorder (SD-U), and the intersection of single disorder (SD-I) continuous trait kernel machine model analyses over 100 simulations. Largest values within each category are in bold font.

γ_G	% causal shared	% variance explained	threshold	True Positive Rate									False Positive Rate		
				CD	shared SD-U	SD-I	unshared			all			CD	SD-U	SD-I
1	40	66.8	0.02	1 (0)	0.52 (0.25)	0.01 (0.03)	0.68 (0.08)	0.2 (0.33)	0 (0)	0.81 (0.05)	0.33 (0.3)	0.004 (0.01)	0.1 (0.03)	0.18 (0.38)	0 (0)
			0.03	1 (0)	0.46 (0.27)	0.001 (0.01)	0.46 (0.09)	0.18 (0.33)	0 (0)	0.67 (0.06)	0.29 (0.3)	0 (0.01)	0.01 (0.01)	0.17 (0.35)	0 (0)
			0.02	0.93 (0.03)	0.42 (0.31)	0.004 (0.02)	0.81 (0.1)	0.23 (0.4)	0 (0)	0.88 (0.05)	0.33 (0.35)	0.002 (0.01)	0.12 (0.04)	0.2 (0.39)	0 (0)
	60	69.4	0.02	0.918 (0.01)	0.38 (0.32)	0.002 (0.01)	0.62 (0.12)	0.21 (0.39)	0 (0)	0.79 (0.05)	0.3 (0.35)	0.001 (0.01)	0.02 (0.01)	0.18 (0.36)	0 (0)
			0.03	1 (0)	0.81 (0.16)	0.11 (0.05)	0.87 (0.05)	0.38 (0.32)	0 (0)	0.92 (0.03)	0.55 (0.24)	0.04 (0.02)	0.28 (0.05)	0.25 (0.43)	0 (0)
			0.03	1 (0)	0.78 (0.17)	0.11 (0.05)	0.84 (0.05)	0.36 (0.31)	0 (0)	0.9 (0.03)	0.53 (0.24)	0.04 (0.02)	0.09 (0.03)	0.25 (0.42)	0 (0)
2	60	90.1	0.02	0.94 (0.04)	0.5 (0)	0.08 (0)	0.98 (0.04)	0.22 (0)	0 (0)	0.96 (0.03)	0.38 (0)	0.048 (0)	0.3 (0.04)	0 (0)	0 (0)
			0.02	0.92 (0.01)	0.57 (0.21)	0.07 (0.04)	0.95 (0.06)	0.33 (0.32)	0 (0)	0.93 (0.03)	0.47 (0.26)	0.04 (0.02)	0.11 (0.03)	0.18 (0.37)	0 (0)
			0.03	0.92 (0.01)	0.57 (0.21)	0.07 (0.04)	0.95 (0.06)	0.33 (0.32)	0 (0)	0.93 (0.03)	0.47 (0.26)	0.04 (0.02)	0.11 (0.03)	0.18 (0.37)	0 (0)

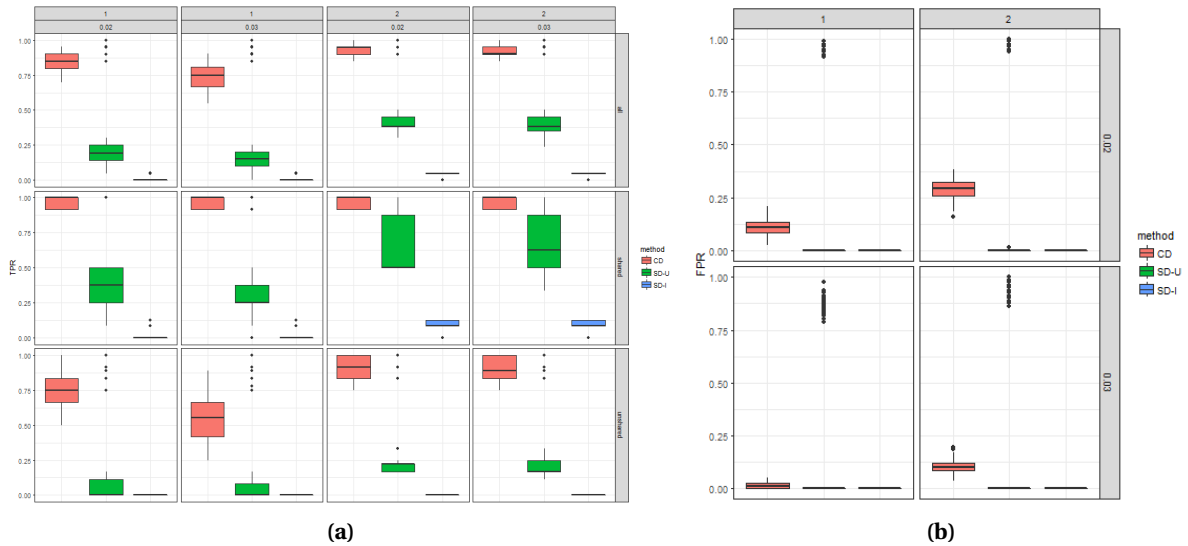


Figure 3.2 True positive and false positive rates for cross disorder (CD), the union of single disorder (SD-U), and the intersection of single disorder (SD-I) continuous trait kernel machine model analyses over 100 simulations

3.6 Discussion

In this paper, we consider the benefits of leveraging information from multiple correlated traits when conducting genetic association studies. Namely, we note that looking for association between a set of variants and a set of phenotypes/disorders allows us to gain a better understanding of the underlying pleiotropy and true genetic architecture for these disorders, leading to the potential for improved diagnosis, classification, and treatment. Further, by increasing our effective sample size and by allowing incorporation of comorbidity and coheritability directly into our analyses, we show that we increase our power to detect true causal variants (those that are associated with at least one trait) while having nearly identical, or occasionally lower, false positive rates. This additional power is especially helpful when trying to detect rare variant associations.

While there are many existing approaches to incorporate multiple traits into an analysis, not many are able to pinpoint the genes/variants most likely to be associated with at least one trait. Most focus on either single-variant tests, which lead to high multiple testing burden, or overall genome-wide tests of association. We propose the fastLasso method to efficiently perform gene-selection while estimating relative effects of association between said genes and at least one of the disorders that allows data to come from different studies, not requiring overlapping individuals, in a way that is easy and valid to apply to both continuous and binary traits using existing group lasso software. We note that as the number of genetic variants increases, it becomes infeasible to perform this type of analysis without the fastKM decomposition.

In our simulations, we suggest using a hard threshold on fastLasso to decrease false positive rates stemming from the sparse genotype design matrix. In our simulation we choose this threshold using the null model fit, looking at the distribution of nonzero model coefficients. We note that this could also be used for real data applications by fitting the fastLasso model with permuted phenotype values, creating an effective null model for comparison.

While we focus on continuous trait SNV-level analysis for genetic main effects, we note it is straightforward to extend to binary traits (also handled in the fastKM and grpreg R packages). It is also straightforward to add terms to our model to incorporate other genetic information, e.g. common single nucleotide polymorphisms (SNPs) and copy number variants (CNVs), leading to a full pathway model to further understand the true biological network of the disorders studied. Further, an additional kernel term could allow for incorporation of population substructure or gene-environment (GxE) interaction, as is demonstrated in the fastKM methodology.

Chapter 4

Rare Variant Prioritization Using Structure-Supervised Kernel Association Tests

4.1 Abstract

Rare variants are of increasing interest to genetic association studies, largely due to their part in explaining common complex diseases. However, rare variants are difficult to detect individually since they are uncommon and often have weak effects. Traditional aggregation analyses improve signal compared to single variant tests but suffer from loss of power from incorporating noise variants and are not able to pinpoint causal variants within a variant set. One way to overcome these issues is to incorporate additional information, e.g. on structure or predicted function. We propose a rare variant association test which utilizes protein tertiary structure to increase signal and identify likely causal variants. Following the biological hypothesis that important variants are likely to cluster together in 3D protein space, we perform structure-guided collapsing, leading to

variant-level tests which borrow information from neighboring variants on a protein. We use a kernel machine framework along with resampling to obtain variant-level significance, and show that our method performs at least as well as single variant level tests in the absence of clustered causal variants.

4.2 Introduction

Rare genetic variants, which occur infrequently, e.g. in less than 1-3% of a population, play an important role in understanding complex diseases. As Next Generation Sequencing (NGS) studies become cheaper and data becomes more widely available, we are better able to study these less common variants, which may be a part of uncovering the “missing heritability” from examining only common variants in complex disease studies (Manolio et al., 2009). Studying rare variants can be a challenging task, as individual rare variants may be difficult to detect due to low minor allele frequency (MAF) and weak individual effects (Li and Leal, 2008). Detecting associations with rare variants requires a much larger odds ratio and sample size than their common counterparts (Manolio et al., 2009, Morris and Zeggini, 2010) and often requires incorporation of additional information for sufficient signal.

One way to facilitate rare variant analysis is by aggregating information across variants. Much rare variant research to date has focused on “global” or variant set level tests, which collapse genetic information across a set of variants (e.g., a gene or pathway) in order to gain sufficient power to detect association. This can be done in a burden-based fashion, modeling phenotype as a function of a (weighted) sum of genetic markers, as in Madsen and Browning (2009), Morris and Zeggini (2010), Price et al. (2010), and Li and Leal (2008), or using kernels to model phenotype as a function of the genetic similarity between all pairs of individuals, as in Kwee et al. (2008), Lin et al. (2011), Liu et al. (2008, 2007), and Wu et al. (2010, 2011, 2013). These global aggregation tests improve signal over traditional single variant tests but lose power from collapsing over noise variants, and possibly

over variants with opposite effect and are not able to pinpoint the causal rare variants.

Causal rare variant prioritization, or “localization,” either on a genome-wide scale or from a previously identified significant variant set, can be done using local tests. Local tests may help us to guide follow-up studies to better understand the biological etiology of diseases and lead towards more targeted therapeutics for genetic influenced disorders. Existing localization methods often use penalization or most promising subset methodology.

Penalization methods help us to model associations with large quantities of rare variants without direct collapsing using group and/or individual-level penalties, enforcing sparsity by shrinking the estimates of noncausal variants or groups of variants. They can readily allow modeling over multiple genetic regions of interest simultaneously through use of group-level penalties (Larson and Schaid, 2014). Xu et al. (2012) found that sparse methods, such as the least absolute shrinkage and selection operator (lasso), ridge regression, and sparse partial least squares methodology outperformed traditional global testing, noting the increased power of the lasso in particular when a higher proportion of rare variants are causal. Some notable uses of lasso for rare variant association modeling includes that of Zhou et al. (2011, 2010) and Larson and Schaid (2014). Zhou et al. (2010) looked at using a mixture of penalties – both Euclidean group (e.g., gene or pathway-level) and individual lasso penalties, to allow rare variants with weak individual effect to enter the model, while avoiding forced collapsing. They further discussed use of weighted penalties to incorporate additional external biological information (Zhou et al., 2011). Larson and Schaid (2014) extended group penalty approaches, looking at both gene-based and exon-based lasso (GB-L and EB-SGL), examining the performance of using group penalties to localize most promising genes and exons, respectively, from whole exome data. While their methods performed well overall, they found variant-level sparse group lasso to have low power due to insufficient sequencing information, where many exons studied contained at most one rare variant in their simulation. Further, penalization approaches tend to focus on improving the power of global analyses, rather than variant level, and do not perform any explicit testing of association.

Most promising subset methodology focuses on clustering variants to find the top group of variants most likely to be causal, using the ideology that causal variants are more likely to be nearby, e.g. on a DNA sequence or functional domain (Larson and Schaid, 2014). Fier et al. (2012) highlights spatial clustering motivation, namely how variants that are close on the DNA sequence may be more likely to appear in the same protein functional domain or same gene regulatory element and thus are more likely to be deleterious or protective as a group based on genomic clustering. Ionita-Laza et al. (2012) gives many examples of diseases, both Mendelian and complex, which have shown clustering of causal variants, e.g. within domains or “mutational hot spots” (Ionita-Laza et al., 2012, Yue et al., 2010). Clustering localization approaches do not require outside functional information, just location of variants, e.g. along a sequence (Fier et al., 2012).

Fier et al. (2012) uses spatial clustering ideology to obtain a set-level test for rare variants, distance-based measure (DBM). In particular, they follow the ideology of and examine the distribution of distances between rare variants in both cases and controls, weighting on minor allele frequency in cases and controls, and on minimum distance to nearest neighbor in each group. They use the maximum of the nonparametric Ansari-Bradley test statistic for cases and controls (Ansari and Bradley, 1960), along with permutation of case/control status in individuals, to test for significance of a set of rare variants. They show that using spatial location increases power over many burden and kernel tests, but works best for coding regions with high coverage (Fier et al., 2012) and does not tell us which variants are more likely to be the true causal variants.

Kulldorff (1997) and Ionita-Laza et al. (2012) further the idea of using rare variants’ position on a DNA sequence, but rather than trying to improve power of global tests, they try to identify the window within a sequence that contains the highest degree of clustering of “disease-risk variants.” They expand upon the exhaustive scan statistic methodology, calculating a likelihood ratio statistic for all possible windows of variants based on the proportion of affected individuals at each site, within and outside of a window, and determine the window with highest probability of containing a cluster of causal variants by selecting the maximum LR statistic over all maximum statistics for each

window size. Significance of this window is calculated using Monte Carlo simulations, permuting case/control status (Ionita-Laza et al., 2012). This method is useful for localizing a most promising subset of rare variants, but has similar difficulties as burden-based collapsing approaches, i.e. it assumes causal variants affect the disease in the same direction, and its power is dependent on window and gene size, decreasing for large windows, especially proportional to the studied genomic region, and for smaller genes with less variability (Ionita-Laza et al., 2012).

Kernel-based approaches are also used for spatial clustering to find most promising subsets. Tango (2010) uses a kernel approach to examine the geographic clustering of case and control subjects using with kernel $A_{ij}(\tau) = e^{-4(d_{ij}/\tau)^2}$. Here they allow τ , which determines the “maximum distance between clusters” (Schaid et al., 2013), to vary, calculating a minimum p value test statistic from the quadratic statistic $Q = (O - E)^T A (O - E)$. Schaid et al. (2013) extends this approach to spatial clustering of genetic variants for case/control binary phenotype, noting that the kernel matrix acts as a smoother of rare variants, helping to pinpoint most likely causal variants when global tests of association are found to be significant (Schaid et al., 2013). Schaid et al. (2013) uses the tri-weight kernel, which gives similar shape from distances as the Gaussian kernel used by Tango (2010), scaling based on maximum distance of interest, and calculating a minimum p test statistic based on proportions of the maximum distance of interest. Significance is calculated using moments of scaled χ^2 distribution or Davies’ method (Davies, 1980, Schaid et al., 2013). Schaid et al. (2013) compared the methods of Fier et al. (2012) and Ionita-Laza et al. (2012) with their method, finding using spatial information of genes increases power over SKAT when any genomic correlation exists, and that the scan statistic of Ionita-Laza et al. (2012) outperformed the other methods. Lin (2014) also uses a tri-weight kernel to incorporate spatial information in their method CLUSTER. CLUSTER creates an aggregate test that only combines information from variants most likely to be causal by thresholding p-values from variant-level Fisher’s exact tests. This approach works particularly well when all variants are protective, but is outperformed by the scan statistic when variants are of opposite effects or all deleterious (Lin, 2014).

We note that existing methods have some things in common. Namely, they use variants' proximity on the DNA sequence to penalize or cluster variants, pinpointing which areas of the sequence are most likely to be causal. Also, they do not explicitly determine variant-level significance or calculate variant-level p-values. We propose a method to explicitly prioritize rare variants using biostructural information to facilitate borrowing of information between neighboring variants. Here, however, we define neighboring as those variants which are close-by on the protein tertiary space. Protein tertiary structure refers to the three-dimensional folding of amino acids. We incorporate protein tertiary structure into a traditional kernel machine regression framework for genetic association studies to enable powerful variant-level tests via supervised borrowing of signal from neighbors. Incorporating this information allows variants to borrow information from neighbors in the 3D protein space, but potentially far apart in DNA sequence, stemming from the biological hypothesis that variants that are near in the structural space have a similar effect on phenotype. This hypothesis is supported by work done by Song et al. (2012), who studied variants within phospholipase A2 group VII (*PLA2G7*) in association with lipoprotein-associated phospholipase A₂ (Lp-PLA₂), and found that important variants tended to cluster on the surface of the protein structure, while null variants tended to be near the core.

We define local kernels to enable variant-specific borrowing as a function of distance between pairs of variants in the 3-dimensional protein space and conduct local score tests for each variant over a range of maximum distances allowed for borrowing of information. We calculate a minimum p statistic for each variant, evaluating significance using a resampling approach. We evaluate the performance of our method using a simulation study based on the *PLA2G7* sequencing data, comparing its power and type I error to the single variant-level score test and to the scan statistic. Further, we apply our method to the Action to Control Cardiovascular Risk in Diabetes (ACCORD) clinical trial data, finding promising variants in *PCSK9* for targeting for reduction in low-density lipoprotein (LDL).

4.3 Methods

4.3.1 Structure-Supervised Kernel Machine Association Testing

4.3.1.1 Kernel Machine Regression Models

We consider a study of n subjects with phenotype $Y_{n \times 1} = (Y_1, \dots, Y_n)^T$ which we assume to follow an exponential family distribution with canonical link $g(\mu) = E(Y|X, G)$. $G_{n \times M}$ is a matrix of genotype information and $X_{n \times p}$ is a matrix of non-genetic covariates.

A global kernel machine (KM) regression model to associate the relationship between phenotype and genotype, accounting for non-genetic covariates, is of the form $g(\mu) = \beta_0 + X\beta_X + h(G)$, where $h(G)$ is a smooth function of genotype, which may be expressed in dual form as $h = K\alpha$ (Kimeldorf and Wahba, 1971). Here $K_{n \times n}$ is a kernel summarizing genetic similarity amongst all pairs of individuals and $\alpha_{n \times 1}$ is a vector of unknown parameters with length equal to the number of subjects. This may be expressed as a random effects term, $h \sim N(0, \tau K)$.

A test for global association over a gene set looks at the null hypothesis $H_0 : h(G) = 0$. In dual form, this is equivalent to the null hypothesis $H_0 : \tau = 0$, examined using a variance component score test (Liu et al., 2008, 2007).

After we've already determined gene-level significance, it may be of interest to prioritize the rare variants within a gene. We propose local variant-level kernel machine models which test for association between a given variant and the phenotype of interest, borrowing information for neighboring variants with amount of borrowing determined by nearness in structural space, expressed in a local kernel matrix. For the sake of this paper, we focus on 3-dimensional protein structure, but note that the model could be easily extended to other structure types, such as chromosome looping structure or network structure.

For each variant $j = 1, \dots, M$, a variant-level KM model is of the form $g(\mu) = \beta_0 + X\beta_X + h_j$, $h_j = K_j\alpha_j$, $\alpha_j \sim N(0, \tau K_j^{-1})$ for invertible local kernel K_j generated using protein tertiary structure.

We note the parallel to a linear model framework, where $g(\mu) = \beta_0 + X\beta_X + h_j$, $h_j = G\beta$, $\beta \sim N(0, \tau R)$.

This notation emphasizes that the structure is included only as a prior.

4.3.1.2 Obtaining Protein Tertiary Structure Information

Protein tertiary structure and corresponding amino acid coordinates in the 3D protein space are obtained from the RCSB Protein Data Bank (PDB) (Berman et al., 2000).

In order to do so, we must first obtain annotation for all exonic protein-coding variants of interest in order to map the variant position on the DNA sequence to an associated amino acid and corresponding position on the amino acid sequence. When variants of interest are well studied (i.e., with reference single nucleotide polymorphism (SNP) IDs and accessions), this mapping can be obtained using the Innovation Center for Biomedical Informatics (IBCI) structure database online query tool SNP2Structure (Wang et al., 2015a). SNP2Structure allows us to obtain a full list of SNP to amino acid mapping for all known PDB entries.

When we only know variant position and NCBI genome build, further annotation is necessary. This can be done using the functional annotation tool ANNOVAR (Wang et al., 2010), which requires as input a file listing the chromosome, start and stop position, and major and minor allele for each variant of interest, e.g. in Variant Calling Format (VCF). From this we can extract type of mutation (nonsynonymous single nucleotide polymorphism (SNV), synonymous SNV, CNV, frameshift deletion, insertion, etc.), gene name, RefSeq accession number, exon, DNA position and mutation, and amino acid position and mutation. This must be manually checked against PDB entries for proportion of variants included, as well as aligned and manually checked. The alignment is an important step as PDB entries usually do not cover the whole protein structure and may have their own position numbering.

Once annotation has been completed, we must choose the “best” PDB entry, which maximizes the number of variants of interest with known 3D protein position foremost, and secondarily is most representative of a true wild type protein, with relatively high resolution and low percent

outliers, clashscore, and Rfree score. We also note the possibility of combining information from multiple PDB entries, combining positional information, e.g. using the align command in the PyMOL software (Schrödinger, LLC, 2015) in order to get a fuller coverage of all variants of interest.

4.3.1.3 Building a Structural Kernel

From the 3D Cartesian coordinates, we can build a pairwise distance matrix $D_{M \times M}$, where $D_{(l,k)} = d(l,k)$ is the Euclidean distance between two variants on the protein tertiary structure, $d(l,k) = 0$ for $l = k$. Similar to the Tango statistic (Tango, 2010), we summarize the amount of borrowing from neighboring variants using a variant similarity matrix based on the Gaussian kernel of the form $R_{M \times M} = e^{-D^2/2h^2}$. This allows smooth, gradual drop off in amounts of borrowing between variants as distance between variants increases, rather than hard thresholding as in scan statistic approaches of Kulldorff (1997) and Ionita-Laza et al. (2012) (Schaid et al., 2013).

Here h is the unknown scale parameter, which determines how fast of a drop off, or how much structural information is incorporated. Rather than parameter tuning, we follow a similar approach to Tango (2010) and Schaid et al. (2013) and examine a grid of values. Rather than proportion of maximum distance, however, we look at a proportion of the standard deviation of all pairwise Euclidean distances between variants, $h = c \cdot s d(D)$. The idea behind using a function of the standard deviation of pairwise distances is borrowed from nonparametric theory for choosing the optimal bandwidth of kernels. It is used as a proxy for separating clusters of variants, such that variants within the same cluster on the protein structure are likely to borrow information from each other, while variants in different clusters have minimal effect on one another. Larger c values encourage large amounts of sharing between neighboring variants, and approaches a global-level test as $c \rightarrow \infty$. Smaller c values decrease the amount of borrowed information, and approach a single-variant test as $c \rightarrow 0$.

We choose to consider a grid of c values from 0 to 0.5, representing half of the standard deviation in distances between the variants. The choice of $c = 0.5$ as the maximum borrowing distance is due

to our desire to borrow information only from variants who are neighbors and may be considered to cluster together in the protein tertiary space. Choosing larger maximum c values forces sharing from all variants, whether they are close enough to be expected to share biological architecture or not, which may lead to higher false positive rates, finding significance of noncausal variants. An illustration of variant sharing for smaller and larger c values may be found in Figure 4.2.

By using a grid of scale values, we can allow the data to speak for itself, adaptively choosing the best scale, borrowing information only if the data is consistent with the effective prior set by the protein structure. Therefore, structure can be used as a prior where, rather than forcing sharing of information between variants which may not have related genetic effects or may in fact confer opposite effects on phenotype, neighboring information can be shared only if there appears to be sufficient consistency/evidence to do so. Variant similarity matrix R therefore works to induce smoothing between neighboring variants, decaying with distance between variants, rather than inducing correlation.

Finally, we can create local kernel K_j to summarize the genetic similarity between all pairs of individuals at variant j ($j = 1, \dots, M$) and neighboring variants, weighted by distance and, optionally, another feature such as minor allele frequency (MAF). One way to construct this kernel is as a weighted sum of partitions of a non-weighted genetic similarity matrix as $K_j = \sum_{k=1}^M r_{kj} w_k S^{(k)}$, where r_{kj} is the $(k, j)^{th}$ entry of R , w_k is a weight for variant $k = 1, \dots, M$, and $S^{(k)}$ is the genotype similarity at variant k . As with global kernels, there are many options for the form of this kernel. Some commonly used forms include the IBS kernel, which for additive effects has entries $S_{i,i'}^{(k)} = 2 - |G_{ik} - G_{i'k}|$, and the linear kernel with entries $S_{i,i'}^{(k)} = G_{ik} G_{i'k}$, where G_{ik} is the $(i, k)^{th}$ entry of G . Local kernels may also be formed by incorporating the j^{th} row/column of R into the typical kernel form, e.g. $K_j = G W^{1/2} \text{diag}(R_j) W^{1/2} G^T$ for the linear kernel, $K_j = (1 + G W^{1/2} \text{diag}(R_j) W^{1/2} G^T)^d$ for the polynomial kernel, and $K_j = e^{[-\psi \sum_{k=1}^M (G_k - G_j)^T R^{-1} (G_k - G_j)]}$ for the Gaussian kernel. Here $W = \text{diag}(w_1, \dots, w_M)$ is a diagonal weight matrix, G_j is the j^{th} row of G , $\text{diag}(R_j)$ is a diagonal matrix of the j^{th} row of R , and ψ is the kernel bandwidth.

4.3.1.4 Local Kernel Score Test

Once we obtain a local (variant-level) kernel, a score test statistic for $H_0 : \tau = 0$ for a fixed c is: $T_{j,c} = \frac{1}{n}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)^T K_{j,c}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n) = \frac{1}{n}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n) Z_{j,c} Z_{j,c}^T (\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)^T$ where $\hat{\epsilon}_i$ is the fitted residual for the KM model under the null hypothesis of no genetic effect, $\hat{\epsilon}_i = Y_i - g^{-1}(X\hat{\beta}_X)$. Under the null hypothesis, $\frac{1}{\sqrt{n}}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n) Z_{j,c} \xrightarrow{d} N(0, \Sigma)$, $\Sigma = E(\psi\psi^T)$, as defined in Appendix B. Equivalently, we note that $T_{j,c} \xrightarrow{d} \sum_d \xi_d \chi_1^2$, a mixture of Chi-square distributions, weighted by the positive eigenvalues of Σ . Thus, for a fixed c , localized single variant p-values can be calculated, much like global level test p-values, using the Davies method (Davies, 1980) or by resampling.

4.3.1.5 Minimum P Test Statistic and Resampling Procedure

Other than the incorporation of a localized kernel, the distinction from global KM testing lies in our borrowing weight parameter c . Because it is not known for real data how much variants should borrow from one another, we perform this test over a grid of c values, and allow the data to speak for itself and choose the necessary amount of borrowing. This is done using a minimum p value approach.

If we are interested in m variants and ℓ c values, we can calculate $m \cdot \ell$ test statistics $T_{j,c} = \frac{1}{n}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n) Z_{j,c} Z_{j,c}^T (\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)^T$, or equivalently $T_{j,c} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\psi}_{j,c} \right)^T \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\psi}_{j,c} \right)$. Our test statistic asymptotically follows the mixture of χ^2 distribution $\sum_d \xi_d \chi_1^2$, where ξ_d are the nonzero eigenvalues of $\Sigma_{j,c} = \Sigma_{Z(j,c)}$, estimated by $\hat{\Sigma}_{Z(j,c)}$. For each, corresponding p-values can be obtained using the Davies method or by resampling the nonzero eigenvalues.

In order to get an overall p-value for each variant (over the different c values), we first create a minP statistic by taking the minimum of p-values over all c values considered, $\min P_j = \min\{p_{j,c_1}, \dots, p_{j,c_\ell}\}$.

The p-value of the minP statistic is obtained via resampling, by perturbing the influence functions. We generate B random vectors of length n : O_1, \dots, O_B from a standard Normal distribution

and create B perturbed test statistics for each variant/ c value combination

$$T_{j,c}^{(b)} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n O_{ib} \hat{\psi}_{j,c} \right)^T \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n O_{ib} \hat{\psi}_{j,c} \right), b = 1, \dots, B$$

The p-values for each of these perturbed statistics, $p_{j,c}^{(b)}$, can be calculated using Davies method or resampling the nonzero eigenvalues of $\hat{\Sigma}_{j,c}$, $b = 1, \dots, B$. We can then calculate B minP statistics for each variant: $\min P_j^{(b)} = \min_c \{p_{j,c}^{(b)}\}$, $b = 1, \dots, B$.

Finally, our minP p-value for each variant is $p_j^* = \frac{1}{B} \sum_{b=1}^B I(\min P_j^{(b)} < \min P_j)$, where $I(\cdot)$ is the indicator function. These can be compared to the single SNP p-values p_1, \dots, p_m . Adjustments for multiple testing (either via Bonferroni/FDR or the resampling methodology of Westfall and Young (1993)) should be applied to both minP and single SNP p-values.

4.4 Simulation Study

4.4.1 Simulation Set Up

We design a simulation following the work of Song et al. (2012), which examined the effect of single nucleotide polymorphisms (SNPs) within Phospholipase A2 Group VII, *PLA2G7*, on protein function and enzyme activity of Lipoprotein-associated phospholipase A₂ (Lp-PLA₂). Song et al. (2012) performed Sanger resequencing of *PLA2G7* on 2000 individuals from the CoLaus study, a clinical trial examining psychiatric, cardiovascular, and metabolic disorders in 6188 Caucasians aged 35-75 from Lausanne, Switzerland (Firmann et al., 2008, Preisig et al., 2009), and found variants that were deemed likely to be noise variants clustered in the core of the protein tertiary space, whereas causal variants were all near to the surface (Song et al., 2012). We utilize the sequencing genotypes from this study, and the protein tertiary structure from PDB entry 3F96 (Samanta et al., 2009), obtaining 3D coordinates by averaging the positions of all side chains for each variant. We focus only on the 13 variants from their study that were both rare (or less common) and had available protein coordinate

information from PDB. A summary of the studied variants and their relative positions in the protein tertiary structure can be found in Figure 4.1 below.

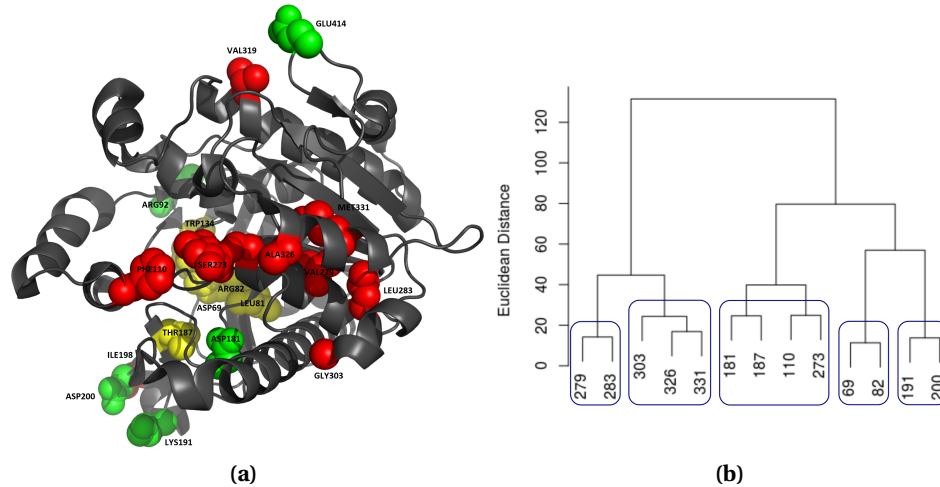


Figure 4.1 *PLA2G7* (a) rare variants location on the protein tertiary structure, and (b) corresponding Euclidean distance-based clustering of variants. Protein tertiary structure published with permission from Dr. Fourches.

We generate phenotype for 1,000 individuals from exponential model $g(\mu) = \beta_0 + X\beta_X + G\beta_G$, with identity link $g(\mu) = \mu$, for continuous traits ($y \stackrel{iid}{\sim} N(\mu, \sigma = 1)$), and logit link $g(\mu_i) = \frac{e^\mu}{1+e^\mu}$ for binary traits. For simplicity, we do not include any non-genetic covariates (i.e., $\beta_X = 0$), and use intercept $\beta_0 = 0.5$ for continuous traits, and $\beta_0 = -0.05$ for binary traits. Coefficients for genetic variants are generated as $\beta_G = (b \cdot |\log_{10}(MAF_k)| I\{\text{marker } k \text{ is causal}\})$, where MAF_k is the minor allele frequency of marker $k = 1, \dots, 13$, to upweight the effect of the rare variants. We choose a variety of scenarios with different causal variants from the 13 studied variants (9 rare, 4 less common, as examined in Song et al. (2012)) based on the true protein structure. We first consider scenarios where the generated causal variants cluster close together on the tertiary protein structure, with varying closeness on the amino acid sequence, choosing: (69, 82), (303, 326, 331), (191, 200), and (110, 273) to be causal variant clusters. We further consider scenarios where only two of three closely clustered

variants are causal, choosing: (326, 331) as causal and (303, 326) as causal. Finally, we consider a scenario where two clusters of variants are causal: one, (69, 82), positively conferring phenotype risk, and one, (303, 326, 331), negatively conferring phenotype risk. We also examine the type I error, where no variants were deemed causal.

All scenarios were examined with two levels of gridding for scale variable c : with 3 values considered $c = (0, 0.25, 0.5)$, and with 6 values $c = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$, using burden and linear kernels for fitted model, with 500 simulations per scenario, and 1000 resamples.

We compare our local kernel test to the single variant score test and the scan statistic of Ionita-Laza et al. (2012). The scan statistic is only used as a comparison for the case/control data scenario, however, as it is not directly applicable to continuous phenotype data.

4.4.2 Simulation Study Results

4.4.2.1 Variant Borrowing Example

In Figure 4.2 we see a visual representation of the proportion of information borrowed from each neighboring variant (obtained from the R matrix), over various values of c , for the local test for *PLA2G7* variant 110, overlaid on the variants' relative positions in the 3-dimensional protein space. We see that for smaller values of c (i.e., $c \in (0, 0.4)$), variant 110 borrows very little, if any, information from any other variants. As c increases, it begins borrowing more and more signal from variants further away on the 3-dimensional space, as can be seen from figures 4.2 and 4.1. For large values of c (e.g., $c \geq 2$), the local test is using information from all considered variants, and begins approaching a global test as c approaches 10.

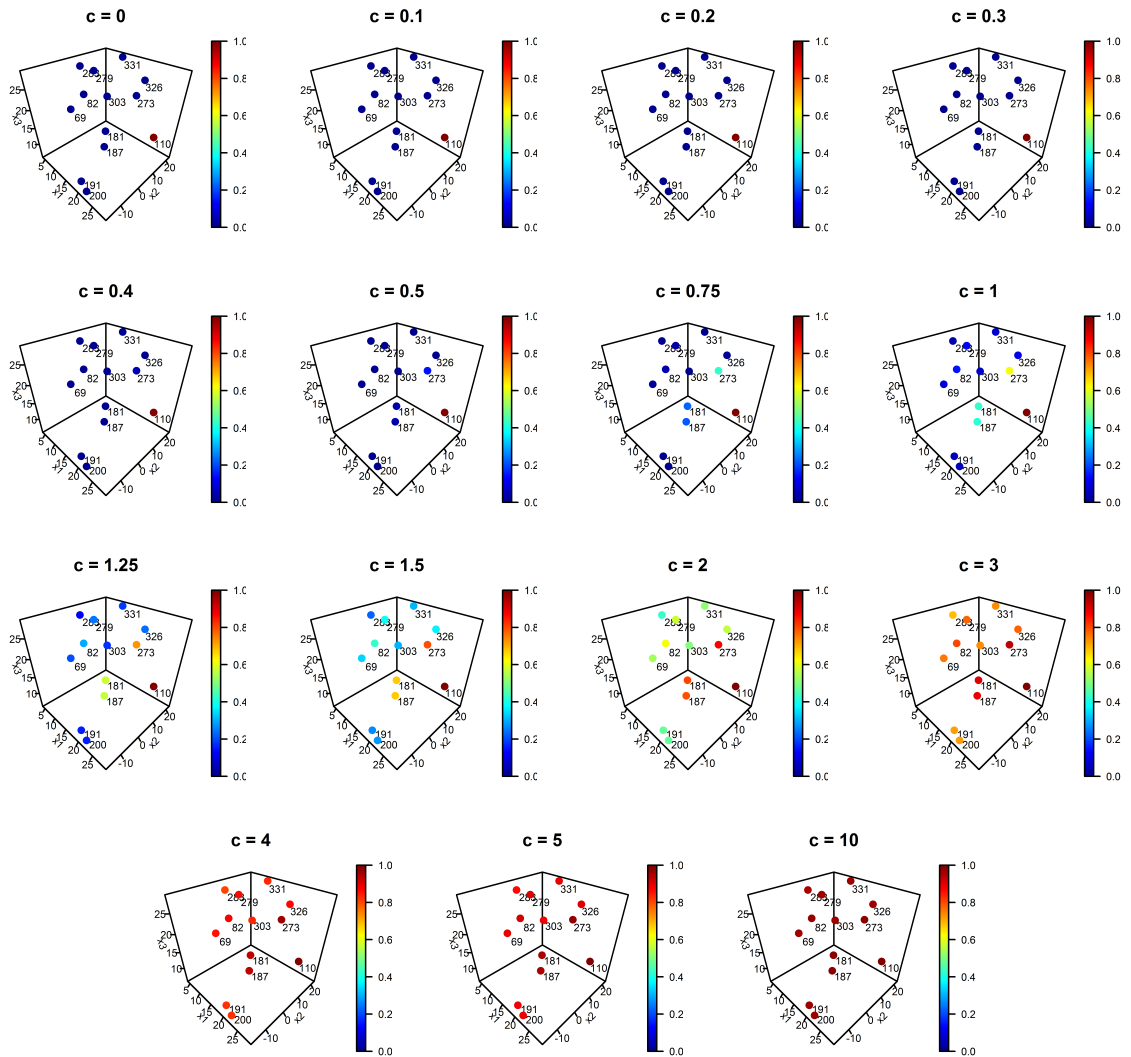


Figure 4.2 Amount of borrowing from neighboring variants for *PLA2G7* variant 110 for different values of c .

4.4.2.2 Single Variant Score and Localized Kernel Score Tests

The results of the scenarios for tests weighted by minor allele frequency with a grid of 6 c values for quantitative and binary traits with local burden and local linear kernels are presented in tables 4.1-

4.4 below. Results of unweighted tests and those with 3 grid values for c are similar and may be found in Appendix C.

4.4.2.3 Local Burden Kernel Test Results

Table 4.1 Type I error and power for single-variant and local burden kernel tests for quantitative traits over various simulation scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
type I error	0	N/A	0.044	N/A	0.042
(69,82)	0.35	0.736	0.103	0.971	0.113
	1.05	1.000	0.147	1.000	0.200
	0.35	0.802	0.052	0.930	0.053
(303,326,331)	0.55	0.987	0.059	1.000	0.068
	0.8	0.996	0.050	1.000	0.051
	1.5	1.000	0.060	1.000	0.064
(191,200)	0.45	0.910	0.049	0.977	0.049
	0.6	0.993	0.051	0.998	0.051
	0.35	0.778	0.046	0.904	0.045
(326,331)	0.55	0.990	0.049	1.000	0.049
	0.35	0.870	0.051	0.879	0.053
	0.55	0.990	0.055	0.990	0.067
(303,326)	0.35	0.823	0.135	0.970	0.157
	0.55	0.990	0.164	1.000	0.213
	0.35	0.823	0.135	0.970	0.157
(69,82) pos, (303,326,331) neg	0.55	0.990	0.164	1.000	0.213

In tables 4.1 and 4.2 we see that both our local kernel method with burden kernel and the single variant score test are quite conservative, with type I error rates of approximately 0.04 for quantitative continuous and binary traits. Our method is more powerful than the single variant score test in all scenarios except for the case where (303,326) are the causal variants, where the two methods perform relatively equivalently. The false positive rate is slightly more elevated in the local kernel method, but is on the same level as the single variant test. We note an increased false positive rate

Table 4.2 Type I error and power for single-variant and local burden kernel tests for binary traits over various simulation scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
type I error	0	N/A	0.045	N/A	0.043
(69,82)	0.7	0.705	0.111	0.965	0.114
	0.85	0.821	0.126	0.998	0.130
(303,326,331)	0.6	0.725	0.050	0.827	0.050
	0.75	0.855	0.050	0.937	0.051
(191,200)	1.4	0.973	0.045	0.993	0.042
	1.75	0.993	0.045	0.998	0.043
(110,273)	0.85	0.858	0.047	0.934	0.047
	1.1	0.961	0.049	0.987	0.048
(326,331)	0.9	0.923	0.047	0.972	0.045
	1.1	0.980	0.048	0.993	0.047
(303,326)	0.6	0.779	0.048	0.774	0.050
	0.75	0.894	0.049	0.887	0.054
(69,82) pos,	0.6	0.732	0.126	0.892	0.135
(303,326,331) neg	0.75	0.862	0.144	0.969	0.158

for both the single variant method and the local kernel method for scenarios in which variants at amino acid positions 69 and 82 are causal, which could be due to high genetic correlation between variant 82 and variant 181, as can be seen in the variant correlation plot in Figure 4.3 below.

In Figure 4.4, we see that our method chooses not to borrow information (i.e., chooses $c = 0$) a majority of the time for noncausal variants, and mostly chooses either large amounts of borrowing between variants ($c = 0.4$) or no borrowing ($c = 0$) for causal variants.

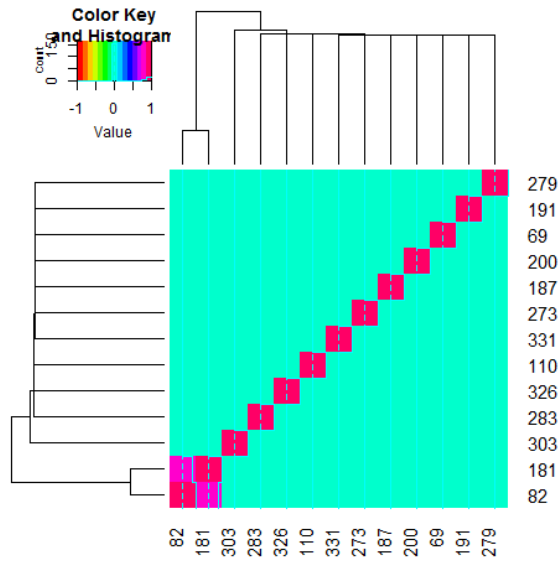


Figure 4.3 Variant correlation for *PLA2G7* SNVs

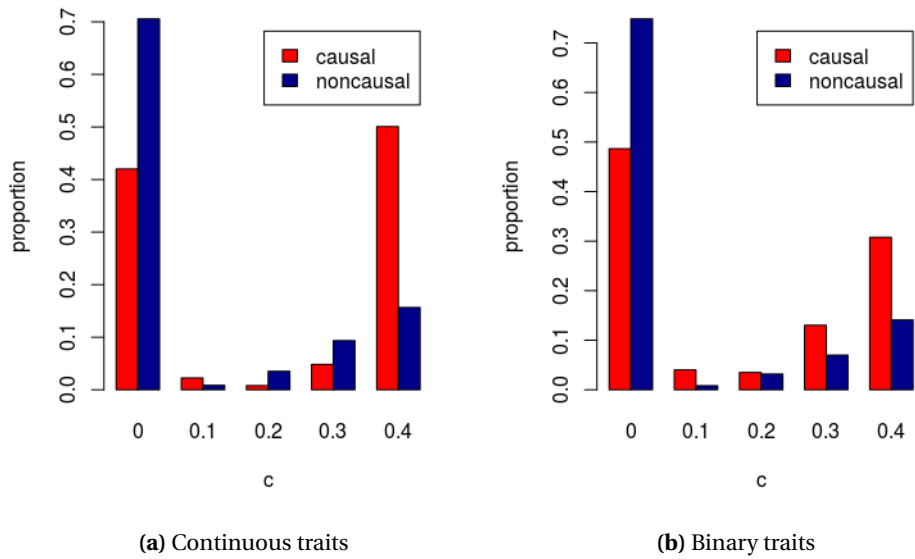


Figure 4.4 Summary of proportion of times a given c value was chosen as optimal for the local burden kernel test for causal and noncausal variants

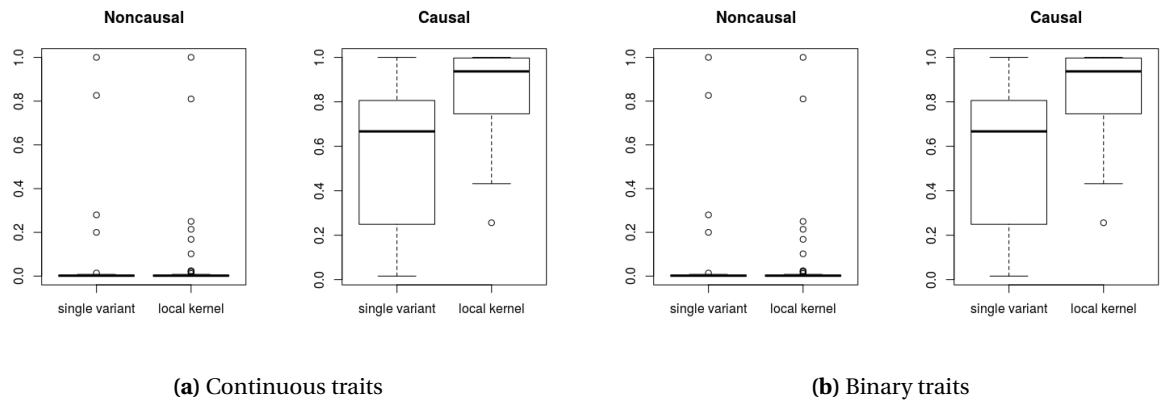


Figure 4.5 Variant selection probabilities for single variant and local burden kernel tests, separated by noncausal variants (effective type I error), and causal variants (effective power)

4.4.2.4 Local Linear Kernel Test Results

Table 4.3 Type I error and power for single-variant and local linear kernel tests for quantitative traits over various simulation scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
type I error	0	N/A	0.046	N/A	0.045
(69,82)	0.35	0.743	0.107	0.951	0.148
	1.05	1.000	0.151	1.000	0.333
(303,326,331)	0.35	0.807	0.053	0.943	0.071
	0.55	0.989	0.062	1.000	0.170
(191,200)	0.8	0.996	0.053	1.000	0.067
	1.5	1.000	0.062	1.000	0.146
(110,273)	0.45	0.918	0.051	0.977	0.057
	0.6	0.993	0.054	0.999	0.070
(326,331)	0.35	0.787	0.050	0.887	0.051
	0.55	0.990	0.053	0.999	0.071
(303,326)	0.35	0.876	0.053	0.914	0.110
	0.55	0.990	0.057	0.999	0.204
(69,82) pos,	0.35	0.827	0.138	0.970	0.259
(303,326,331) neg	0.55	0.992	0.165	1.000	0.488

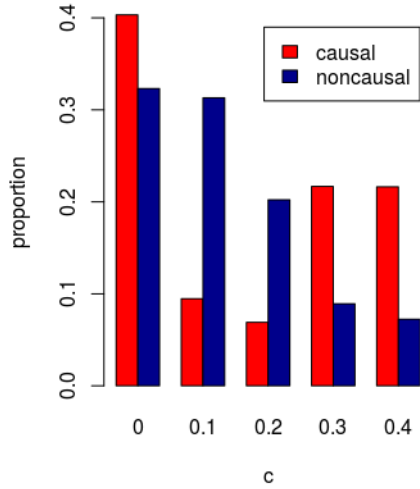
In tables 4.3 and 4.4 we see that the localized kernel test with linear kernel is also conservative, with type I error rate of approximately 0.045 for continuous and binary traits. As with the local burden tests, the local linear kernel tests outperform the single variant score test, here under all simulated scenarios. However, there are slightly increased false positive rates of our method compared to the single variant test, especially with larger causal effect size, and again both have difficulty when high genetic correlations exist between causal and noncausal loci, as with variants 82 and 181.

In Figure 4.6 we observe increased level of borrowing from noncausal variants, with noncausal variant simulations choosing all values of c at reasonable probabilities ($c = 0$ only chosen slightly

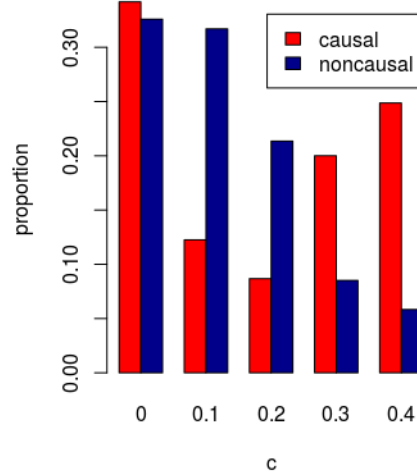
Table 4.4 Type I error and power for single-variant and local linear kernel tests for binary traits over various simulation scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
type I error	0	N/A	0.047	N/A	0.046
(69,82)	0.7	0.712	0.114	0.950	0.149
	0.85	0.822	0.128	0.991	0.180
(303,326,331)	0.6	0.734	0.052	0.836	0.059
	0.75	0.858	0.052	0.955	0.066
(191,200)	1.4	0.973	0.047	0.993	0.046
	1.75	0.993	0.047	0.998	0.047
(110,273)	0.85	0.866	0.050	0.930	0.051
	1.1	0.963	0.051	0.990	0.053
(326,331)	0.9	0.926	0.050	0.968	0.055
	1.1	0.980	0.051	0.995	0.056
(303,326)	0.6	0.785	0.050	0.815	0.078
	0.75	0.900	0.050	0.930	0.106
(69,82) pos,	0.6	0.742	0.128	0.893	0.178
(303,326,331) neg	0.75	0.868	0.146	0.976	0.233

over 30% of the time), decreasing with increasing c . We also see increased amount of causal variants that do not borrow strength from other variants.

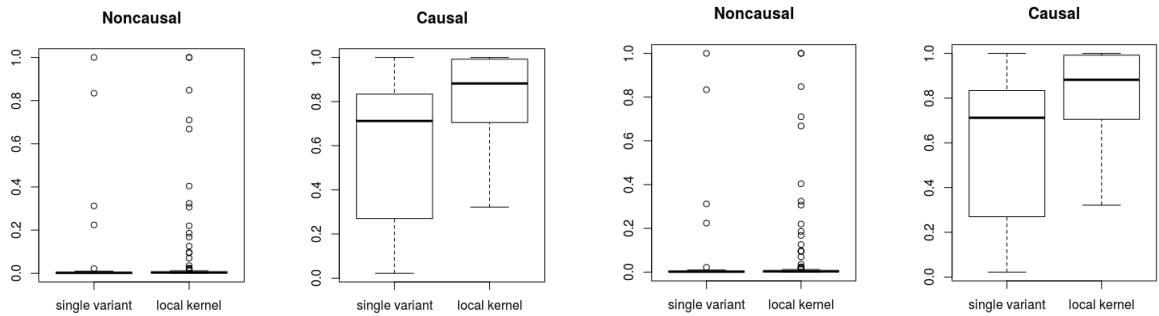


(a) Continuous traits



(b) Binary traits

Figure 4.6 Summary of proportion of times a given c value was chosen as optimal for the local linear kernel test for causal and noncausal variants



(a) Continuous traits

(b) Binary traits

Figure 4.7 Variant selection probabilities for single variant and local linear kernel tests, separated by non-causal variants (effective type I error), and causal variants (effective power)

4.4.2.5 Scan Statistic

In Table 4.5 we see a summary of the performance of the scan statistic, including the proportion of times (over 500 simulations) the best window was significant at the $\alpha = 0.05$ level, the average false positive and true positive rate of the test (i.e., the proportion of times the noncausal and causal variants were included in the best window and the best window was significant, over 500 simulations), the proportion of times (over 500 simulations) the best window chosen was the exact true model (defined as one where all generated causal variants are included in the model with no noise variants), and a summary of window sizes (mean, standard deviation, median, and range), where window size is said to be zero for runs where the most promising window is not significant.

We see that the scan statistic performs very well at choosing true causal variants within the best window when variants are relatively nearby on the amino acid sequence compared to other variants considered, but, as would be expected, can never choose the true best window when variants are not sequential on the amino acid sequence, with low true positive rate in these scenarios. Much like the single variant and local kernel tests, the scan statistic has well controlled false positive rate with the exception of the scenarios where causal variants are genetically correlated with noncausal variants. On average the chosen most promising window is relatively close to the true window size, again generally overestimating window size only in scenarios with high genetic correlation between causal and noncausal variants.

Table 4.5 Scan statistic results for binary traits, summarized over 500 replications

causal variants	b	% significant window	true pos. rate	false pos. rate	% correct model	window size		
						avg (std. err)	median	range
type I error	0	0.07	N/A	0.024	0.93	0.3 (0.06)	0	(0,12)
	(69,82)	0.7	0.81	0.767	0.135	0.228	3.02 (0.1)	4
(303,326,331)	0.85	0.92	0.911	0.136	0.33	3.32 (0.08)	4	(0,11)
	0.6	0.84	0.784	0.034	0.57	2.69 (0.07)	3	(0,9)
(191,200)	0.75	0.95	0.917	0.026	0.76	3.01 (0.05)	3	(0,9)
	1.4	0.99	0.982	0.013	0.88	2.11 (0.03)	2	(0,9)
(110,273)	1.75	0.99	0.995	0.008	0.93	2.07 (0.02)	2	(0,6)
	0.85	0.71	0.395	0.055	0	1.39 (0.09)	1	(0,12)
(326,331)	1.1	0.90	0.489	0.047	0	1.49 (0.08)	1	(0,9)
	0.9	0.93	0.897	0.005	0.85	1.84 (0.03)	2	(0,9)
(303,326)	1.1	0.99	0.97	0.002	0.95	1.96 (0.02)	2	(0,6)
	0.6	0.74	0.707	0.048	0.45	1.95 (0.08)	2	(0,12)
(69,82) pos,	0.75	0.91	0.886	0.042	0.66	2.23 (0.06)	2	(0,9)
	0.6	1	0.397	0.705	0	7.62 (0.12)	9	(2,11)
(303,326,331) neg	0.75	1	0.398	0.719	0	7.74 (0.12)	9	(2,10)

4.5 Application to ACCORD Study

4.5.1 ACCORD Trial Background

The Action to Control Cardiovascular Risk in Diabetes (ACCORD) clinical trial was a multi-clinic trial with intent to test for effect of intensive glycemic, blood pressure, and fenofibrate treatments versus their corresponding standard controls on cardiovascular disease endpoints in diabetic subjects (Buse et al., 2007, Genuth and Ismail-Beigi, 2012, Goff et al., 2007, Marvel et al., 2017). The trial enrolled 10,251 subjects with type II diabetes and a risk or history of cardiovascular disease (CVD) from 77 centers around North America and found that intensive treatments were not found to be beneficial and even potentially risky for the CVD endpoints studied (Genuth and Ismail-Beigi, 2012). A secondary goal of this trial was to understand how genotype plays into individual risk for diabetes and CVD by trying to understand the relationship between a subject's genes and their diabetes and CVD risk factors, such as blood lipid levels and hypertension (Marvel et al., 2017). A previous study by Marvel et al. (2017) looked at testing for association between over 8 million genetic variants covering 16,538 genes and blood lipid levels in 7844 genotyped ACCORD trial participants using a multitude of statistical methods (linear regression for common variants, and modified burden tests: weighted sum statistic (Madsen and Browning, 2009), RVT1, and RVT2 (Morris and Zeggini, 2010), as well as SKAT (Wu et al., 2011) and SKAT-O (Lee et al., 2012a) for rare variants, correcting for multiple testing by combining p-values for each gene using the "corrected Lancaster procedure" (Hongying Dai and Cui, 2014)). Using the rare variant analysis methodology, they found 11 genes to be significantly associated with blood lipid levels (in terms of total cholesterol (TC), low-density lipoprotein (LDL), high-density lipoprotein (HDL), and total triglycerides (TG), with 2-22 rare variants/gene. Proprotein convertase subtilisin kexin 9 (*PCSK9*), which was found to be significantly associated with total cholesterol and LDL (Liu et al., 2014, Marvel et al., 2017) using rare variant analysis is particularly of interest, as it has well-studied protein structure and is under

study for targeted therapy for reducing LDL levels in diabetics (Lange et al., 2014, Marvel et al., 2017, Stroes et al., 2014).

4.5.2 ACCORD Analysis

We apply our method to the baseline data from the ACCORD clinical trial (Goff et al., 2007). We focus on the baseline pre-intervention data, aiming to prioritize variants within the *PCSK9* gene, which was found to have significant gene-level association with LDL (Liu et al., 2014, Marvel et al., 2017) and total cholesterol levels in diabetics (Marvel et al., 2017).

Following the work of Marvel et al. (2017), we consider rare variants to be those with $< 3\%$ minor allele frequency (MAF), and use only individuals with less than 15% missingness. Missing genotype information was imputed previously by Marvel et al. (2017). All variants studied were functional and annotated in build GRCh37p13. We use the IBCI data base SNP2Structure (Wang et al., 2015a) to match variant RSID with corresponding amino acid position in the 2D protein sequence as well as which protein chain they are situated on in all PDB entries for *PCSK9*. We use carbon alpha coordinates from PDB entry 4K8R (Schiele et al., 2014) as it was found to be most representative of wild type protein that also maximized the number of variants of interest with known protein tertiary position, 19 of 22. Local kernel tests were carried out using the burden kernel with 26 baseline covariates, including patient age, gender, body mass index (BMI), presence of cardiovascular history, trial treatment arm assignment, top three principal components of ethnic background, years since diabetes and since hyperlipidemia diagnoses, fasting glucose level, and indicators of use of different treatments (e.g., insulin, lipid-lowering drugs, etc.). A full list of these covariates can be found in the Supplementary Materials of Marvel et al. (2017). A summary of variants used and corresponding 3D coordinates from PDB can be found in Table D.1 in Appendix D.

4.5.3 ACCORD Results

The rare variants from *PCSK9* were globally significant, with SKAT test statistic and p-value of 2.74×10^5 and 8.11×10^{-3} , respectively. SKAT-O had similar significance, with test statistic and p-value of 0.005 and 0.008, respectively. Global burden-based methods were also significant, with p-values of 0.008, 0.018, and 0.024 for RVT1, RVT2, and the weighted sum tests, respectively. The scan statistic method applied to this data, consistent with the global significance from SKAT and SKAT-O, chose a rather large best window which included variants 93-425. This window was not significant after accounting for multiple testing, however, with p-value 0.607.

Two variants, with RSIDs rs28362263 and rs28362270 at amino acid positions 443 and 553 were found to be significant using both the single variant score test and the local kernel approach. Three others, as can be seen in Table 4.6, were identified only by the local kernel approach with burden kernel: rs143117125, rs146471967, and rs28362263 at amino acid positions 157, 391, and 425, respectively. The local linear kernel test additionally found variant rs149311926 at position 554 to be significant, borrowing strength from significant variant 553; however, this was not significant after correcting for False Discovery Rate.

Table 4.6 *PCSK9* rare variant single SNP and localized kernel test summary using PDB entry 4K8R. Significant results are given in bold font. Q-values are calculated assuming $\pi_0 = 0$ (e.g., Benjamini-Hochberg corrected values) due to low variant pool size.

AA coord	SNP ID	single variant		local burden kernel			local linear kernel		
		p-value	q-value	best c	p-value	q-value	best c	p-value	q-value
93	rs151193009	0.159	0.389	0	0.195	0.337	0.4	0.171	0.271
96	rs185392267	0.397	0.471	0.5	0.391	0.408	0.4	0.309	0.367
157	rs143117125	0.213	0.389	0.4	0.023	0.152	0.5	0.099	0.257
252	rs149139428	0.554	0.619	0.5	0.147	0.323	0.5	0.143	0.271
253	rs72646508	0.130	0.389	0.5	0.135	0.323	0.5	0.134	0.271
279	rs72646509	0.318	0.431	0.4	0.408	0.408	0.1	0.498	0.498
283	rs72646510	0.637	0.672	0.5	0.262	0.356	0.5	0.163	0.271
391	rs146471967	0.091	0.389	0.3	0.010	0.152	0.2	0.076	0.241
417	rs143275858	0.318	0.431	0.5	0.106	0.323	0.5	0.108	0.257
425	rs28362261	0.051	0.322	0.5	0.034	0.162	0.5	0.042	0.241
443	rs28362263	0.043	0.322	0.5	0.048	0.182	0.1	0.066	0.241
466	rs72646517	0.988	0.988	0.5	0.379	0.408	0.5	0.367	0.387
469	rs141502002	0.225	0.389	0.5	0.237	0.346	0	0.232	0.315
498	rs145468572	0.317	0.431	0.5	0.153	0.323	0.5	0.281	0.356
525	rs140286279	0.347	0.440	0.4	0.374	0.408	0.5	0.075	0.241
553	rs28362270	0.016	0.306	0	0.024	0.152	0.3	0.018	0.241
554	rs149311926	0.124	0.389	0	0.194	0.337	0.4	0.026	0.241
619	rs28362277	0.183	0.389	0	0.218	0.345	0.5	0.190	0.278
659	rs147182054	0.201	0.389	0	0.359	0.408	0.5	0.328	0.367

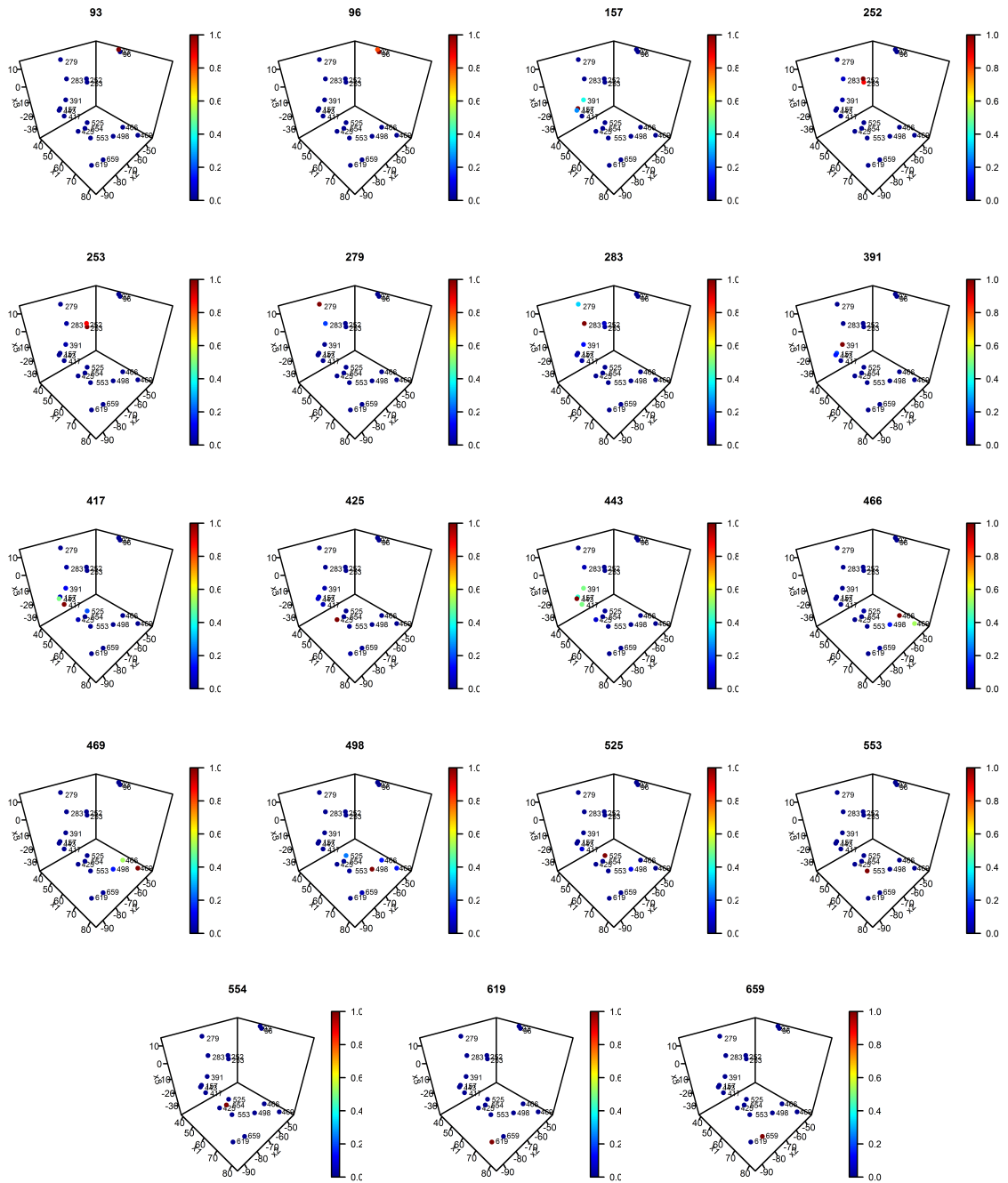


Figure 4.8 Variant borrowing for the best scale c for local kernel tests of association between rare variants in *PCSK9* and LDL

In Figure 4.8 we see how much signal each local burden kernel test borrowed from neighboring variants (incorporated as weights in the local kernel), with no borrowing in dark blue and full borrowing in dark red. We see that many of the significant variants from the local kernel test cluster close together and choose to borrow information from one another. In particular, we see mutual borrowing of information between variants 157, 391, and 443, with variant 443 also borrowing from variant 417. We also notice the close-by variant 425 borrowed information from 157, 417, and 443. We note that variant 417 does not choose to borrow information from its neighboring variants, and is not found to be significant. This demonstrates how the sharing of information between neighboring variants needs not be symmetric, allowing each variant to choose whether to borrow information based on how consistent the prior set by the local kernel is with the raw data. We further see how significant variant 553 (which is farther away from the aforementioned cluster of variants) has large signal on its own without needing or choosing to borrow from its neighboring variants.

In Figure 4.9 we see the 2 dimensional clustering of all considered variants from their 3 dimensional amino acid positions and corresponding variant positions on the 3 dimensional protein structure, with the two variants found using single variant and local kernel tests in light blue, and those found only using the local burden kernel test in pink. Here we note that the variants found were all reasonably closely clustered together relative to the rest of the variants. Further, the newly identified variants using the local burden kernel are all the closest to the protein-binding domain for gene *LDLR* (low-density lipoprotein receptor), which is a known target for treatment aiming to lower LDL (Lagace, 2014).

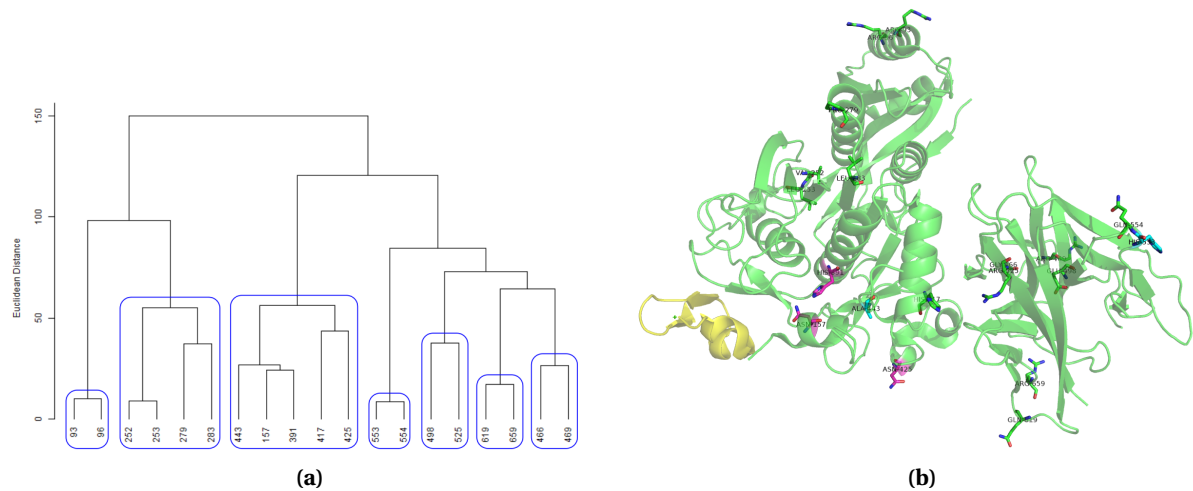


Figure 4.9 *PCSK9* (a) distance-based clustering of *PCSK9* variants, and (b) rare variants location on the protein tertiary structure of *PCSK9* and binding site of *LDLR* (in yellow); Significant associations from the single variant approach are blue and new associations found using the local kernel test are pink. Protein tertiary structure published with permission from Dr. Kuenemann.

4.6 Discussion

Rare variant association testing can be quite tricky due to low minor allele frequencies and weak individual effects. In this work we provide a framework to explicitly prioritize rare variants using local structural kernels. We describe how to incorporate protein tertiary structure into a formal testing framework, allowing for borrowing information from variants which are more likely to have similar effect on the disease trait of interest. Our approach is adaptive, utilizing the minimum p statistic to choose the best scale with which neighboring variants are used and at what amount. In doing so, we allow the data to speak for itself, only borrowing from neighboring variants if the data is consistent with the prior set by the protein structure.

This is highlighted from examination of variant-borrowing maps. We see that while neighboring variants do tend to borrow from one another to gain strength, this borrowing only occurs when the data is supportive of the structure. We note that the borrowing is not equal for a pair of variants,

strengthening our understanding of structure as a prior only, and strengthening our argument for having a data adaptive scale c that allows the data to speak for itself rather than forcing sharing of unrelated variants who may have clear differing effects.

We find that our method has better power to detect rare variant associations while maintaining similar false positive rates to single variant score testing. Both methods have difficulties with inflated false positive, however, when causal variants are highly genetically correlated with noncausal variants, though this is regardless of incorporation of protein structure information. The scan statistic also does a reasonable job of finding the true causal loci, but is unable to select the correct causal variants when they are not all within a nearby window on the DNA sequence.

Applying our local kernel method to the ACCORD clinical trial, we are able to pinpoint three new rare variants that were not found by single variant testing, all near to the protein-binding domain between *PLA2G7* and *LDLR*. This highlights the strength of our method to find additional signal that cannot be found by existing rare variant association testing techniques.

While we only demonstrate our approach to variants whose position in 3-dimensional protein structure are known, variants with no known structure can also be incorporated by considering the Euclidean distance between it and all other variants to be infinite, effectively using a single variant test for these variants. Further, our method can be straightforwardly extended to incorporate other types of biological information, such as functional annotation, chromosome looping structure (e.g., from 3C or Hi-C technology), or network structure.

REFERENCES

- Achim, A. M., Maziade, M., Raymond, É., Olivier, D., Mérette, C., and Roy, M.-A. (2009). How prevalent are anxiety disorders in schizophrenia? A meta-analysis and critical review on a significant association. *Schizophrenia Bulletin*, 37(4):811–821.
- Andreassen, O. A., Djurovic, S., Thompson, W. K., Schork, A. J., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., van de Bunt, M., Morris, A. P., et al. (2013). Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *The American Journal of Human Genetics*, 92(2):197–209.
- Ansari, A. R. and Bradley, R. A. (1960). Rank-sum tests for dispersions. *The Annals of Mathematical Statistics*, 31(4):1174–1189.
- Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G., Gormley, P., Malik, R., Patsopoulos, N., et al. (2016). Analysis of shared heritability in common disorders of the brain. *bioRxiv*, page 048991.
- Aschard, H., Vilhjálmsson, B. J., Greliche, N., Morange, P.-E., Trégouët, D.-A., and Kraft, P. (2014). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *The American Journal of Human Genetics*, 94(5):662–676.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, 28(1):235–242.
- Bodmer, W. and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, 40(6):695–701.
- Bolormaa, S., Pryce, J. E., Reverter, A., Zhang, Y., Barendse, W., Kemper, K., Tier, B., Savin, K., Hayes, B. J., and Goddard, M. E. (2014). A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLoS Genetics*, 10(3):e1004198.
- Breheeny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, 2(3):369.
- Breheeny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25(2):173–187.
- Broadaway, K. A., Cutler, D. J., Duncan, R., Moore, J. L., Ware, E. B., Jhun, M. A., Bielak, L. F., Zhao, W., Smith, J. A., Peyser, P. A., et al. (2016). A statistical approach for testing cross-phenotype effects of rare variants. *The American Journal of Human Genetics*, 98(3):525–540.
- Buckley, P. F., Miller, B. J., Lehrer, D. S., and Castle, D. J. (2008). Psychiatric comorbidities and schizophrenia. *Schizophrenia Bulletin*, 35(2):383–402.

- Buse, J. B., Group, A. S., et al. (2007). Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial: design and methods. *The American Journal of Cardiology*, 99(12):S21–S33.
- Cai, T., Tonini, G., and Lin, X. (2011). Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics*, 67(3):975–986.
- Casale, F. P., Rakitsch, B., Lippert, C., and Stegle, O. (2015). Efficient set tests for the genetic analysis of correlated traits. *Nature Methods*, 12(8):755–758.
- Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, 381(9875):1371–1379.
- Davenport, C. A., Maity, A., Sullivan, P. E., and Tzeng, J.-Y. (2017). A powerful test for snp effects on multivariate binary outcomes using kernel machine regression. *Statistics in Biosciences*, pages 1–22.
- Davies, R. (1980). The distribution of a linear combination of chi-square random variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(3):323–333.
- De Los Campos, G., Gianola, D., and Allison, D. B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics*, 11(12):880.
- Dekou, V., Gudnason, V., Hawe, E., Miller, G., Stansbie, D., and Humphries, S. (2001). Gene-environment and gene-gene interaction in the determination of plasma homocysteine levels in healthy middle-aged men. *Thrombosis and Haemostasis*, 85(1):67–74.
- Duchesne, P. and Lafaye De Micheaux, P. (2010). Computing the distribution of quadratic forms: further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54(4):858–862.
- Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446.
- Fawzi, M. H. and Fawzi, M. M. (2012). Disordered eating attitudes in Egyptian antipsychotic naive patients with schizophrenia. *Comprehensive Psychiatry*, 53(3):259–268.
- Ferreira, M. A. and Purcell, S. M. (2008). A multivariate test of association. *Bioinformatics*, 25(1):132–133.
- Fier, H., Won, S., Prokopenko, D., AlChawa, T., Ludwig, K. U., Fimmers, R., Silverman, E. K., Pagano, M., Mangold, E., and Lange, C. (2012). ‘Location, Location, Location’: a spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft palate. *Bioinformatics*, 28(23):3027–3033.

- Firmann, M., Mayor, V., Vidal, P. M., Bochud, M., Pécoud, A., Hayoz, D., Paccaud, F., Preisig, M., Song, K. S., Yuan, X., et al. (2008). The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovascular Disorders*, 8(1):6.
- Foulon, C. (2003). Schizophrenia and eating disorders. *L'Encephale*, 29(5):463–466.
- Galesloot, T. E., Van Steen, K., Kiemeny, L. A., Janss, L. L., and Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. *PLoS One*, 9(4):e95923.
- Genuth, S. and Ismail-Beigi, F. (2012). Clinical implications of the ACCORD trial. *The Journal of Clinical Endocrinology & Metabolism*, 97(1):41–48.
- Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, pages 1440–1450.
- Godart, N. T., Flament, M. F., Lecrubier, Y., and Jeammet, P. (2000). Anxiety disorders in anorexia nervosa and bulimia nervosa: co-morbidity and chronology of appearance. *European Psychiatry*, 15(1):38–45.
- Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99.
- Goff, D. C., Gerstein, H. C., Ginsberg, H. N., Cushman, W. C., Margolis, K. L., Byington, R. P., Buse, J. B., Genuth, S., Probstfield, J. L., Simons-Morton, D. G., et al. (2007). Prevention of cardiovascular disease in persons with type 2 diabetes mellitus: current knowledge and rationale for the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial. *The American Journal of Cardiology*, 99(12):S4–S20.
- Götestam, K. G., Eriksen, L., and Hagen, H. (1995). An epidemiological study of eating disorders in Norwegian psychiatric institutions. *International Journal of Eating Disorders*, 18(3):263–268.
- Harmon, D., Shields, D., Woodside, J., McMaster, D., Yarnell, J., Young, I., Peng, K., Shane, B., Evans, A., and Whitehead, A. (1999). Methionine synthase D919G polymorphism is a significant but modest determinant of circulating homocysteine concentrations. *Genetic Epidemiology*, 17(4):298–309.
- Hoff, P. (2012). Eugen Bleuler's concept of schizophrenia and its relevance to present-day psychiatry. *Neuropsychobiology*, 66(1):6–13.
- Hongying Dai, J. and Cui, Y. (2014). A modified generalized Fisher method for combining probabilities from dependent tests. *Frontiers in Genetics*, 5.
- Hsu, F., Sides, E., Mychaleckyj, J., Worrall, B., Elias, G., Liu, Y., Chen, W.-M., Coull, B., Toole, J., Rich, S., et al. (2011). Transcobalamin 2 variant associated with poststroke homocysteine modifies recurrent stroke risk. *Neurology*, 77(16):1543–1550.

- Hu, J. X., Thomas, C. E., and Brunak, S. (2016). Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics*, 17(10):615–629.
- Hudson, J. I., Hiripi, E., Pope, H. G., and Kessler, R. C. (2007). The prevalence and correlates of eating disorders in the National Comorbidity Survey Replication. *Biological Psychiatry*, 61(3):348–358.
- Hyndman, M., Bridge, P., Warnica, J., Fick, G., and Parsons, H. (2000). Effect of heterozygosity for the methionine synthase 2756 A → G mutation on the risk for recurrent cardiovascular events. *The American Journal of Cardiology*, 86(10):1144–1146.
- Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., and Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *The American Journal of Psychiatry*, 167(7):748–751.
- Ionita-Laza, I., Makarov, V., Buxbaum, J. D., Consortium, A. A. S., et al. (2012). Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. *The American Journal of Human Genetics*, 90(6):1002–1013.
- Kang, H., Sul, J., Service, S., Zaitlen, N., Kong, S., Freimer, N., Sabatti, C., Eskin, E., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354.
- Kaye, W. H., Bulik, C. M., Thornton, L., Barbarich, N., and Masters, K. (2004). Comorbidity of anxiety disorders with anorexia and bulimia nervosa. *The American Journal of Psychiatry*, 161(12):2215–2221.
- Khalil, R. B., Hachem, D., and Richa, S. (2011). Eating disorders and schizophrenia in male patients: a review. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, 16(3):e150–e156.
- Kiezun, A., Garimella, K., Do, R., Stitzel, N. O., Neale, B. M., McLaren, P. J., Gupta, N., Sklar, P., Sullivan, P. E., Moran, J. L., et al. (2012). Exome sequencing and the genetic basis of complex traits. *Nature Genetics*, 44(6):623–630.
- Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95.
- Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology*, 32(1):9–19.
- Korte, A., Vilhjálmsón, B. J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, 44(9):1066–1071.
- Kouidrat, Y., Amad, A., Lalau, J.-D., and Loas, G. (2014). Eating disorders in schizophrenia: implications for research and management. *Schizophrenia Research and Treatment*, 2014.

- Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496.
- Kwee, L., Liu, D., Lin, X., Ghosh, D., and Epstein, M. (2008). A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397.
- Lagace, T. A. (2014). PCSK9 and LDLR degradation: regulatory mechanisms in circulation and in cells. *Current Opinion in Lipidology*, 25(5):387.
- Lange, L. A., Hu, Y., Zhang, H., Xue, C., Schmidt, E. M., Tang, Z.-Z., Bizon, C., Lange, E. M., Smith, J. D., Turner, E. H., et al. (2014). Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *The American Journal of Human Genetics*, 94(2):233–245.
- Larson, N. and Schaid, D. (2013). A kernel regression approach to gene-gene interaction detection for case-control studies. *Genetic Epidemiology*, 37(7):695–703.
- Larson, N. B. and Schaid, D. J. (2014). Regularized rare variant enrichment analysis for case-control exome sequencing data. *Genetic Epidemiology*, 38(2):104–113.
- Leclerc, D., Wilson, A., Dumas, R., Gafuik, C., Song, D., Watkins, D., Heng, H., Rommens, J., Scherer, S., Rosenblatt, D., et al. (1998). Cloning and mapping of a cDNA for methionine synthase reductase, a flavoprotein defective in patients with homocystinuria. *Proceedings of the National Academy of Sciences*, 95(6):3059–3064.
- Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23.
- Lee, S., Wu, M. C., and Lin, X. (2012a). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775.
- Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., and Wray, N. R. (2012b). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542.
- Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321.
- Li, C., Yang, C., Gelernter, J., and Zhao, H. (2014). Improving genetic risk prediction by leveraging pleiotropy. *Human Genetics*, 133(5):639–650.
- Lievers, K., Afman, L., Kluijtmans, L., Boers, G., Verhoef, P., den Heijer, M., Trijbels, F., and Blom, H. (2002). Polymorphisms in the transcobalamin gene: association with plasma homocysteine in healthy individuals and vascular disease patients. *Clinical Chemistry*, 48(9):1383–1389.

- Lin, W.-Y. (2014). Association testing of clustered rare causal variants in case-control studies. *PLoS One*, 9:e94337.
- Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326.
- Lin, X., Cai, T., Wu, M., Zhou, Q., Liu, G., Christiani, D., and Lin, X. (2011). Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genetic Epidemiology*, 35(7):620–631.
- Lin, X., Lee, S., Christiani, D., and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, page kxt006.
- Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9(1):292.
- Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088.
- Liu, D. J., Peloso, G. M., Zhan, X., Holmen, O. L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P. L., Goel, A., Zhang, H., et al. (2014). Meta-analysis of gene-level tests for rare variant association. *Nature Genetics*, 46(2):200–204.
- Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance components analysis. *Nature Genetics*, 47(12):1385.
- Lysaker, P. H. and Whitney, K. A. (2009). Obsessive–compulsive symptoms in schizophrenia: prevalence, correlates and treatment. *Expert Review of Neurotherapeutics*, 9(1):99–107.
- Madsen, B. and Browning, S. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384.
- Maier, R., Moser, G., Chen, G.-B., Ripke, S., Coryell, W., Potash, J. B., Scheftner, W. A., Shi, J., Weissman, M. M., Hultman, C. M., et al. (2015). Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *The American Journal of Human Genetics*, 96(2):283–294.
- Maity, A., Sullivan, P., and Tzeng, J. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic Epidemiology*, 36(7):686–695.

- Makowsky, R., Pajewski, N. M., Klimentidis, Y. C., Vazquez, A. I., Duarte, C. W., Allison, D. B., and de Los Campos, G. (2011). Beyond missing heritability: prediction of complex traits. *PLoS Genetics*, 7(4):e1002051.
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.
- Marceau, R., Lu, W., Holloway, S., Sale, M. M., Worrall, B. B., Williams, S. R., Hsu, F.-C., and Tzeng, J.-Y. (2015). A fast multiple-kernel method with applications to detect gene-environment interaction. *Genetic Epidemiology*, 39(6):456–468.
- Marvel, S. W., Rotroff, D. M., Wagner, M. J., Buse, J. B., Havener, T. M., McLeod, H. L., and Motsinger-Reif, A. A. (2017). Common and rare genetic markers of lipid variation in subjects with type 2 diabetes from the ACCORD clinical trial. *PeerJ*, 5:e3187.
- Mellman, I., Lin, P.-F., Ruddle, F., and Rosenberg, L. (1979). Genetic control of cobalamin binding in normal and mutant cells: assignment of the gene for 5-methyltetrahydrofolate: L-homocysteine S-methyltransferase to human chromosome 1. *Proceedings of the National Academy of Sciences*, 76(1):405–409.
- Morris, A. P. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, 34(2):188–193.
- Morris, S. E. and Cuthbert, B. N. (2012). Research Domain Criteria: cognitive systems, neural circuits, and dimensions of behavior. *Dialogues in Clinical Neuroscience*, 14(1):29.
- Mukhopadhyaya, K., Krishnaiah, R., Taye, T., Nigam, A., Bailey, A., Sivakumaran, T., and Fineberg, N. (2009). Obsessive-compulsive disorder in UK clozapine-treated schizophrenia and schizoaffective disorder: a cause for clinical concern. *Journal of Psychopharmacology*, 23(1):6–13.
- Nygård, O., Vollset, S., Refsum, H., Stensvold, I., Tverdal, A., Nordrehaug, J., Ueland, P., and Kvåle, G. (1995). Total plasma homocysteine and cardiovascular risk profile: the Hordaland Homocysteine Study. *JAMA*, 274(19):1526–1533.
- O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, pages 1079–1087.
- O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M.-R., and Coin, L. J. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One*, 7(5):e34861.
- Pang, H., Kim, I., and Zhao, H. (2014). Random effects model for multiple pathway analysis with applications to type II diabetes microarray data. *Statistics in Biosciences*, pages 1–20.

- Pongpanich, M., Neely, M., and Tzeng, J. (2011). On the aggregation of multimarker information for marker-set and sequencing data analysis: genotype collapsing vs. similarity collapsing. *Frontiers in Genetics*, 2:110.
- Poyurovsky, M., Weizman, R., Weizman, A., and Koran, L. (2005). Memantine for treatment-resistant OCD. *The American Journal of Psychiatry*, 162(11):2191–2192.
- Poyurovsky, M., Zohar, J., Glick, I., Koran, L. M., Weizman, R., Tandon, R., and Weizman, A. (2012). Obsessive-compulsive symptoms in schizophrenia: implications for future psychiatric classifications. *Comprehensive Psychiatry*, 53(5):480–483.
- Preisig, M., Waeber, G., Vollenweider, P., Bovet, P., Rothen, S., Vandeleur, C., Guex, P., Middleton, L., Waterworth, D., Mooser, V., et al. (2009). The PsyCoLaus study: methodology and characteristics of the sample of a population-based survey on psychiatric disorders and their association with genetic and cardiovascular risk factors. *BMC Psychiatry*, 9(1):9.
- Price, A., Kryukov, G., de Bakker, P., Purcell, S., Staples, J., Wei, L.-J., and Sunyaev, S. (2010). Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838.
- Qiu, Y. and Mei, J. (2016). *rARPACK: Solvers for Large Scale Eigenvalue and SVD Problems*. R package version 0.11-0.
- Raychaudhuri, S., Korn, J., McCarroll, S., Consortium, I. S., Altshuler, D., Sklar, P., Purcell, S., Daly, M., et al. (2010). Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genetics*, 6(9):e1001097.
- Rosenblatt, D., Schmutz, S., Cooper, B., Zaleski, W., and Casey, R. (1985). Prenatal vitamin B12 therapy of a fetus with methylcobalamin deficiency (cobalamin E disease). *The Lancet*, 325(8438):1127–1129.
- Rubenstein, C. S., Pigott, T. A., L'Heureux, F., Hill, J. L., and Murphy, D. (1992). A preliminary investigation of the lifetime prevalence of anorexia and bulimia nervosa in patients with obsessive compulsive disorder. *The Journal of Clinical Psychiatry*.
- Ruscio, A., Stein, D., Chiu, W., and Kessler, R. (2010). The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. *Molecular Psychiatry*, 15(1):53–63.
- Sakoda, L. C., Jorgenson, E., and Witte, J. S. (2013). Turning of COGS moves forward findings for hormonally mediated cancers. *Nature Genetics*, 45(4):345–348.
- Samanta, U., Kirby, S. D., Srinivasan, P., Cerasoli, D. M., and Bahnson, B. J. (2009). Crystal structures of human group-VIIA phospholipase A2 inhibited by organophosphorus nerve agents exhibit non-aged complexes. *Biochemical Pharmacology*, 78(4):420–429.

- Sanislow, C. A., Pine, D. S., Quinn, K. J., Kozak, M. J., Garvey, M. A., Heinssen, R. K., Wang, P. S.-E., and Cuthbert, B. N. (2010). Developing constructs for psychopathology research: research domain criteria. *Journal of Abnormal Psychology*, 119(4):631–639.
- Schaffner, S., Foo, C., Gabriel, S., Reich, D., Daly, M., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, 15(11):1576–1583.
- Schaid, D. J., Sinnwell, J. P., McDonnell, S. K., and Thibodeau, S. N. (2013). Detecting genomic clustering of risk variants from sequence data: cases versus controls. *Human Genetics*, 132(11):1301–1309.
- Schiele, F., Park, J., Redemann, N., Luippold, G., and Nar, H. (2014). An antibody against the C-terminal domain of PCSK9 lowers LDL cholesterol levels in vivo. *Journal of Molecular Biology*, 426(4):843–852.
- Schirmbeck, F. and Zink, M. (2013). Comorbid obsessive-compulsive symptoms in schizophrenia: contributions of pharmacological and genetic factors. *Frontiers in Pharmacology*, 4.
- Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.
- Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.8.
- Schuh, S., Rosenblatt, D., Cooper, B., Schroeder, M.-L., Bishop, A., Seargeant, L., and Haworth, J. (1984). Homocystinuria and megaloblastic anemia responsive to vitamin B12 therapy: an inborn error of metabolism due to a defect in cobalamin metabolism. *The New England Journal of Medicine*, 310(11):686–690.
- Seeman, M. V. (2014). Eating disorders and psychosis: Seven hypotheses. *World Journal of Psychiatry*, 4(4):112.
- Seetharam, B., Bose, S., and Li, N. (1999). Cellular import of cobalamin (vitamin B-12). *The Journal of Nutrition*, 129(10):1761–1764.
- Song, K., Nelson, M., Aponte, J., Manas, E., Bacanu, S., Yuan, X., Kong, X., Cardon, L., Mooser, V., Whittaker, J., et al. (2012). Sequencing of Lp-PLA2-encoding PLA2G7 gene in 2000 europeans reveals several rare loss-of-function mutations. *The Pharmacogenomics Journal*, 12(5):425–431.
- Spence, J., Cordy, P., Kortas, C., and Freeman, D. (1999). Effect of usual doses of folate supplementation on elevated plasma homocyst(e)ine in hemodialysis patients: no difference between 1 and 5 mg daily. *American Journal of Nephrology*, 19(3):405–410.
- Stanislawski-Sachadyn, A., Woodside, J. V., Sayers, C., Yarnell, J., Young, I., Evans, A., Mitchell, L., and Whitehead, A. (2010). The transcobalamin (TCN2) 776C>G polymorphism affects homocysteine concentrations among subjects with low vitamin B(12) status. *European Journal of Clinical Nutrition*, 64(11):1338–1343.

- Stroes, E., Colquhoun, D., Sullivan, D., Civeira, F., Rosenson, R. S., Watts, G. F., Bruckert, E., Cho, L., Dent, R., Knusel, B., et al. (2014). Anti-PCSK9 antibody effectively lowers cholesterol in patients with statin intolerance: the GAUSS-2 randomized, placebo-controlled phase 3 clinical trial of evolocumab. *Journal of the American College of Cardiology*, 63(23):2541–2548.
- Swinbourne, J., Hunt, C., Abbott, M., Russell, J., St Clare, T., and Touyz, S. (2012). The comorbidity between eating disorders and anxiety disorders: Prevalence in an eating disorder sample and anxiety disorder sample. *Australian & New Zealand Journal of Psychiatry*, 46(2):118–131.
- Tango, T. (2010). *Statistical methods for disease clustering*. Springer Science & Business Media.
- Toole, J., Malinow, M. R., Chambless, L., Spence, J., Pettigrew, L., Howard, V., Sides, E., Wang, C.-H., and Stampfer, M. (2004). Lowering homocysteine in patients with ischemic stroke to prevent recurrent stroke, myocardial infarction, and death: the Vitamin Intervention for Stroke Prevention (VISP) randomized controlled trial. *JAMA*, 291(5):565–575.
- Tsai, M., Bignell, M., Yang, F., Welge, B., Graham, K., and Hanson, N. (2000). Polygenic influence on plasma homocysteine: association of two prevalent mutations, the 844ins68 of cystathionine β -synthase and A2756 G of methionine synthase, with lowered plasma homocysteine levels. *Atherosclerosis*, 149(1):131–137.
- Tzeng, J., Lu, W., and Hsu, F. (2014). Gene-level pharmacogenetic analysis on survival outcomes using gene-trait similarity regression. *The Annals of Applied Statistics*, 8(2):1232–1255.
- Tzeng, J., Zhang, D., Chang, S.-M., Thomas, D., and Davidian, M. (2009). Gene-trait similarity regression for multimarker-based association analysis. *Biometrics*, 65(3):822–832.
- Tzeng, J., Zhang, D., Pongpanich, M., Smith, C., McCarthy, M., Sale, M., Worrall, B., Hsu, F., Thomas, D., and Sullivan, P. (2011). Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *The American Journal of Human Genetics*, 89(2):277–288.
- Tzeng, J.-Y., Magnusson, P. K., Sullivan, P. F., Szatkiewicz, J. P., Consortium, S. S., et al. (2015). A new method for detecting associations with rare copy-number variants. *PLoS Genetics*, 11(10):e1005403.
- van der Griend, R., Biesma, D., and Banga, J.-D. (2002). Postmethionine-load homocysteine determination for the diagnosis hyperhomocysteinaemia and efficacy of homocysteine lowering treatment regimens. *Vascular Medicine*, 7(1):29–33.
- Van der Sluis, S., Posthuma, D., and Dolan, C. V. (2013). TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genetics*, 9(1):e1003235.
- Vattikuti, S., Guo, J., and Chow, C. C. (2012). Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genetics*, 8(3):e1002637.

- Voorman, A., Lumley, T., McKnight, B., and Rice, K. (2011). Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PLoS One*, 6(5):e19416.
- Wang, D., Song, L., Singh, V., Rao, S., An, L., and Madhavan, S. (2015a). SNP2Structure: a public and versatile resource for mapping and three-dimensional modeling of missense SNPs on human protein structures. *Computational and Structural Biotechnology Journal*, 13:514–519.
- Wang, J., Huff, A., Spence, J., and Hegele, R. (2004). Single nucleotide polymorphism in CTH associated with variation in plasma homocysteine concentration. *Clinical Genetics*, 65(6):483–486.
- Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164.
- Wang, X., Duarte, N., Cai, H., Adachi, T., Sim, A., Cranney, G., and Wilcken, D. (1999). Relationship between total plasma homocysteine, polymorphisms of homocysteine metabolism related enzymes, risk factors and coronary artery disease in the Australian hospital-based population. *Atherosclerosis*, 146(1):133–140.
- Wang, X., Epstein, M., and Tzeng, J. (2014). Analysis of gene-gene interactions using gene-trait similarity regression. *Human Heredity*, 78(1):17–26.
- Wang, Y., Liu, A., Mills, J. L., Boehnke, M., Wilson, A. F., Bailey-Wilson, J. E., Xiong, M., Wu, C. O., and Fan, R. (2015b). Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. *Genetic Epidemiology*, 39(4):259–275.
- Wang, Z., Maity, A., Hsiao, C. K., Voora, D., Kaddurah-Daouk, R., and Tzeng, J.-Y. (2015c). Module-based association analysis for omics data with network structure. *PLoS One*, 10(3):e0122309.
- Wang, Z., Maity, A., Luo, Y., Neely, M., and Tzeng, J. (2015d). Complete effect-profile assessment in association studies with multiple genetic and multiple environmental factors. *Genetic Epidemiology*, 39:122–133.
- Wei, C. and Lu, Q. (2015). A generalized similarity u test for multivariate analysis of sequencing data. *arXiv preprint arXiv:1505.01179*.
- Wei, L. and Johnson, W. E. (1985). Combining dependent tests with incomplete repeated measurements. *Biometrika*, 72(2):359–364.
- Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.
- Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., and Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14(7):507–515.

- Wu, M., Kraft, P., Epstein, M., Taylor, D., Chanock, S., Hunter, D., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942.
- Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93.
- Wu, M., Maity, A., Lee, S., Simmons, E., Harmon, Q., Lin, X., Engel, S., Mollidrem, J., and Armistead, P. (2013). Kernel machine SNP-set testing under multiple candidate kernels. *Genetic Epidemiology*, 37(3):267–275.
- Xu, C., Ladouceur, M., Dastani, Z., Richards, J. B., Ciampi, A., and Greenwood, C. M. (2012). Multiple regression methods show great potential for rare variant association tests. *PLoS One*, 7(8):e41694.
- Yang, Q. and Wang, Y. (2012). Methods for analyzing multivariate phenotypes in genetic association studies. *Journal of Probability and Statistics*, 2012:1–13.
- Yang, Q., Wu, H., Guo, C.-Y., and Fox, C. S. (2010). Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic Epidemiology*, 34(5):444–454.
- Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.
- Yue, P., Forrest, W. F., Kaminker, J. S., Lohr, S., Zhang, Z., and Cavet, G. (2010). Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Human Mutation*, 31(3):264–271.
- Yum, S. Y., Caracci, G., and Hwang, M. Y. (2009). Schizophrenia and eating disorders. *Psychiatric Clinics of North America*, 32(4):809–819.
- Zavadáková, P., Fowler, B., Suormala, T., Novotna, Z., Mueller, P., Hennermann, J., Zeman, J., Vilaseca, M., Vilarinho, L., Gutsche, S., et al. (2005). cblE type of homocystinuria due to methionine synthase reductase deficiency: functional correction by minigene expression. *Human Mutation*, 25(3):239–247.
- Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, 4(1):57–74.
- Zhao, G., Marceau, R., Zhang, D., and Tzeng, J.-Y. (2015). Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. *Genetics*, 199(3):695–710.

Zhou, H., Alexander, D. H., Sehl, M. E., Sinsheimer, J. S., Sobel, E., and Lange, K. (2011). Penalized regression for genome-wide association screening of sequence data. In *Pacific Symposium on Biocomputing*, page 106. NIH Public Access.

Zhou, H., Sehl, M. E., Sinsheimer, J. S., and Lange, K. (2010). Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 26(19):2375.

Zhou, X. and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11(4):407–409.

APPENDICES

Appendix A

Additional Information for Cross Disorder Kernel Machine Simulation

A.0.1 Case Control Sampling of Genotype Matrix for Binary Traits

Below we discuss a case control framework for binary trait simulations for multiple traits that enables true controls (i.e., individuals who are cases for all of the considered traits). We note the model fitting would be the same as for quantitative traits using the generalized model framework.

Given the randomly sampled genotype matrix G^* , we consider a case control sampling framework to generate simulated genotype and phenotype for all three disorders, giving us $CS_d = 1000$ cases for each disorder $d = 1, 2, 3$ and $CN = 3000$ “true controls,” defined as those which are controls for all three disorders simultaneously – 1000 per disorder. Here causal variants are determined in the same manner as for continuous phenotype simulations.

1. Sample one individual (row) from G^* , which we denote as G_i^* .
2. For disorder $d = 1, 2, 3$ do:
 - (a) If number of accumulated sampled cases for disorder d is less than the desired number

of cases, or if the number of true controls is less than the desired number of controls:

- i. Generate probability of case for individual i , disorder d as: $p_{i,d} = \frac{\exp(\beta_0 + X\beta_X + G_i^*\beta_d)}{1 + \exp(\beta_0 + X\beta_X + G_i^*\beta_d)}$
 - ii. Generate phenotype for individual i , disorder d as: $y_{i,d} \sim \text{Bin}(1, p_{i,d})$.
 - iii. If $y_{i,d} = 1$, save individual i as a case for disorder d , and sample the next individual.
Otherwise, continue.
3. If $y_{i,d} = 0 \forall d$, save individual i as a true control.
 4. Continue until all cases and controls are determined.

A.0.2 Cross Disorder and Single Disorder Tuning Parameter Summaries

Table A.1 Average optimal tuning parameter (and corresponding standard deviation) for the cross disorder and single disorder continuous trait kernel machine models over 100 simulations

γ_G	% causals shared	% variance explained	Lambda			
			cross disorder	disorder 1	disorder 2	disorder 3
1	40	66.8	0.03 (6×10^{-4})	0.06 (0.01)	0.06 (0.03)	0.07 (0.003)
	60	69.4	0.03 (0.002)	0.06 (0.02)	0.07 (0.02)	0.08 (0.004)
2	40	88.8	0.03 (4×10^{-4})	0.09 (0.02)	0.1 (0.06)	0.12 (0.005)
	60	90.1	0.04 (5×10^{-4})	0.09 (0.03)	0.11 (0.04)	0.13 (0.007)

Appendix B

Local Score Test Limiting Distribution

Recall the local kernel score test statistic for a test of no effect of variant j for fixed c is

$$T_{j,c} = \frac{1}{n}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)^T K_{j,c}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n) = \frac{1}{n}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n) Z_{j,c} Z_{j,c}^T (\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)^T$$

or

$$T_{j,c} = \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \hat{\epsilon}_i \right]^T \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \hat{\epsilon}_i \right]$$

where $\hat{\epsilon}_i = Y_i - g^{-1}(X\hat{\beta})$ is the fitted residual from the KM model under the null hypothesis of no genetic effect.

To derive the limiting distribution of the local kernel test statistic, we perform a Taylor expansion around the residual under the null hypothesis, $\epsilon_i = y_i - g(x_i^T \beta_*)$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \hat{\epsilon}_i = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \epsilon_i + \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \frac{\partial \epsilon_i}{\partial \beta^T} (\hat{\beta} - \beta_*) + o_p(1) \quad (\text{B.1})$$

where

$$\frac{\partial \epsilon_i}{\partial \beta^T} = \begin{cases} -x_i^T, & \text{for quantitative traits} \\ -x_i^T \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} & \text{for binary traits} \end{cases}$$

Further, we know the score function $\frac{\partial \ell(\beta, \sigma)}{\partial \beta} \Big|_{\beta=\hat{\beta}} = 0$ at the maximum likelihood estimator $\hat{\beta}$.

For quantitative traits,

$$\begin{aligned} \frac{\partial \ell(\beta, \sigma)}{\partial \beta} \Big|_{\beta=\hat{\beta}} &= \frac{\partial}{\partial \beta} \left[-\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i^T \beta)^2 \right] \Big|_{\beta=\hat{\beta}} = \left[\frac{1}{\sigma^2} \sum_{i=1}^n x_i (y_i - x_i^T \beta) \right] \Big|_{\beta=\hat{\beta}} = 0 \\ &\Rightarrow \sum_{i=1}^n x_i (y_i - x_i^T \hat{\beta}) = 0 \end{aligned}$$

Performing a Taylor expansion about the true regression coefficient β gives

$$\begin{aligned} 0 &= \sum_{i=1}^n x_i (y_i - x_i^T \beta_*) - \sum_{i=1}^n x_i x_i^T (\hat{\beta} - \beta_*) + o_p(1) \\ &\Rightarrow (\hat{\beta} - \beta_*) = \left(\sum_{i=1}^n x_i x_i^T \right)^{-1} \sum_{i=1}^n x_i (y_i - x_i^T \beta_*) + o_p(1) \\ &\Rightarrow \sqrt{n}(\hat{\beta} - \beta_*) = \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^T \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i (y_i - x_i^T \beta_*) + \sqrt{n} o_p(1) \\ &\Rightarrow \sqrt{n}(\hat{\beta} - \beta_*) \xrightarrow{p} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^T \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i (y_i - x_i^T \beta_*) \end{aligned} \tag{B.2}$$

Plugging B.2 into equation B.1,

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \hat{\epsilon}_i &= \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \epsilon_i - \frac{1}{n} \sum_{i=1}^n Z_{j,c}^{(i)} x_i^T \sqrt{n}(\hat{\beta} - \beta_*) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \epsilon_i - \frac{1}{n} \sum_{i=1}^n Z_{j,c}^{(i)} x_i^T \left[\left(\frac{1}{n} \sum_{i=1}^n x_i x_i^T \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i (y_i - x_i^T \beta_*) \right] + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \epsilon_i - \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} x_i^T \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^T \right)^{-1} x_i \epsilon_i + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ Z_{j,c}^{(i)} - \left(Z_{j,c}^{(i)} x_i^T \right) \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^T \right)^{-1} x_i \right\} \epsilon_i + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i + o_p(1) \tag{B.3}
\end{aligned}$$

Here $\psi_i = \{Z_{j,c}^{(i)} - A_2 A_1^{-1} x_i\} \epsilon_i$ is an influence function, estimated by $\hat{\psi}_i = \{Z_{j,c}^{(i)} - \hat{A}_2 \hat{A}_1^{-1} x_i\} \epsilon_i$,

where

$$\begin{aligned}
A_1 &= E[x_i x_i^T] & \hat{A}_1 &= \frac{1}{n} \sum_{i=1}^n x_i x_i^T \\
A_2 &= E[Z_{j,c}^{(i)} x_i^T] & \hat{A}_2 &= \frac{1}{n} \sum_{i=1}^n Z_{j,c}^{(i)} x_i^T
\end{aligned}$$

Similarly, for binary traits,

$$\begin{aligned}
\frac{\partial \ell(\beta, \sigma)}{\partial \beta} \Big|_{\beta=\hat{\beta}} &= \frac{\partial}{\partial \beta} \sum_{i=1}^n [y_i \beta^T x_i - \log(1 + e^{\beta^T x_i})] \Big|_{\beta=\hat{\beta}} = \sum_{i=1}^n x_i \left(y_i - \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \right) \Big|_{\beta=\hat{\beta}} = 0 \\
&\Rightarrow \sum_{i=1}^n x_i \left(y_i - \frac{e^{\hat{\beta}^T x_i}}{1 + e^{\hat{\beta}^T x_i}} \right) = 0
\end{aligned}$$

A Taylor expansion of the score function about the true regression coefficient β gives

$$\begin{aligned}
0 &= \sum_{i=1}^n x_i \left(y_i - \frac{e^{\beta_*^T x_i}}{1 + e^{\beta_*^T x_i}} \right) - \sum_{i=1}^n \left[\frac{e^{\beta_*^T x_i}}{(1 + e^{\beta_*^T x_i})^2} x_i x_i^T \right] (\hat{\beta} - \beta_*) + o_p(1) \\
\Rightarrow \sqrt{n}(\hat{\beta} - \beta_*) &= \left[\sum_{i=1}^n \frac{e^{\beta_*^T x_i}}{(1 + e^{\beta_*^T x_i})^2} x_i x_i^T \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \left(y_i - \frac{e^{\beta_*^T x_i}}{1 + e^{\beta_*^T x_i}} \right) + \sqrt{n} o_p(1) \\
\Rightarrow \sqrt{n}(\hat{\beta} - \beta_*) &\xrightarrow{p} \left[\sum_{i=1}^n \frac{e^{\beta_*^T x_i}}{(1 + e^{\beta_*^T x_i})^2} x_i x_i^T \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \left(y_i - \frac{e^{\beta_*^T x_i}}{1 + e^{\beta_*^T x_i}} \right) \quad (\text{B.4})
\end{aligned}$$

Plugging B.4 into equation B.1,

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \hat{\epsilon}_i &= \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \epsilon_i - \frac{1}{n} \sum_{i=1}^n Z_{j,c}^{(i)} \left[\frac{e^{\beta_*^T x_i}}{1 + e^{\beta_*^T x_i}} x_i^T \right] \sqrt{n}(\hat{\beta} - \beta_*) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \epsilon_i - \frac{1}{n} \sum_{i=1}^n Z_{j,c}^{(i)} \frac{e^{\beta_*^T x_i}}{1 + e^{\beta_*^T x_i}} x_i^T \\
&\quad \times \left[\left(\sum_{i=1}^n \frac{e^{\beta_*^T x_i} x_i x_i^T}{(1 + e^{\beta_*^T x_i})^2} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \left(y_i - \frac{e^{\beta_*^T x_i}}{1 + e^{\beta_*^T x_i}} \right) \right] + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \epsilon_i - \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \frac{e^{\beta_*^T x_i}}{1 + e^{\beta_*^T x_i}} x_i^T \left(\sum_{i=1}^n \frac{e^{\beta_*^T x_i}}{(1 + e^{\beta_*^T x_i})^2} x_i x_i^T \right)^{-1} x_i \epsilon_i + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ Z_{j,c}^{(i)} - \left(Z_{j,c}^{(i)} \frac{e^{\beta_*^T x_i}}{1 + e^{\beta_*^T x_i}} x_i^T \right) \left(\sum_{i=1}^n \frac{e^{\beta_*^T x_i}}{(1 + e^{\beta_*^T x_i})^2} x_i x_i^T \right)^{-1} x_i \right\} \epsilon_i + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i + o_p(1) \quad (\text{B.5})
\end{aligned}$$

Again $\psi_i = \{Z_{j,c}^{(i)} - A_2 A_1^{-1} x_i\} \epsilon_i$ is an influence function, estimated by $\hat{\psi}_i = \{Z_{j,c}^{(i)} - \hat{A}_2 \hat{A}_1^{-1} x_i\} \epsilon_i$

wherer

$$\begin{aligned}
A_1 &= E \left[\frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} x_i x_i^T \right] & \hat{A}_1 &= \frac{1}{n} \sum_{i=1}^n \frac{e^{\hat{\beta}^T x_i}}{1 + e^{\hat{\beta}^T x_i}} x_i x_i^T \\
A_2 &= E \left[Z_{j,c}^{(i)} \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} x_i^T \right] & \hat{A}_2 &= \frac{1}{n} \sum_{i=1}^n Z_{j,c}^{(i)} \frac{e^{\hat{\beta}^T x_i}}{1 + e^{\hat{\beta}^T x_i}} x_i^T
\end{aligned}$$

By the central limit theorem, $\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i \xrightarrow{d} N(0, \Sigma)$. Then by Slutsky's theorem, $\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \hat{\epsilon}_i \xrightarrow{d} N(0, \Sigma)$, where

$$\Sigma = E(\psi_i \psi_i^T) = E \left[\left\{ Z_{j,c}^{(i)} - A_2 A_1^{-1} x_i \right\}^{\otimes 2} \epsilon_i^2 \right] \quad (\text{B.6})$$

estimated by

$$\hat{\Sigma}_n = \frac{1}{n} \sum_i^n \left\{ Z_{j,c}^{(i)} - \hat{A}_2 \hat{A}_1^{-1} x_i \right\}^{\otimes 2} \hat{\epsilon}_i^2 \quad (\text{B.7})$$

where we define $a^{\otimes 2} = a a^T$.

Thus,

$$T_{j,c} = \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \hat{\epsilon}_i \right]^T \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{j,c}^{(i)} \hat{\epsilon}_i \right] \xrightarrow{d} \sum_d \xi_d \chi_{1,d}^2$$

where ξ_d are the nonzero eigenvalues of Σ . This can be estimated by $\sum_d \hat{\xi}_d \chi_{1,d}^2$ where $\hat{\xi}_d$ are the nonzero eigenvalues of $\hat{\Sigma}_n$.

Appendix C

Additional Local Kernel Simulation

Results

C.0.1 Additional Simulation Results: Quantitative Traits Local Burden Kernel Test

Table C.1 Type I error and power for continuous trait local burden kernel test, weighted by minor allele frequency, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
(69,82)	0.35	0.736	0.103	0.972	0.111
	1.05	1.000	0.147	1.000	0.200
(303,326,331)	0.35	0.802	0.052	0.930	0.052
	0.55	0.987	0.059	1.000	0.067
(191,200)	0.8	0.996	0.050	1.000	0.051
	1.5	1.000	0.060	1.000	0.064
(110,273)	0.45	0.910	0.049	0.977	0.049
	0.6	0.993	0.051	0.998	0.051
(326,331)	0.35	0.778	0.046	0.903	0.046
	0.55	0.990	0.049	1.000	0.049
(303,326)	0.35	0.870	0.051	0.879	0.053
	0.55	0.990	0.055	0.990	0.067
(69,82) pos, (303,326,331) neg	0.35	0.823	0.135	0.970	0.156
	0.55	0.990	0.164	1.000	0.212

Table C.2 Type I error and power for continuous trait local burden kernel test, unweighted, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
type I error	0	0.000	0.044	0.000	0.043
(69,82)	0.35	0.736	0.103	0.971	0.113
	1.05	1.000	0.147	1.000	0.211
(303,326,331)	0.35	0.802	0.052	0.943	0.052
	0.55	0.987	0.059	1.000	0.067
(191,200)	0.8	0.996	0.050	1.000	0.050
	1.5	1.000	0.060	1.000	0.065
(110,273)	0.45	0.910	0.049	0.977	0.048
	0.6	0.993	0.051	0.998	0.051
(326,331)	0.35	0.778	0.046	0.898	0.045
	0.55	0.990	0.049	0.999	0.049
(303,326)	0.35	0.870	0.051	0.890	0.057
	0.55	0.990	0.055	0.994	0.078
(69,82) pos,	0.35	0.823	0.135	0.979	0.164
(303,326,331) neg	0.55	0.990	0.164	1.000	0.226

Table C.3 Type I error and power for continuous trait local burden kernel test, unweighted, with $c = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$ over various causal variant scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
type I error	0	0.000	0.044	0.000	0.042
(69,82)	0.35	0.736	0.103	0.971	0.113
	1.05	1.000	0.147	1.000	0.211
(303,326,331)	0.35	0.802	0.052	0.943	0.053
	0.55	0.987	0.059	1.000	0.068
(191,200)	0.8	0.996	0.050	1.000	0.050
	1.5	1.000	0.060	1.000	0.064
(110,273)	0.45	0.910	0.049	0.977	0.048
	0.6	0.993	0.051	0.998	0.052
(326,331)	0.35	0.778	0.046	0.899	0.045
	0.55	0.990	0.049	0.999	0.049
(303,326)	0.35	0.870	0.051	0.890	0.057
	0.55	0.990	0.055	0.994	0.079
(69,82) pos,	0.35	0.823	0.135	0.979	0.164
(303,326,331) neg	0.55	0.990	0.164	1.000	0.226

C.0.2 Additional Simulation Results: Binary Traits Local Burden Kernel Test

Table C.4 Type I error and power for binary trait local burden test, weighted by minor allele frequency, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
(69,82)	0.7	0.705	0.111	0.965	0.114
	0.85	0.821	0.126	0.998	0.131
(303,326,331)	0.6	0.725	0.050	0.828	0.050
	0.75	0.855	0.050	0.937	0.051
(191,200)	1.4	0.973	0.045	0.993	0.043
	1.75	0.993	0.045	0.998	0.043
(110,273)	0.85	0.858	0.047	0.934	0.047
	1.1	0.961	0.049	0.987	0.048
(326,331)	0.9	0.923	0.047	0.972	0.045
	1.1	0.980	0.048	0.993	0.047
(303,326)	0.6	0.779	0.048	0.774	0.050
	0.75	0.894	0.049	0.887	0.054
(69,82) pos,	0.6	0.732	0.126	0.892	0.135
(303,326,331) neg	0.75	0.862	0.144	0.969	0.158

Table C.5 Type I error and power for binary trait local burden kernel test, unweighted, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
type I error	0	0.000	0.045	0.000	0.042
(69,82)	0.7	0.705	0.111	0.964	0.115
	0.85	0.821	0.126	0.998	0.131
(303,326,331)	0.6	0.725	0.050	0.847	0.050
	0.75	0.855	0.050	0.953	0.052
(191,200)	1.4	0.973	0.045	0.993	0.042
	1.75	0.993	0.045	0.998	0.043
(110,273)	0.85	0.858	0.047	0.933	0.046
	1.1	0.961	0.049	0.989	0.047
(326,331)	0.9	0.923	0.047	0.970	0.046
	1.1	0.980	0.048	0.992	0.047
(303,326)	0.6	0.779	0.048	0.791	0.052
	0.75	0.894	0.049	0.908	0.058
(69,82) pos, (303,326,331) neg	0.6	0.732	0.126	0.903	0.139
	0.75	0.862	0.144	0.976	0.164

Table C.6 Type I error and power for binary trait local burden kernel test, unweighted, with $c = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$ over various causal variant scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
type I error	0	0.000	0.045	0.000	0.043
(69,82)	0.7	0.705	0.111	0.964	0.116
	0.85	0.821	0.126	0.998	0.131
(303,326,331)	0.6	0.725	0.050	0.847	0.050
	0.75	0.855	0.050	0.953	0.053
(191,200)	1.4	0.973	0.045	0.993	0.042
	1.75	0.993	0.045	0.998	0.043
(110,273)	0.85	0.858	0.047	0.933	0.047
	1.1	0.961	0.049	0.989	0.047
(326,331)	0.9	0.923	0.047	0.970	0.046
	1.1	0.980	0.048	0.992	0.047
(303,326)	0.6	0.779	0.048	0.791	0.052
	0.75	0.894	0.049	0.907	0.058
(69,82) pos, (303,326,331) neg	0.6	0.732	0.126	0.904	0.140
	0.75	0.862	0.144	0.976	0.164

C.0.3 Additional Simulation Results: Quantitative Traits Local Linear Kernel Test

Table C.7 Type I error and power for continuous trait local linear kernel test, weighted by minor allele frequency, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
(69,82)	0.35	0.740	0.107	0.951	0.149
	1.05	1.000	0.150	1.000	0.334
(303,326,331)	0.35	0.813	0.055	0.945	0.072
	0.55	0.988	0.062	1.000	0.172
(191,200)	0.8	0.995	0.053	1.000	0.067
	1.5	1.000	0.062	1.000	0.146
(110,273)	0.45	0.914	0.051	0.977	0.056
	0.6	0.994	0.055	0.999	0.069
(326,331)	0.35	0.785	0.049	0.891	0.051
	0.55	0.990	0.053	0.999	0.071
(303,326)	0.35	0.876	0.052	0.915	0.111
	0.55	0.990	0.057	0.999	0.205
(69,82) pos, (303,326,331) neg	0.35	0.828	0.137	0.971	0.260
	0.55	0.992	0.164	1.000	0.493

Table C.8 Type I error and power for continuous trait local linear kernel test, unweighted, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
type I error	0	0.000	0.047	0.000	0.045
(69,82)	0.35	0.742	0.107	0.952	0.157
	1.05	1.000	0.150	1.000	0.345
(303,326,331)	0.35	0.812	0.054	0.960	0.101
	0.55	0.988	0.061	1.000	0.245
(191,200)	0.8	0.997	0.053	1.000	0.065
	1.5	1.000	0.062	1.000	0.147
(110,273)	0.45	0.916	0.052	0.978	0.055
	0.6	0.993	0.054	0.999	0.071
(326,331)	0.35	0.789	0.049	0.884	0.051
	0.55	0.991	0.053	0.999	0.067
(303,326)	0.35	0.878	0.052	0.942	0.155
	0.55	0.991	0.057	1.000	0.291
(69,82) pos,	0.35	0.828	0.138	0.980	0.316
(303,326,331) neg	0.55	0.992	0.164	1.000	0.582

Table C.9 Type I error and power for continuous trait local linear kernel test, unweighted, with $c = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$ over various causal variant scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
type I error	0	0.000	0.046	0.000	0.045
(69,82)	0.35	0.741	0.107	0.951	0.158
	1.05	1.000	0.150	1.000	0.343
(303,326,331)	0.35	0.810	0.054	0.961	0.099
	0.55	0.989	0.061	1.000	0.243
(191,200)	0.8	0.996	0.052	1.000	0.065
	1.5	1.000	0.063	1.000	0.145
(110,273)	0.45	0.915	0.052	0.975	0.055
	0.6	0.993	0.054	0.999	0.069
(326,331)	0.35	0.785	0.050	0.884	0.050
	0.55	0.990	0.052	0.999	0.068
(303,326)	0.35	0.876	0.051	0.938	0.154
	0.55	0.988	0.057	1.000	0.289
(69,82) pos,	0.35	0.829	0.137	0.978	0.313
(303,326,331) neg	0.55	0.992	0.164	1.000	0.580

C.0.4 Additional Simulation Results: Binary Traits Local Linear Kernel Test

Table C.10 Type I error and power for binary trait local linear kernel test, weighted by minor allele frequency, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
(69,82)	0.7	0.713	0.115	0.951	0.149
	0.85	0.823	0.128	0.991	0.180
(303,326,331)	0.6	0.732	0.052	0.841	0.060
	0.75	0.861	0.052	0.956	0.067
(191,200)	1.4	0.973	0.047	0.993	0.046
	1.75	0.993	0.047	0.998	0.046
(110,273)	0.85	0.868	0.050	0.931	0.052
	1.1	0.965	0.052	0.991	0.053
(326,331)	0.9	0.926	0.050	0.969	0.055
	1.1	0.980	0.051	0.995	0.057
(303,326)	0.6	0.788	0.049	0.817	0.078
	0.75	0.898	0.050	0.933	0.108
(69,82) pos, (303,326,331) neg	0.6	0.740	0.128	0.893	0.179
	0.75	0.867	0.147	0.977	0.237

Table C.11 Type I error and power for binary trait local linear kernel test, unweighted, with $c = (0, 0.25, 0.5)$ over various causal variant scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
type I error	0	0.000	0.047	0.000	0.047
(69,82)	0.7	0.713	0.115	0.949	0.157
	0.85	0.823	0.128	0.991	0.188
(303,326,331)	0.6	0.732	0.052	0.887	0.072
	0.75	0.859	0.052	0.977	0.095
(191,200)	1.4	0.973	0.047	0.993	0.047
	1.75	0.993	0.047	0.998	0.047
(110,273)	0.85	0.868	0.051	0.934	0.051
	1.1	0.964	0.052	0.991	0.052
(326,331)	0.9	0.926	0.050	0.968	0.055
	1.1	0.980	0.050	0.994	0.056
(303,326)	0.6	0.787	0.050	0.855	0.115
	0.75	0.900	0.050	0.959	0.156
(69,82) pos, (303,326,331) neg	0.6	0.741	0.128	0.915	0.205
	0.75	0.868	0.146	0.988	0.288

Table C.12 Type I error and power for binary trait local linear kernel test, unweighted, with $c = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$ over various causal variant scenarios using *PLA2G7* protein tertiary structure

causal variants	b	single SNP		local kernel	
		causal	noncausal	causal	noncausal
type I error	0	0.000	0.047	0.000	0.046
(69,82)	0.7	0.712	0.113	0.949	0.157
	0.85	0.824	0.128	0.991	0.189
(303,326,331)	0.6	0.735	0.051	0.886	0.071
	0.75	0.860	0.052	0.979	0.092
(191,200)	1.4	0.973	0.047	0.994	0.046
	1.75	0.993	0.047	0.998	0.047
(110,273)	0.85	0.863	0.050	0.934	0.052
	1.1	0.964	0.051	0.991	0.053
(326,331)	0.9	0.927	0.049	0.966	0.055
	1.1	0.980	0.050	0.993	0.056
(303,326)	0.6	0.785	0.050	0.852	0.115
	0.75	0.899	0.050	0.959	0.154
(69,82) pos, (303,326,331) neg	0.6	0.741	0.128	0.914	0.203
	0.75	0.867	0.147	0.987	0.286

Appendix D

ACCORD Rare Variants Summary

Table D.1 PCSK9 rare variants summary using PDB entry 4K8R

AA coord	SNP ID	imputed	3D coordinate			chain
			x1	x2	x3	
93	rs151193009	No	57.346	-46.239	14.168	A
96	rs185392267	Yes	55.728	-42.242	11.566	A
157	rs143117125	No	38.018	-76.553	-23.43	B
252	rs149139428	No	47.715	-55.573	-7.178	B
253	rs72646508	No	40.829	-54.25	-10.678	B
279	rs72646509	No	40.775	-73.533	10.343	B
283	rs72646510	Yes	43.215	-72.831	-0.779	B
391	rs146471967	No	41.198	-73.649	-16.112	B
417	rs143275858	No	53.053	-84.877	-15.42	B
425	rs28362261	No	53.144	-76.191	-32.719	B
443	rs28362263	Yes	44.693	-81.951	-18.093	B
466	rs72646517	Yes	74.752	-63.34	-21.448	B
469	rs141502002	No	83.385	-63.452	-20.617	B
498	rs145468572	No	78.379	-77.199	-15.736	B
525	rs140286279	Yes	66.35	-81.468	-12.955	B
553	rs28362270	Yes	80.897	-94.859	-2.167	B
554	rs149311926	Yes	78.241	-94.757	0.603	B
619	rs28362277	Yes	71.961	-85.137	-38.086	B
659	rs147182054	Yes	75.068	-79.678	-34.965	B