ABSTRACT

MARVEL, SKYLAR WARD. Exome-Wide Association Study of the ACCORD Clinical Trial. (Under the direction of Alison Motsinger-Reif.)

The Action to Control Cardiovascular Risk in Diabetes (ACCORD) clinical trial was designed to be the first definitive clinical trial on the possible benefits of targeting near-normal glycemic control, blood pressure and dyslipidemia management in patients with type 2 diabetes and whether the benefits will exceed the potential risks. The results were disappointing, with no difference found overall between standard and intensive treatments for controlling blood pressure and blood lipids and an increase in mortality when intensively controlling blood glucose. Observations of heterogeneity of response encouraged the investigation of genetic effects that could identify subgroups of patients who do significantly benefit from or are harmed by more intensive therapeutic regimens.

The research presented in this document focuses on analysis of the ACCORD genotype dataset. The first contribution of this work is creating a clean genotype dataset by implementing an extensive quality control pipeline applied to new exome chip technology. This was accomplished by using pre- and post-genotype calling metrics to identify poor quality plates, samples and probes. The second contribution is developing an analysis plan to streamline association studies. This was accomplished by developing a highly-automated pipeline that accounts for standard issues in association studies such as population structure, cryptic relatedness, covariate selection and multiple-hypothesis testing. The third contribution is a novel method that was developed to detect gene-gene (epistatic) interactions. This method uses grammatical evolution to select features and classifier architecture to construct a support vector machine classifier.

The ACCORD clinical trial produced a rich data source that can be mined for many interesting phenotype association studies. A clean genotype dataset is crucial for reliable association results. The highly-automated analysis pipeline will greatly reduce the amount of time required to test new phenotypes as they emerge from the clinical data. Methods that can detect gene-gene interactions will be useful in finding heritability that has been missing from common variant analyses.

Exome-Wide Association Study of the ACCORD Clinical Trial

by
Skylar Ward Marvel

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

Bioinformatics

Raleigh, North Carolina

2014

APPROVED BY:

Dahlia Nielsen	Jung-Ying Tzeng
Jorge Piedrahita	Alison Motsinger-Reif Chair of Advisory Committee

BIOGRAPHY

I used to be funny and happy, then I went to grad school. But now that's done... for the third time.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Alison Motsinger-Reif for picking me up as I was trying to transfer into the department and encouraging my accelerated graduation plan. I'd also like to thank Dr. Zhao-Bang Zeng for allowing me to transfer and the financial support from the National Institute of Environmental Health Sciences (NIEHS) training grant 5T32ES007329-12. I also thank my committee for spending the time to listen to my presentations and giving very appreciated feedback.

I thank my parents for always telling me what movies they just saw or what trip they just came back from while I'm stuck writing this thing until 4 in the morning. I don't thank my brother this time, take that! I thank my girlfriend for making sure I don't die when I forget to drink water all day.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter

1

Introduction

1.1 Overview

The Action to Control Cardiovascular Risk in Diabetes (ACCORD) clinical trial was designed to be the first definitive clinical trial on the possible benefits of targeting near-normal glycemic control, blood pressure and managing dyslipidemia in patients with type 2 diabetes and whether the benefits will exceed the potential risks [Buse, 2007; Goff Jr. et al., 2007]. The clinical trial was a large investment, spanning 8 years and being conducted at 77 clinical centers across the United States and Canada involving 10,251 middle-aged and older participants. The results were disappointing, with no difference found overall between standard and intensive treatments for controlling blood pressure and blood lipids and an increase in mortality when intensively controlling blood glucose [Cushman et al., 2010; Gerstein et al., 2008; Ginsberg et al., 2010].

Although the results of the trial were negative overall, there was evidence of differential response to treatment, e.g. an interaction between treatment and gender and heterogeneity of response in certain patients administered fenofibrate. These observations encouraged the investigation of genetic effects that could identify subgroups of patients who do significantly benefit from or are harmed by more intensive therapeutic regimens.

The research described in this document focuses on analysis of the ACCORD genotype dataset. This work focuses on three main areas: creating a clean genotype dataset using an extensive quality control pipeline, developing an analysis plan to streamline association studies, and methods development to detect gene-gene interactions. Best practices guidelines were followed to produce high-quality pipelines, however, this was not accomplished by simply following instructions or processing data using existing software. The vast majority of all steps were implemented in custom Perl, Bash, and R scripts. The scripts were written to allow a large amount of automation and easy adaptation for future datasets. Creating the QC and analysis pipelines has set the stage for rapid analysis of interesting ACCORD phenotypes, which will lead to multiple publications built on this foundation.

1.2 Contributions of This Work

Contribution 1

This research provides a new genotype dataset consisting of middle-aged or older men and women with type 2 diabetes who are at an increased risk for cardiovascular disease. This genetic data is accompanied by a large set of clinical measurements and can be used for many genetic association studies in the future. An extensive quality control pipeline was developed to remove any poor samples and genotyping probes. The genotype data was pre-phased to accelerate any future data imputations. In addition, functional variants were grouped into gene sets for use in rare variant analysis.

Contribution 2

This research provides an analysis pipeline that can be utilized to accelerate future analyses using the ACCORD dataset. This pipeline automates many of the steps used in a typical genetic association study, including covariate selection, removal of cryptically related individuals and adjusting for population structure using principal components. The ACCORD trial is a rich data source for future analyses and having an established analysis plan will be a great help as new phenotypes are investigated.

Contribution 3

This research presents a new analysis method for detecting gene-gene (epistatic) interactions. The difficulty in detecting epistatic effects is the number of genetic variant combinations. This new method uses an evolutionary algorithm (grammatical evolution) to explore the search space.

1.3 Organization

Chapter 2 provides an overview of the ACCORD clinical trial. The clinical trial was designed to test if intensive treatments that target normal levels of blood glucose, blood pressure and blood lipids reduce the risk of cardiovascular disease (CVD) in type 2 diabetic patients at high risk for CVD. Overall,

no difference was found between standard and intensive treatments for controlling blood pressure and blood lipids and there was an increase in mortality when intensively controlling blood glucose.

Chapter 3 describes the quality control (QC) pipeline that was applied to the raw ACCORD genotype data. QC metrics were used to identify and/or filter out problematic plates of samples, individual samples on plates and genotyping probes. Those samples and probes that passed the QC filters were then used to impute a large number of variants from a 1000 Genomes reference panel. Genotyped and imputed variants that were annotated as functional for a specific gene were then grouped together to create a rare variant analysis dataset.

Chapter 4 describes the analysis pipeline that was established to accelerate future analyses using the ACCORD dataset. The pipeline was written to automate much of the standard procedures for common and rare variant analyses. The pipeline includes steps that extract phenotype data from the clinical trial records, detect and remove cryptically related individuals, compute principal components, adjust p-values for genomic inflation, conduct common and rare variant analyses, and produce a variety of diagnosis and result plots. The pipeline was tested by analyzing the baseline lipid levels of all ACCORD samples after the quality control process. Significant variants and genes were compared to those previously reported in the literature.

Chapter 5 covers the development of the grammatical evolution support vector machine (GESVM) algorithm. The ultimate goal of GESVM is to detect epistatic effects in large datasets where brute force methods are infeasible. The approach of GESVM is to use an evolutionary algorithm (GE) to explore the search space of features and architectures that define the supervised learning model (SVM).

Chapter 6 summarizes the work presented in this document and presents ideas for future research.

1.4 Computing Environment

Most of the software and scripts were run on the Bioinformatics Research Core (BRC) server. Others were run on the High Performance Computing (HPC) cluster when indicated. The HPC uses CentOS

release 5.4 and has 1149 dual socket servers with Intel Xeon processors, comprised of a mixture of single-, dual-, quad-, six- and eight-cores, with 2-4 GB per core distributed memory. The BRC uses Ubuntu 12.04.4 and has 17 compute nodes. Each compute node has 2 Xeon E5 2670 “Sandy Bridge” processors with 8 cores, 128 GB of 1600 MHz RAM and two 1 TB 7.2k RPM near-line SAS 6 Gbps hard drives.

Chapter

2

ACCORD Clinical Trial

This chapter provides an overview of the Action to Control Cardiovascular Risk in Diabetes (ACCORD) clinical trial. The clinical trial was designed to test if intensive treatments that target normal levels of blood glucose, blood pressure and blood lipids reduce the risk of cardiovascular disease (CVD) in type 2 diabetic patients at high risk for CVD. Overall, no difference was found between standard and intensive treatments for controlling blood pressure and blood lipids and there was an increase in mortality when intensively controlling blood glucose.

2.1 Rationale, Design and Methods

The original purpose of the Action to Control Cardiovascular Risk in Diabetes (ACCORD) clinical trial is discussed in detail in [Buse, 2007] and [Goff Jr. et al., 2007]. Briefly, patients with type 2 diabetes have an increased risk of cardiovascular disease (CVD) that scales with increases in glycosylated hemoglobin (HbA_{1c}), blood pressure, low-density lipoprotein cholesterol (LDL), and triglycerides and with a decrease in high-density lipoprotein cholesterol (HDL). The prevalence of diabetes is drastically increasing, leading to increases in death, disability, and healthcare costs [Buse, 2007]. The overall goal of the ACCORD clinical trial was to determine if CVD event rates could be reduced in patients with type 2 diabetes who were at high risk for CVD events by targeting hyperglycemia, hypertension, and dyslipidemia. The ACCORD clinical trial was a randomized, double 2×2 factorial design study conducted at 77 clinical centers across the United States and Canada involving 10,251 middle-aged and older participants with type 2 diabetes who were at high risk for CVD events because of existing CVD or additional risk factors. Table 2.1 gives the sample sizes for each of the 8 treatment groups. Patients were treated and followed for 4-8 years.

The primary outcome measure for all 3 research questions was the first occurrence of a major CVD event, specifically nonfatal myocardial infarction (MI), nonfatal stroke, or cardiovascular death. Secondary outcomes included 1) macrovascular disease (primary outcome + coronary revascularization + hospitalization for heart failure), 2) total mortality, 3) each individual primary outcome, 4) congestive

Table 2.1 Clinical Trial Sample Sizes

	Blood Pressure Trial		Lipid Trial		
Glycemia Trial	SBP < 120	SBP < 140	Fibrate	Placebo	Total
HbA_{1c} < 6.0%	1,178	1,193	1,383	1,374	5,128
HbA_{1c} 7.0%-7.9%	1,184	1,178	1,370	1,391	5,123
	2,362	2,371	2,753	2,765	
Total	4,733		5,518		10,251

SBP = systolic blood pressure (mm Hg); HbA_{1c} = glycosylated hemoglobin

heart failure death or hospitalization for congestive heart failure, 5) composite microvascular disease outcome, including kidney and eye disease, 6) health-related quality of life, 7) cost-effectiveness.

Initial evaluation of recruitment and treatment protocols was done in a vanguard phase that began in January 2001. This phase recruited 1,174 participants (491 in the blood pressure arm, 683 in the lipid arm) after which protocol changes were made. The main trial began in February 2003 with the last patient randomized on October 29, 2005. The final visit was planned for June 30, 2009.

Measurement frequencies varied by treatment group, but were at least made at baseline, the end of the trial, and every 2 years in-between. White blood cells were stored for future DNA extractions for those patients who consented to genetic studies. Data were collected in 2 sub-studies to examine intervention effects on visual [Chew et al., 2007] (3,537 participants) and cognitive function [Williamson et al., 2007] (2,977 participants).

Analysis was performed on each of the three research hypotheses separately based on the χ^2 statistic from a likelihood ratio test obtained from proportional-hazards models. These models contained separate indicator variables for certain treatment groups, secondary participants, and clinical center networks.

2.1.1 Glycemia Trial

A detailed description of the ACCORD overarching glycemia trail is presented in [Gerstein et al., 2007].

2.1.1.1 Motivation

The specific research question was whether an intensive therapeutic strategy designed to lower HbA_{1c} to $< 6.0\%$ would reduce CVD event rates to a greater extent than a standard strategy that lowers and/or maintains HbA_{1c} between 7.0%-7.9%. Previous clinical trials showed conflicting results for intensive glycemic control groups. One trial showed a trend toward benefit while another showed a trend toward more CVD-related mortality, but neither were powerful enough to demonstrate significant differences between standard and intensive treatment groups. The ACCORD glycemia trial was designed to be the first definitive clinical trial on the possible benefit of targeting near-normal glycemic control in patients with type 2 diabetes and whether the benefit will exceed the potential risks. The glycemia trial of ACCORD was an unmasked, open-label, randomized trial with 10,251 participants.

2.1.1.2 Treatment Goals

Patients were treated with the goal of obtaining an HbA_{1c} level $< 6\%$ or 7.0%-7.9% for intense or standard treatments, respectively.

2.1.1.3 Treatment Regimen

Many classes of agents were permitted: 1) glimepiride (a sulfonylurea), 2) repaglinide (a rapid-acting secretagogue), 3) metformin (a biguanide), 4) rosiglitazone (a thiazolidinedione), 4) acarbose (an α-glucosidase inhibitor), 5) glargine, 6) neutral protamine Hagedorn and premixed insulins (longer-acting insulins), 7) aspart and regular insulin (shorter-acting insulins). In addition, all combinations of drugs were permitted.

2.1.1.4 Treatment Adjustments

Participants in the intensive treatment group had follow-up visits at months 1, 2, 3, 4, and every 2 thereafter with one or more extra visits or between-visits phone calls. Participants in the standard glycemic control group had visits at 1 month and every 2 or 4 months thereafter if they were also allocated to the intensive or standard blood pressure control arm of ACCORD, respectively. Investigators were given flexibility to individualize interventions including lifestyle approaches, behavioral therapies, self-titration, and drug adjustments. Participants in the intensive treatment group were started on at least 2 classes of agents. Dosage increase or addition of a new medication class occurred every month if HbA_{1c} levels were >6% or if >50% of premeal or postmeal capillary glucose readings were > 5.6 mmol/L (100 mg/dL) or > 7.8 mmol/L (140 mg/dL), respectively. For those participants in the standard glycemic control group, lifestyle and/or pharmacologic therapy was intensified whenever HbA_{1c} was > 8%, and antihyperglycemic drugs were reduced or withdrawn if HbA_{1c} persistently decreased to < 7% in patients experiencing hypoglycemia.

2.1.1.5 Hypoglycemia Monitoring and Risk Management

Severe hypoglycemic episodes were expected and extra attention was paid to monitore patient safety [Bonds et al., 2007]. Staff and patients were educated on symptoms and treatments of minor and sever hypoglycemic episodes. Hypoglycemia was monitored at the individual level and group level for each clinical cite. Severe episodes were reviewed by ACCORD investigators, staff memebers, and external review boards. If a patient had 3 or more sever episodes the intensive-group HbA_{1c} goal may have been relaxed to < 6.5% or < 7%.

2.1.2 Blood Pressure Trial

A detailed description of the ACCORD blood pressure trail arm is presented in [Cushman et al., 2007].

2.1.2.1 Motivation

The blood pressure arm of the ACCORD clinical trial was designed to test if an intensive therapeutic strategy that targets a systolic blood pressure of < 120 mm Hg would reduce the rate of CVD events more than a strategy that targets a systolic blood pressure of < 140 mm Hg in the context of good glycemic control for middle-aged or older men and women with type 2 diabetes who are at high risk for having a CVD event. Previous clinical trials demonstrated benefits for patients with diabetes where the treatment goal was a systolic blood pressure of 140 mm Hg. The ACCORD blood pressure trial was designed to be the first definitive clinical trial on the possible benefit of more aggressive blood pressure goals. The blood pressure arm of ACCORD was an unmasked, open-label, randomized trial with 4,733 participants.

2.1.2.2 Treatment Goals

Patients were treated with the goal of obtaining a systolic blood pressure of < 120 mm Hg or < 140 mm Hg for intense or standard treatments, respectively.

2.1.2.3 Treatment Regimen

Many classes of agents and combinations were permitted: 1) angiotensin-converting enzyme (ACE) inhibitors, 2) diuretics, 3) β-blockers, 4) dihydropyridine and nondihydropyridine calcium channel blockers (CCBs), 5) α-blockers, 6) angiotension II receptor blockers (ARBs), 7) sympatholytics, 8) α-/β-blockers, 9) a thiazide diuretic and a potassium-sparing diuretic, 10) a β-blocker and a diuretic, 11) an ACE inhibitor and a diuretic, 12) an ARB and a diuretic, 13) a dihydropyridine CCB and an ACE inhibitor. In addition, other drugs were allowed to be used as deemed appropriate by an investigator.

2.1.2.4 Treatment Adjustments

Trial participants had different follow-up visit schedules depending on the combination of glycemia and blood pressure treatments. Those participants in either the intense blood pressure or standard blood pressure + intense glycemic control groups had postrandomization visits at months 1, 2, 3, 4, and every 2 thereafter with "milepost" visits every 4 months. Those in standard blood pressure + standard glycemic control groups had visits at months 1, 4, and every 4 thereafter. Intense treatment would be modified if the patients systolic blood pressure was $>$ 120 mm Hg. Intense treatment modifications were conditioned on three cases: 1) if it was a milepost visit, another antihypertensive drug from a different class would be added, 2) if it was between milepost visits, another medication may be added or current doses adjusted, 3) if 5 drugs were currently prescribed a different drug class may be substituted. Standard treatment would be modified under two conditions: 1) if systolic blood pressure was $>$ 160 mm Hg at a single visit or $>$ 140 mm Hg at 2 successive visits, then current doses would be titrated or another drug would be added, 2) if systolic blood pressure was $<$ 130 mm Hg at single visit or $<$ 135 mm Hg at 2 successive visits, then downtitration or step-down would be carried out.

2.1.3 Lipid Trial

A detailed description of the ACCORD lipid trail arm is presented in [Ginsberg et al., 2007].

2.1.3.1 Motivation

The lipid arm of the ACCORD clinical trial was designed to test if an intensive therapeutic strategy that uses a fibrate to increase HDL and lower triglyceride levels in addition to a 3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitor (statin) to lower LDL would reduce the rate of CVD events compared with a strategy that uses a statin plus a placebo in the context of good glycemic control for middle-aged or older men and women with type 2 diabetes who are at high risk for having a CVD event. During

development of the lipid trial design there were 2 new strategies for managing dyslipidemia in high-risk populations: 1) reducing LDL to a lower concentration than currently advocated, 2) treat HDL and triglycerides in addition to LDL. The ACCORD lipid trial was designed to be the first definitive clinical trial for the later strategy. The lipid arm of ACCORD was a double-blind, randomized trial with 5,518 participants.

2.1.3.2 Treatment Goals

Patients were treated with the goal of obtaining a minimal LDL level of < 2.84 mmol/L (< 110 mg/dL) during the vanguard phase or < 2.59 mmol/L (< 100 mg/dL) during the main trial.

2.1.3.3 Treatment Regimen

All patients were treated with simvastatin to lower LDL. Patients were randomly assigned fenofibrate to lower serum concentrations of triglyceride and increase HDL.

2.1.3.4 Treatment Adjustments

Several changes to inclusion/exclusion criteria, dosage titration, and patient safety monitoring were made during the trial due to the results of several major clinical trials. During the vanguard phase, statin dose was initially assigned as 0, 5 or 10 mg/day based on LCL cholesterol levels and allowed to be titrated up to 20 mg/day during the vanguard phase. Results of the Medical Research Council/British Heart Foundation (MRC/ BHF) Heart Protection Study (HPS) showed that simvastatin reduced all-cause mortality and cardiovascular events in high-risk patients with low baseline LDL cholesterol [Heart Protection Study Collaborative Group, 2002]. Based on the results of the HPS, the lipid protocol was changed near the end of 2002 to administering a dose of 20 mg/day for all participants and no titration during the beginning of the main trial. Later, results of the PROVE-IT-TIMI 22 and TNT trials showed a benefit of lowering LDL for very high risk patients with regard to cardiovascular event rate [Cannon et al., 2004; LaRosa et al., 2005]. The August 2004 protocol incorporated these findings by

allowing for titration up to 40 mg/day if LDL consistently remained > 2.59 mmol/L (100 mg/dL). Fenofibrate dose was initially 160 mg/day for the vanguard phase and beginning of the main trial. Observations of increases in serum creatinine in some lipid trial participants caused concern about an association between fenofibrate use and deterioration of renal function. The protocol was changed in August 2004 to allow for the fenofibrate dose to be titrated to 54 mg/day if the patient's estimated glomerular filtration rate was between 30 to 50 mL/min per 1.73 m^2.

2.2 Results

Detailed results of the ACCORD glycemia, blood pressure and lipid trial arms are presented in [Gerstein et al., 2008], [Cushman et al., 2010] and [Ginsberg et al., 2010], respectively.

2.2.1 Glycemia Trial

The intensive regimen was terminated in February 2008, 17 months before the scheduled end of the study, due to higher mortality in the intensive-therapy group. The mean duration of follow-up at the time of discontinuation of the intensive regimen was 3.5 years. Stable mean HbA_{1c} levels of 6.4% and 7.5% in the intensive-therapy and standard-therapy groups, respectively, were achieved at 1 year and were maintained throughout the follow-up period. The intensive-therapy group was associated with greater exposure to drugs from every class, more frequent changes in dose or number of drugs used, and significantly higher rates of hypoglycemia, weight gain, and fluid retention.

There were fewer occurrences of the primary outcome in the intensive-therapy group. Rates of primary outcome began to separate in the two study groups after 3 years but the trend was not significant (rates of 6.9% and 7.2% in intensive and standard groups, respectively; hazard ratio 0.90; 95% confidence interval (CI) [0.78, 1.04]; p-value 0.16). Individual primary outcomes showed a significantly lower rate of nonfatal MI and a significantly higher rate of death from cardiovascular causes in the intensive-therapy group. The rates of death from any cause began to separate in the

two study groups after 1 year and persisted throughout the follow-up period and the rate of the intensive-therapy groups was significantly higher than the standard-therapy group (5.0% vs. 4.0%; hazard ratio 1.22; 95% CI [1.01, 1.46]; p-value 0.04).

2.2.2 Blood Pressure Trial

The mean duration of follow-up for the primary outcome was 4.7 years. Stable mean systolic/diastolic blood pressures were 119.3/64.4 mm Hg and 133.5/70.5 mm Hg for the intensive-therapy and standard-therapy groups, respectively, after the first year of therapy and were maintained throughout the follow-up period. The intensive-treatment group had significantly higher rates of serious adverse events, hypokalemia, and elevations in serum creatinine and lower glomerular filtration rates and frequency of macroalbuminuria.

There was no significant difference in the occurrences of the primary outcome between the two groups. The rate was 1.87% per year for the intensive-therapy group and 2.09% per year for the standard-therapy group (hazard ratio 0.88; 95% CI [0.73, 1.06]; p-value 0.20). No significant differences in deaths from any cause or deaths from cardiovascular cases were seen. The only significant differences of the secondary outcomes were for total stroke and nonfatal stroke, which were lower in the intensive-therapy group (p-values 0.01 and 0.03, respectively). Interpretation of the results from this arm of the ACCORD trial are complicated by the fact that the event rate of the standard-therapy group was 50% lower than expected.

2.2.3 Lipid Trial

The mean duration of follow-up for the primary outcome was 4.7 years. From the start to the end of the study, the mean LDL, HDL, and plasma triglyceride levels (all in mg/dL) changed from 100.0, 38.0, and 164 to 81.1, 41.2, and 122, respectively, for the fenofibrate group and from 101.1, 38.2, and 160 to 80.0, 40.5, and 144, respectively, for the placebo group. There was a lower incidence of both microalbuminuria and macroalbuminuria in the fenofibrate group than in the placebo group.

There was no significant difference in the primary or any secondary outcomes. However, there was evidence of an interaction between treatment and gender for the primary outcome where fenofibrate therapy seemed to benefit men and harm women. There was also a nonsignificant suggestion of heterogeneity of response for those patients with highest third triglyceride level and lowest third HDL level when compared to all other patients.

2.3 Motivation for Genetic Association Study

The ACCORD trial was designed to test treatment strategies for lowering levels of blood sugar, blood pressure, and dyslipidemia. Results of the trial were disappointing since there were no significant differences in the primary outcomes between intensive and standard treatment groups. The lack of significant treatment effects could be due to differential response of individuals to particular therapeutic regimens due to polymorphisms creating pharmacological differences. Expanding such a large clinical trial by including a genome wide association study (GWAS) could potentially identify subgroups of patients who do significantly benefit from or are harmed by more intensive therapeutic regimens. However, the variety of drugs and extremely flexible protocols make it hard to identify distinct treatment plans and reduce the power to detect associations due to smaller sample sizes of subgroups and an increased number of statistical tests.

Findings from previous GWAS studies related to the ACCORD trial outcomes and drug formulary may be helpful in replicating any potential findings of an ACCORD GWAS and useful for generating lists of candidate genes that should be closely investigated. Below are a few of the more recent relevant GWAS studies.

2.3.1 CVD risk

Middelberg et al. performed a multivariate GWAS on biochemical traits related to CVD [Middelberg et al., 2011]. The study included adolescent (2548 individuals) and adult (9145 individuals) twins and

relatives in Australia from 4986 families. Analysis was done using the measurements of 13 traits related to CVD (serum alanine aminotransferase, aspartate aminotransferase, butrylycholinesterase (BCHE), C- reactive protein (CRP), ferritin, gamma glutamyltransferase (GGT), glucose, HDL, insulin, LDL, triglycerides, uric acid, and body-mass index). Genotyping was done using Illumina 610k, 317k, or 370k chips and untyped single nucleotide polymorphisms (SNPs) were imputed using MACH.

Several trait measurements were log-transformed and all were adjusted for the effects of age, age^2, gender, gender×age and gender×age^2. Multivariate analysis was performed using PLINK on the standardized residuals. To reduce computational burden, only one individual per family was used for the multivariate analysis. Any locus identified as significant would then be confirmed with univariate association tests using all 11,683 genotyped individuals.

A total of 766 SNPs in 11 loci were associated with the biochemical traits according to the multivariate analysis. Univariate analyses of the most strongly associated SNPs at each locus showed strong or close to genome-wide evidence of associations with more than one trait for 3 of the loci identified in the multivariate analysis. The closest genes to these three loci are *LPL, OASL*, and *TOMM40/APOE-C1-C2-C4*.

2.3.2 Metformin

Zhou et al. performed a GWAS on glycemic response to metformin in 1,024 individuals with type 2 diabetes using a large Scottish observational genetic cohort [Zhou et al., 2011]. Participants were treated with the goal of lowering HbA_{1c} to below 7% in the first 18 months of therapy. Genotyping was done using the Affymetrix Genome-Wide Human SNP Array 6.0 microarray resulting in 705,125 SNPs after quality control filtering. Data were imputed using IMPUTE for a total of 2.2 million imputed SNPs.

Logistic and multiple linear regression models was used to test for association and included covariates for baseline HbA_{1c} level, creatinine clearance, adherence, daily dose, treatment group, and gap between baseline measurement and start of metformin treatment.

Genomic inflation of metformin response was 1.003 and 0.998 for logistic and linear regression modeling, respectively. The strongest evidence for association with metformin response was observed in 14 SNPs that mapped to a 340-kb linkage disequilibrium (LD) block on chromosome 11q22. The most strongly associated SNP (rs11212617) had an allelic odds ratio of 1.64 (95% CI [1.37, 1.99]; p-value=$1.9\text{x}10^{-7}$) and explained 2.5% of the variance in metformin response. This SNP was tested in two replication cohorts and found to be significant in the combined dataset. The LD block contained several genes, but the authors only considered *ATM* as a possible candidate gene.

2.3.3 Simvastatin

Hopewell et al. performed a GWAS on LDL and apolipoprotein B (ApoB) responses to simvastatin in the Heart Protection Study [Hopewell et al., 2013]. Participants were treated with either 40 mg simvastatin or matching placebo daily for ∼5 years. The prespecified primary outcome was the first occurrence after randomization of either nonfatal MI or coronary death, coronary or non-coronary revascularizations, or any stroke. Genotyping was done using the Illumina 610k Quad panel on 3895 individuals resulting in 546,210 SNPs after quality control filtering. Selected SNPs were then genotyped using custom I.PLEX panels in the remaining 14,810 participants. In addition, 36 candidate SNPs in loci that had previously been associated with lipid response, statin pharmacokinetics or pharmacodynamics, statin-related side-effects, LDL levels, or coronary heard disease risk were custom genotyped.

LCL-C and ApoB responses were defined by the change in log-transformed lipid levels from before starting statin therapy to postrandomization visits 4-6 weeks later in compliant individuals. Associations between genetic variants and lipid response to statins were estimated using linear regression. Cox proportional hazard models were used to estimate genotype effects on the primary outcome.

Nine SNPs were considered as genome-wide significant in the 3895 fully genotyped individuals. Eight of these were successfully genotyped in the remaining 14,810 participants and none were

confirmed as significant. Ten of the 36 candidate SNPs, corresponding to 5 loci (*LPA, APOE, SLCO1B1, CELSR2/PSRC1/SORT1, ABCC2*), reached statistical significance and had small effects (0.5–3.0% per allele compared to overall 42% LDL reduction). None of these SNPs were significantly associated with differences in either the proportional or the absolute reductions in major vascular events.

2.3.4 Fenofibrate

Aslibekyan et al. performed a GWAS on the adiponectin response to fenofibrate in a population with a comparatively low prevalence of dyslipidemia [Aslibekyan et al., 2013]. Adiponectin is a protein with moderate to high heritability ($h^2 = 0.35$–0.55) that has been linked to progression of CVD. Participants were from the GOLDN study, consisting of European American pedigrees with at least two siblings, with a final sample size of 793 after applying exclusion criteria. Participants were given a daily treatment of 160 mg of micronized fenofibrate over the period of three weeks. Genotyping was done using the Affymetrix Genome-Wide Human SNP Array 6.0 and untyped SNPs were imputed using MACH for a total of 2,543,887 SNPs, of which 584,029 were from the Affymetrix 6.0 chip after quality control filtering.

Phenotypes were based on log-transformed circulating adiponectin levels. Confounding due to population stratification was assessed using principal component analysis. No associations with the first 10 principal components were found so the final models were not adjusted for ancestry using principal components. Single SNP GWAS was performed using linear mixed models that were adjusted for gender, age, and center as fixed effects and kinship as a random effect. Gene-based association tests were performed using the Versatile Gene-Based Association Study algorithm. Biological pathway analysis was performed using the Meta-Analysis Gene-Set Enrichment of Variant Association software.

Genomic inflation of fenofibrate response was 1.07. The strongest evidence for association with fenofibrate response was observed in the 12q24 genomic region. Only one SNP (rs2384207) in the *DDX54* gene was considered as genome-wide significant (p-value=$5\text{x}10^{-8}$) and explained 2.9% of the phenotype variation. Six variants in the adjacent *TPCN1* gene showed suggestive associations

(p-values $< 3x10^{-6}$), but were in LD ($r^2 > 0.80$) with the lead SNP. Other implicated loci were in the *SHANK2, ENPP3, C6,* and *KDM4C* genes.

2.3.5 Baseline Lipid Levels

While the investigation of genetically influenced drug response is the ultimate goal of the whole-exome study of the ACCORD data, there are a limited number of current GWAS studies with which to replicate any findings. In addition, the treatment regimens of the ACCORD clinical trial are very complex and make it difficult to construct statistical models without confounding from coadministration of drugs. A more straightforward study that could be used to validate the quality control and imputation of genetic data and the analysis plan is the association between genetic variants and baseline lipid levels. Associations between lipid plasma concentrations and genetic markers have been of great interest because total cholesterol (TC), low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL) and triglycerides (TG) are heritable risk factors for cardiovascular disease [Kathiresan et al., 2007]. These traits have been heavily studied and there are large meta-analyses for both common [Consortium, 2013; Teslovich et al., 2010; Waterworth et al., 2010] and rare [Liu et al., 2014] variants.

Although testing the association of baseline lipids has been thoroughly investigated, the ACCORD analysis will contribute results from a large dataset consisting of type 2 diabetic patients. Significant variants can easily be compared to previous findings and used to validate the quality of the genetic data and analysis plan. In addition, the genotype data for the ACCORD trial will be one of the first studies focused on exome variants, providing the potential for finding novel genes with significant effects on baseline lipid levels.

Chapter

3

ACCORD Genotype Datasets

This chapter describes the quality control (QC) pipeline that was applied to the raw ACCORD genotype data. QC metrics were used to identify and/or filter out problematic plates of samples, individual samples on plates and genotyping probes. Those samples and probes that passed the QC filters were then used to impute a large number of variants from a 1000 Genomes reference panel. Genotyped and imputed variants that were annotated as functional for a specific gene were then grouped together to create a rare variant analysis dataset.

3.1 Quality Control Pipeline

All ACCORD samples with available DNA were genotyped using the Affymetrix Axiom Biobank Genotyping Array. This platform became available on November 1, 2012 and this project was one of the first to use the array. The quality control (QC) pipeline for this project is very important since this array has not been heavily used. The array contains several marker sets. There are roughly 250k genome-wide association (GWA) markers, 250k exome markers focusing on rare variants, 70k novel loss-of-function markers, 20k eQTLs and 2k pharmacogenomic markers. The GWA markers were selected to maximize genome-wide coverage of imputed variants in European, Asian, African and Latino populations [*Axiom biobank genotyping arrays*].

Samples were processed at UNC Functional Genomics Core and then grouped into a single batch for genotype calling. Genotype calling and several QC steps were done following the best practices guidelines provided by Affymetrix [*Axiom genotyping solution analysis guide* 2011; *Best practice supplement to axiom genotyping solution data analysis user guide* 2011]. Many of the QC metrics are described in Laurie et. al. 2010 [Laurie et al., 2010]. Genotype calling was done using Affymetrix Power Tools (v1.15.0) with the Axiom GT1 algorithm, which is a modified version of the BRLMM-P algorithm [*BRLMM-P: a Genotype Calling Method for the SNP 5.0 Array* 2007] that adapts generic prior cluster positions to the data using an EM algorithm. A flowchart of the QC pipeline is shown in Fig. 3.1. All metric computations and subsequent filtering were done using custom R [R Core Team, 2014] and

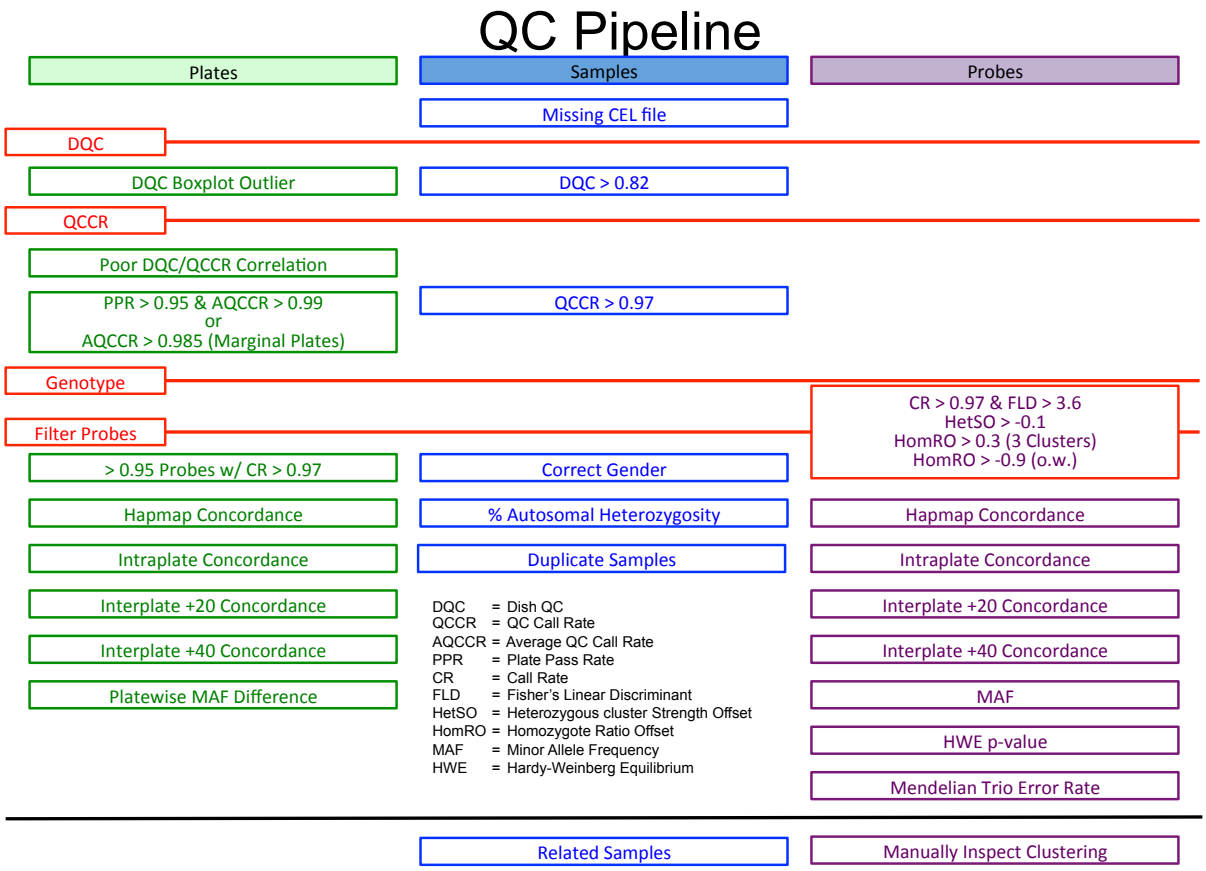


Figure 3.1 Flowchart of QC pipeline. QC metrics are computed for samples, plates and probes. The red boxes and lines represent steps following the best practices guidelines of Affymetrix. QC steps below the black line are analysis-specific.

Perl scripts except for computation of the first set of probe filtering metrics, which was done using the Affymetrix R script `metrics.r` based on the sample code in Appendix D of the Axiom Genotyping Solution Data Analysis Guide.

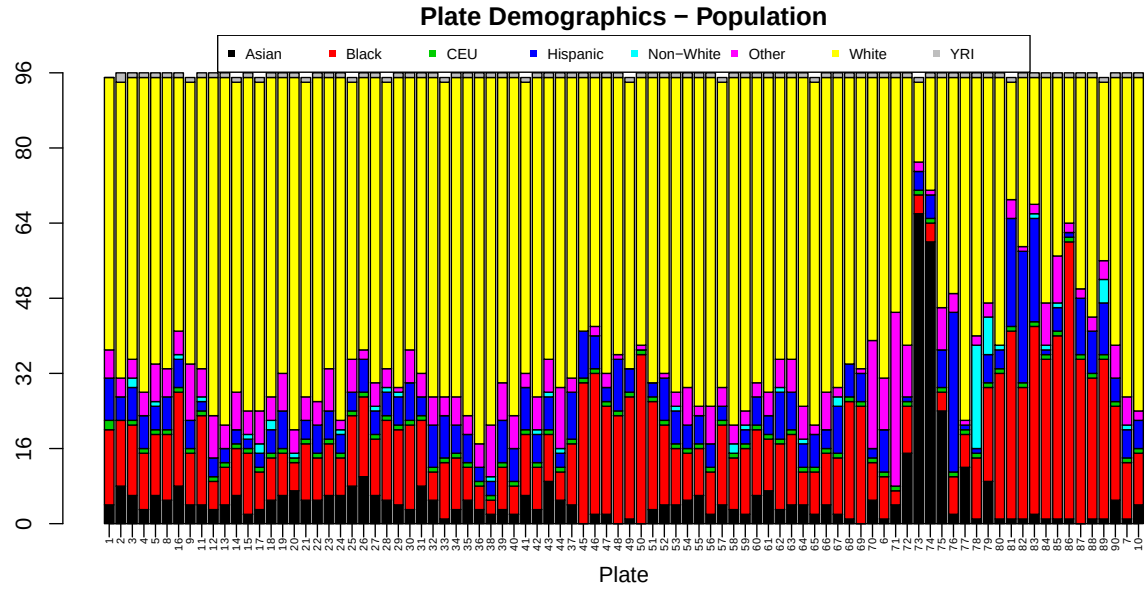
3.1.1 Pre-Genotyping Quality Control

3.1.1.1 Plate Layout

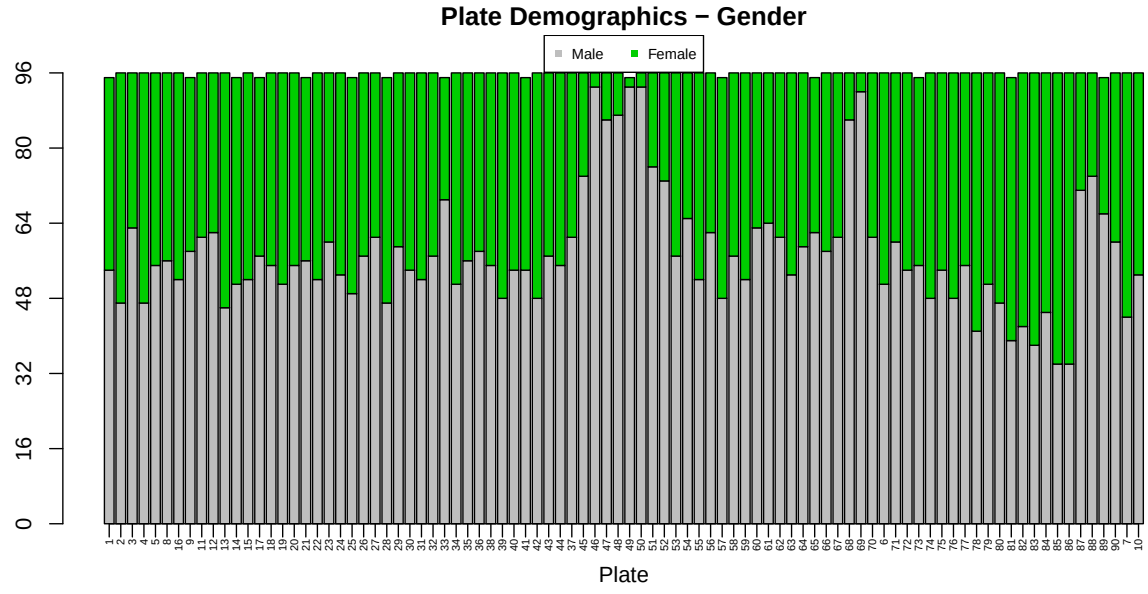
A total of 90 96-well plates were processed. Each plate had 1 CEU (Utah residents with ancestry from northern and western Europe) and 1 YRI (Yoruba in Ibadan, Nigeria) HapMap [The International HapMap Consortium, 2005] sample in wells C8 and G5, respectively. Exceptions to this were plate 1 which had 2 CEU and 0 YRI samples and plate 2 which had 0 CEU and 2 YRI samples. The HapMap samples were used to compute a concordance measurement between genotype calls and expected genotypes. In addition, each plate contained 3 duplicate samples: well E7 is a duplicate of well B2 on the same plate; well C4 on plate x is a duplicate of well G10 on plate $x+20$; well F3 on plate x is a duplicate of well D11 on plate $x+40$. Samples were processed in the order they were provided to the processing center, which prevented the randomization of ethnic background and race among the plates. Figures 3.2a and 3.2b depict the population and gender demographics for each plate, respectively, with the plates being in chronological order of processing. Plates 73 and 74 clearly have a higher proportion of Asian samples. Plates near the end of the processing have a lower proportion of white samples. Several of the plates have a higher proportion of males, which may be the result of samples being collected at Veterans Affairs medical centers.

3.1.1.2 Signal Quality

Dish QC (DQC) was used to remove samples with poor signal quality. DQC measures signal intensity of the two channels at non-polymorphic locations with known genomic sequence. The two channels monitor when a probe ligates to an A or T base (AT Signal) or to a G or C base (GC Signal). The



(a)



(b)

Figure 3.2 a) Population and **b**) gender demographics for genotyped plates. Ethnic background is self-reported. Plates are in chronological order of processing. HapMap samples: CEU - Utah residents with ancestry from northern and western Europe, YRI - Yoruba in Ibadan, Nigeria.

contrast of the two channels is computed as contrast $= \frac{\text{AT Signal}-\text{GC Signal}}{\text{AT Signal}+\text{GC Signal}}$. Distributions of contrast values are computed separately for each channel and DQC is a measure of the resolution between the distributions. A value of 0 indicates no resolution between the distributions while a value of 1 indicates perfect resolution. Twelve samples with a DQC < 0.82 were removed.

DQC was also used to remove plates with poor signal quality. Fig. 3.3 shows the plate-based DQC distributions using box plots in chronological order of plate processing. Any plate whose upper

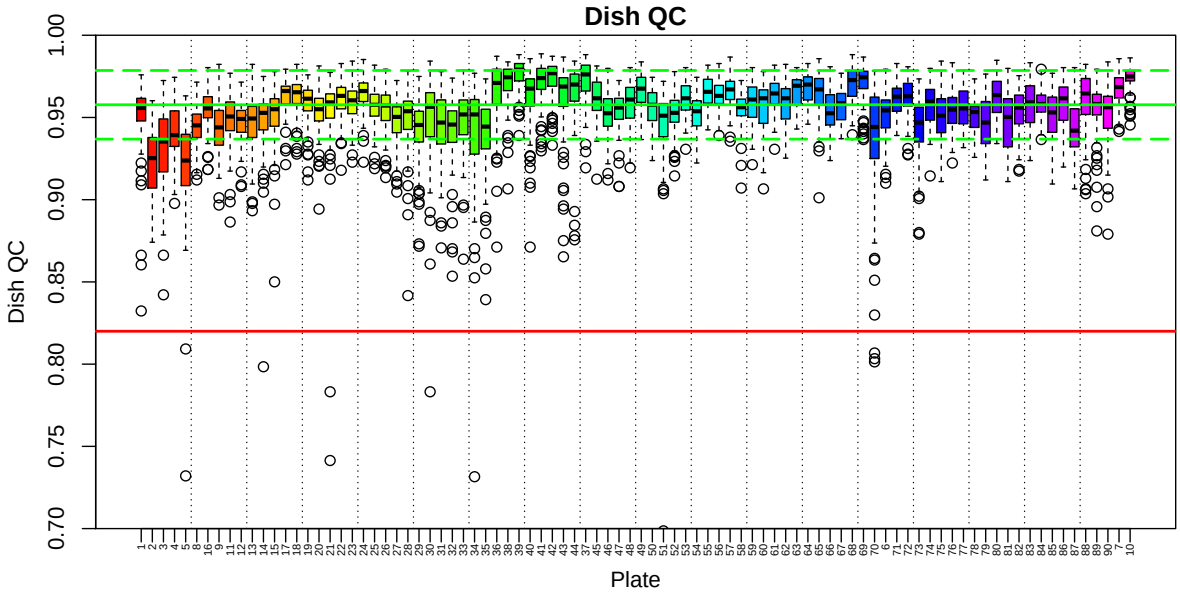


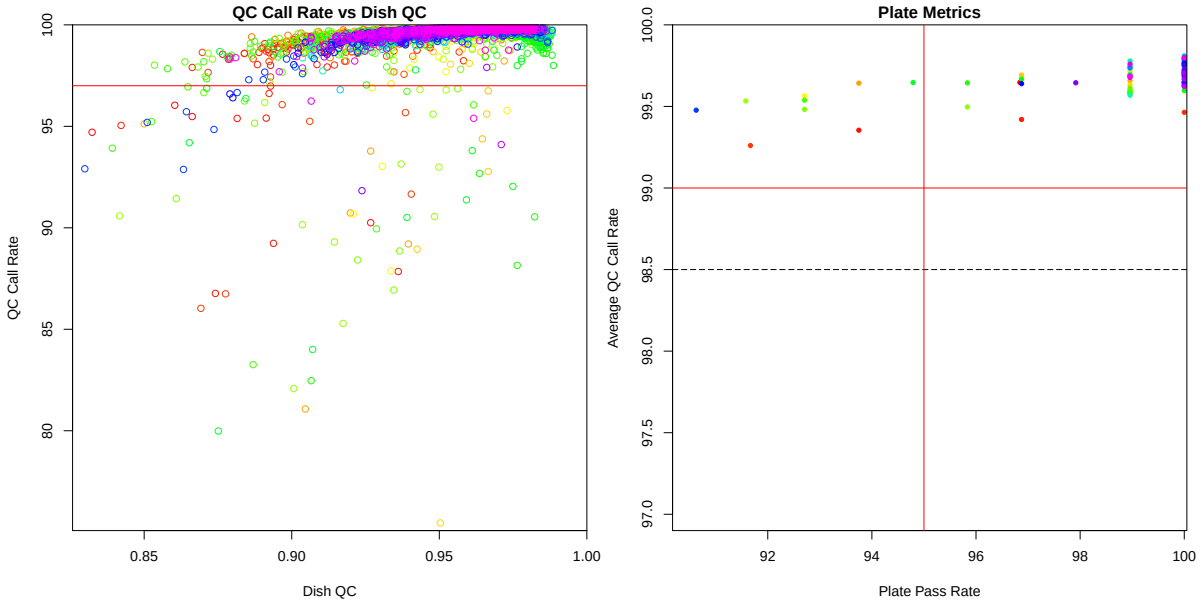
Figure 3.3 Box plots of plate-based DQC distributions in chronological order of plate processing. The red line is the DQC threshold of 0.82. The solid green line is median of plate median DQC values and the dotted green lines are the median of medians $\pm$ 2 standard deviations of medians.

quartile falls below 2 standard deviations below the median of medians is considered an outlier and should be dropped or re-processed. Plates 6, 7 and 10 were identified as outliers and were re-processed. Plates 2 and 5 are nearly classified as outliers, however, they were not re-processed due to monetary and time constraints nor were they dropped.

3.1.1.3 Quality Control Call Rate

While DQC is sufficient for removing samples that produce poor signal quality, it is unable to detect other issues such as sample contamination. Contaminated samples can be identified based on low genotype call rates. A QC call rate metric was computed based on the genotyping performance of a set of 20,000 probes whose genotypes are representative of the expected performance of the standard tiling design. Eighty-seven samples with QC call rates < 0.97 were removed.

The relationship between DQC and QC call rates can be used to check for sample contamination. High DQC values are expected to correlate with high QC call rates. Sample contamination may have occurred if a plate is found to have a large number of samples with high DQC but low QC call rates. Fig. 3.4a is a plot of the QC call rate versus DQC. Most plates behave as expected with an arc going to



(a) **(b)**

Figure 3.4 QC call rate metrics. **a**) Scatter plot of QC call rate and DQC. The red line is the QC call rate threshold of 0.97. **b**) Scatter plot of average QC call rate and plate pass rate. The vertical red line is the plate pass rate threshold of 0.95. The horizontal red line is the average QC call rate threshold of 0.99 and the dashed line is the marginal threshold of 0.985. Marker colors correspond to plate colors in Fig. 3.3.

the left and slightly down. However, there is a small group of samples that drop vertically in QC call rate while maintaining high DQC values. The QC call rate for these samples is still very high and no samples were removed.

The QC call rate was also used to identify potential outlier plates. Two additional metrics were used for plate QC: 1) the plate pass rate, which is the number of samples that passed the DQC and QC call rate metrics divided by the total number of samples on the plate and 2) the average QC call rate, which is the mean QC call rate of samples passing both DQC and QC call rate metrics. The Affymetrix guidelines classify high-quality plates as those that have a plate pass rate > 0.95 and an average QC call rate > 0.99. Marginal plates are those with average QC call rates > 0.985. Fig. 3.4b depicts both of these metrics for each plate. There are several plates with lower plate pass rates indicating poorer sample quality, however, all of the plates have excellent average QC call rates and were retained.

3.1.2 Post-Genotyping Quality Control

3.1.2.1 Genotype Clusters

The Affymetrix script `metrics.r` was used to compute several metrics for evaluating the performance of the genotype calling algorithm. The script removed all mitochondria and Y chromosome probes (101 and 122 probes, respectively), then computed probe call rate (CR), Fisher's linear discriminant (FLD), heterozygous cluster strength offset (HetSO) and homozygote ratio offset (HomRO) for the remaining probes. An example cluster plot is shown in Fig. 3.5 with axes signal size, $(log_2(A)+log_2(B))/2$, and contrast, $log_2(A/B)$, where A and B are the summary intensities for alleles A and B, respectively.

A low CR reflects poor cluster resolution which may result in inaccurate genotypes in those samples that are called. Also, if the no-calls are non-randomly distributed among the genotypes they may lead to false positive associations.

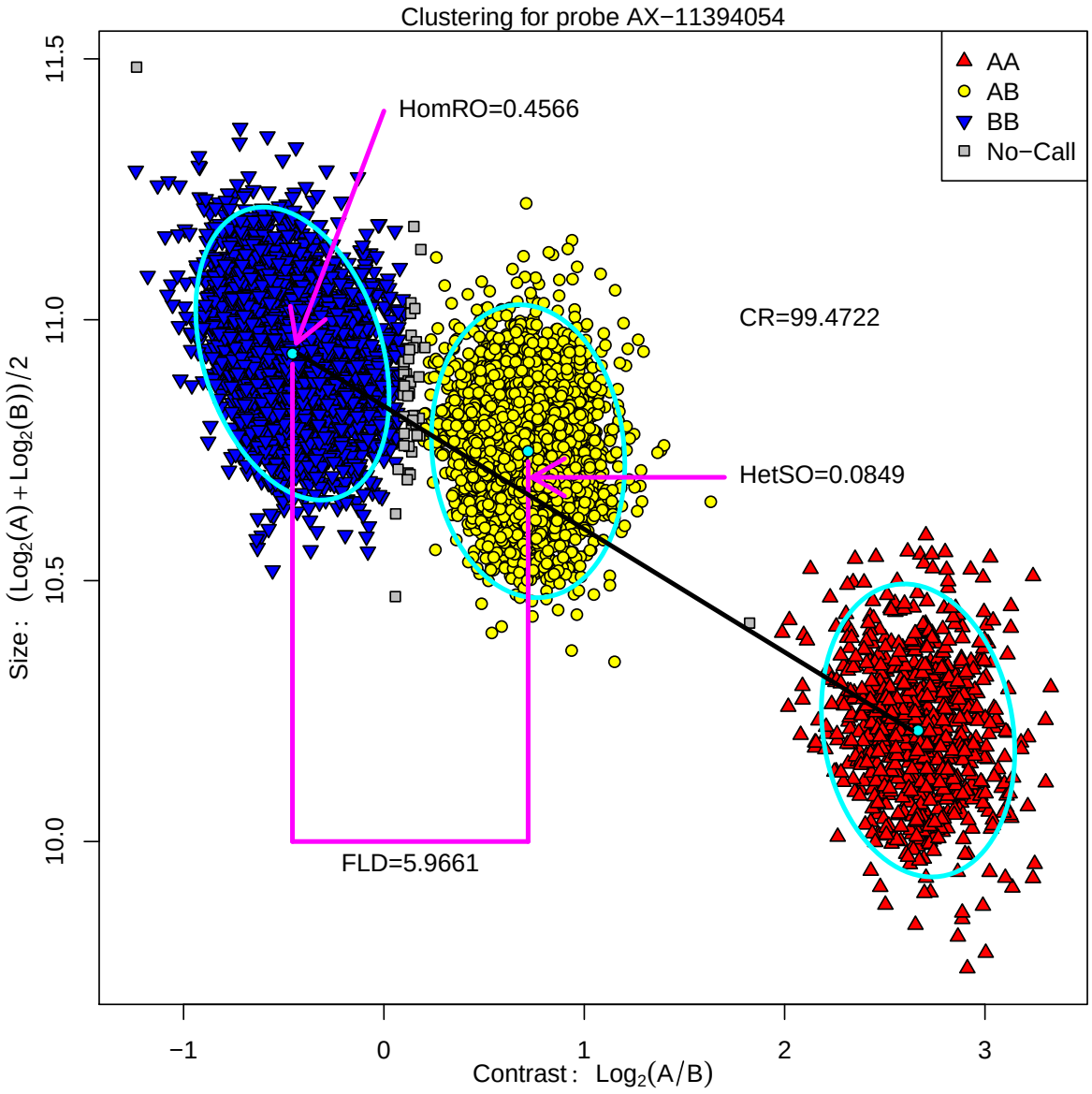


Figure 3.5 Clustering example for probe AX-11394054. Size and contrast are computed using summarized intensities for alleles A and B. Turquoise points and ellipses are diagrammatic representations of the posterior probability distribution location and variance for each cluster, respectively. Magenta lines and arrows indicate important components for metric computations. The black line connects the centers of the AA and BB genotype clusters, which is used when computing HetSO.

FLD is a measure of cluster separation and is computed as

$$FLD = \min_{i \in \{AA, BB\}} \frac{|C_{AB} - C_i|}{s_p},$$

where C_g, $g \in \{AA, AB, BB\}$, is the signal contrast at the center of the genotype clusters and s_p is the square root of variance pooled across all three distributions. Small FLD values will often indicate a mis-clustering event.

HetSO is a measure of probe function and is the vertical distance from the center of the heterozygous cluster to the line connecting the centers of the homozygous clusters computed as

$$HetSO = S_{AB} - S_{BB} + (S_{AA} - S_{BB}) \frac{C_{AB} - C_{BB}}{C_{AA} - C_{BB}},$$

were S_g is the signal size at the center of the genotype clusters. A high degree of mismatch between a sample and the reference genome in the probe sequence will result in low A and B intensities causing the sample to be grouped in the heterozygous cluster. If this occurs for enough samples the algorithm will include them in the AB cluster rather than correctly making them no-calls, decreasing HetSO.

HomRO is a measure of correct cluster labeling. There are 3 cases for computing HomRO: if $C_{AA} > 0$ and $C_{BB} < 0$, then $HomRO = \min(C_{AA}, |C_{BB}|)$; else if $C_{AA} > 0$ and $C_{BB} > 0$, then $HomRO = -C_{BB}$; else if $C_{AA} < 0$ and $C_{BB} < 0$, then $HomRO = C_{AA}$. The contrasts at the center of the AA and BB clusters are expected to be positive and negative, respectively. A small or negative HomRO value tends to indicate that the clusters were mislabeled by the genotyping algorithm.

Probes were retained if they passed the following thresholds: CR $\geq$ 0.97, FLD $\geq$ 3.6 (for $>$ 1 cluster), HetSO $\geq$ -0.1 (for $>$ 1 cluster), HomRO $\geq$ 0.3 (for 3 clusters) or $>$ -0.9 (for $<$ 3 clusters). If the genotyping algorithm resulted in a single cluster the FLD and HetSO metrics were not used for filtering. A total of 36,483 probes were removed for failing at least one of the clustering metrics in addition to the 223 removed before computing the metrics. The filtered probe set containing the remaining 591,973

probes was used in the following post-genotyping QC steps.

3.1.2.2 Plates

It is important to identify any plates that may cause batch effects, which could result in false positive associations. This is accomplished by computing several metrics for each plate and identifying any outlier plates or unexpected patterns.

3.1.2.2.1 Call Rate

An additional metric used to classify high-quality plates requires that at least 95% of the probes should have a call rate ≥ 0.97 over the samples that passed the DQC and QC call rate metrics. Fig. 3.6 plots the percent of probes with a call rate greater than 97% for each plate. Most of the plates pass this metric

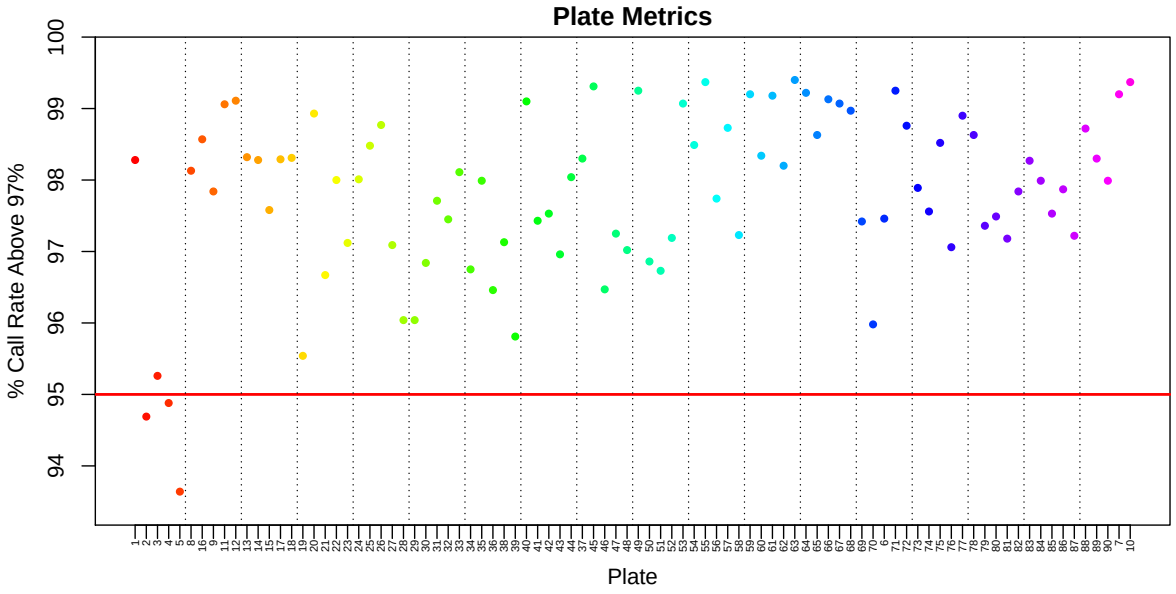


Figure 3.6 Percentage of probe call rates above 97%. The red line is the suggested threshold of 95%.

except plates 2, 4 and 5. Plates 2 and 5 also had relatively lower DQC values when compared to the

other plates, which may explain a lower % of probes exceeding a call rate above 97%. All plates were retained.

3.1.2.2.2 Concordance

Duplicate and HapMap samples were used to compute a set of concordance values for each plate as shown in Fig. 3.7. Concordance was computed as n_{match}/n_{call}, where n_{call} is the total number of

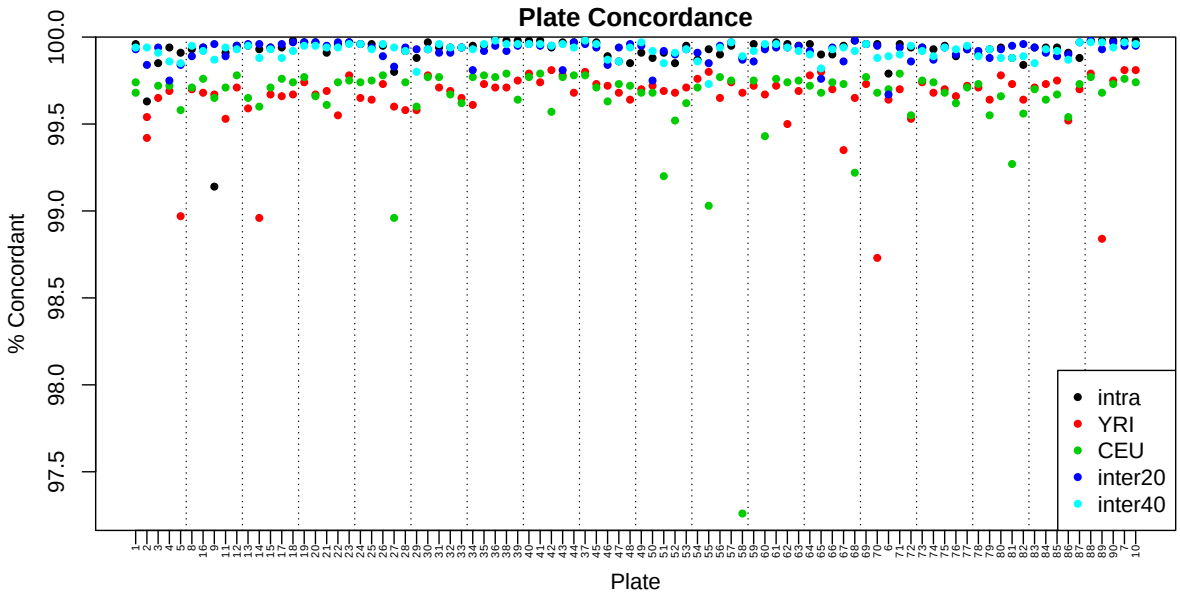


Figure 3.7 Concordance measurements for duplicate and HapMap samples on each plate. Duplicates samples were placed on the same plate (intra), 20 plates apart (inter20) and 40 plates apart (inter40).

probes where both samples had genotype calls and n_{match} is the number of probes with matching calls. A no-call in either sample resulted in the probe being removed from the concordance computation. The concordance for duplicate samples were computed using the entire filtered probe set while those for HapMap samples used a subset of probes annotated with dbSNP rsIDs that matched those in the reference genotypes. All duplicate samples have concordance values > 99%. Concordance of

HapMap samples was slightly lower than the duplicate samples and this presumably reflects the difference in genotyping platforms. Only the CEU sample on plate 58 had a noticeably lower concordance. All plates were retained.

3.1.2.2.3 Plate-Wise Minor Allele Frequency Differences

Under a randomized plate layout the minor allele frequency (MAF) values should not systematically differ on one plate compared to the remainder of the plates. Any shift in MAFs can be detected using a homogeneity test [Pluzhnikov et al., 2008]. Each probe will have a chi-squared test statistic with 1 degree of freedom given by

$$t = n_i n_{(i)} \frac{(\tilde{p}_i - \bar{p}_i)^2}{n\bar{p}(1-\bar{p})},$$

where $\tilde{p}_i$ is the MAF for a probe on the i-th plate (n_i samples), $\bar{p}_i$ is the average MAF over all other plates ($n_{(i)}$ samples), and $\bar{p}$ is the average MAF over all plates (n samples). These test statistics are then averaged across probes. Fig. 3.8 shows the average test statistics for each plate. When looking at the overall sample set (excluding HapMap samples) there are several plates near the end of processing with large average test statistic values, most noticeably plates 73 and 74. However, the pattern of average test statistics mimics the plate demographics change near the end of processing as seen in Fig. 3.2a. The plate-wise MAF difference test statistics were then computed for sample sets grouped by ethnic background. Most plates have average test statistics close to 0.5 with the Asian subset having the most variation. This behavior makes sense because the number of Asian samples on most plates is very small so this test statistic is comparing the MAF of only a few Asian samples with remainder and noise is not averaged out as much. All plates were retained.

3.1.2.3 Samples

Metrics computed for each sample can help identify problems such as sample contamination or mis-labeling.

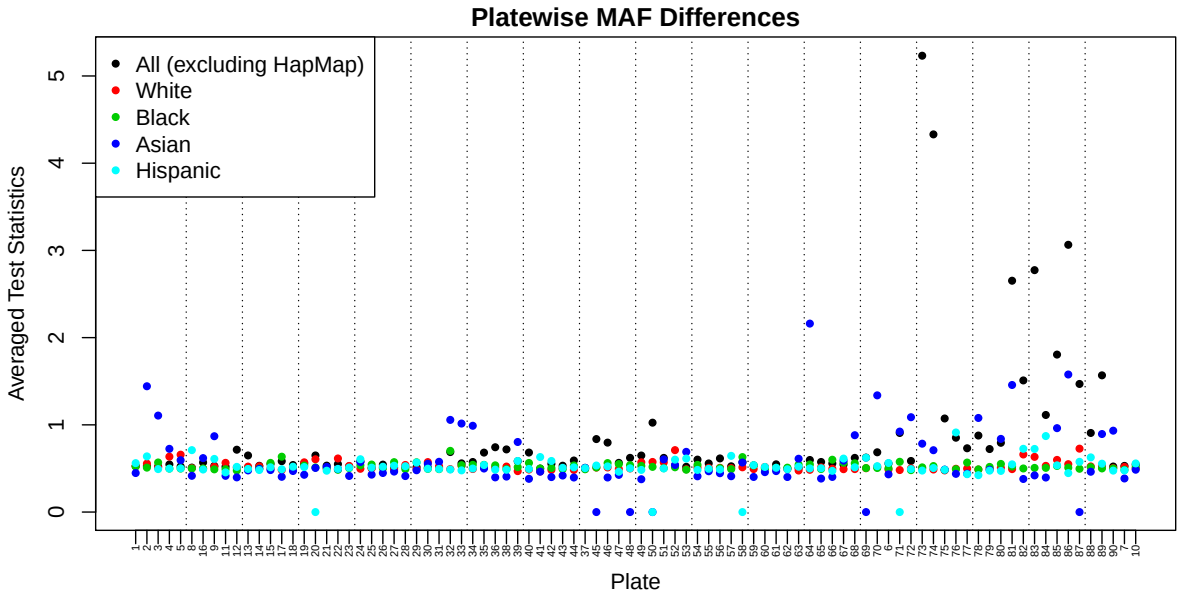


Figure 3.8 Average plate-wise MAF difference test statistics for each plate. Statistics were computed for the entire dataset excluding HapMap samples and separately for each of the four main ethnic backgrounds.

3.1.2.3.1 Predicted Gender

The gender of each sample was predicted using the average intensities of 1500 copy number probes on chromosomes X and Y. Any sample with a Y/X ratio greater than 1.00 was defined as male, a ratio less than 0.54 was defined as female and all other samples were predicted as unknown. The predicted genders for each sample was compared to the gender recorded in the ACCORD clinical trial forms and plotted using the average probe intensities as shown in Fig. 3.9. There were 12 sample with incorrect gender predictions, of which 3 sets were male-female pairs with consecutive laboratory processing IDs. This suggests that there was a sample mixup for each of the pairs at some point between collection and processing. Evidence found during a collaborative project points towards a sample mixup occurring before the samples arrived at the processing center and all 12 samples with incorrect gender predictions were removed. Four samples had relatively low average intensities. These samples had DQC values > 0.92, however, they were in the bottom 1% of QC call rate for those

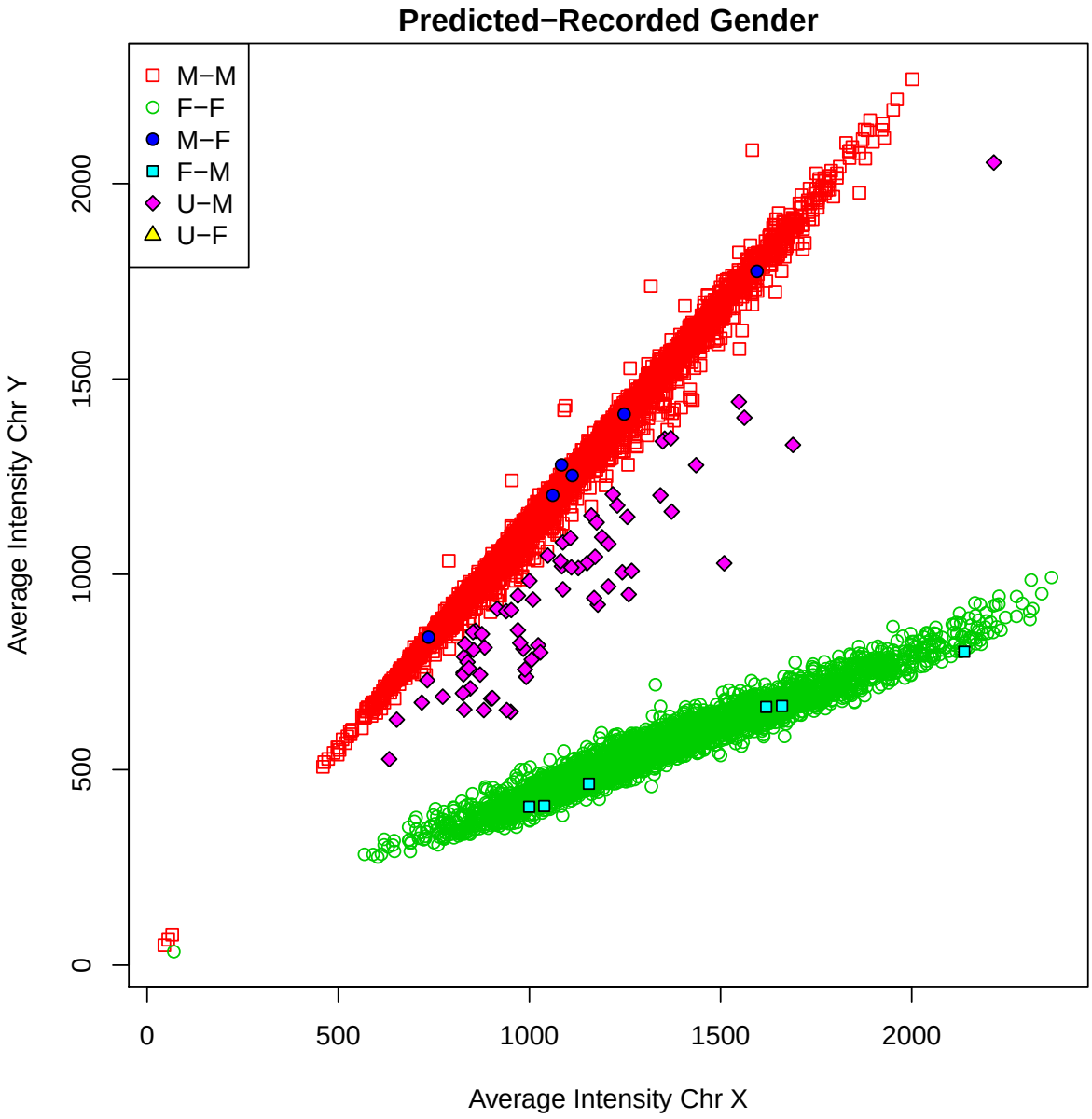
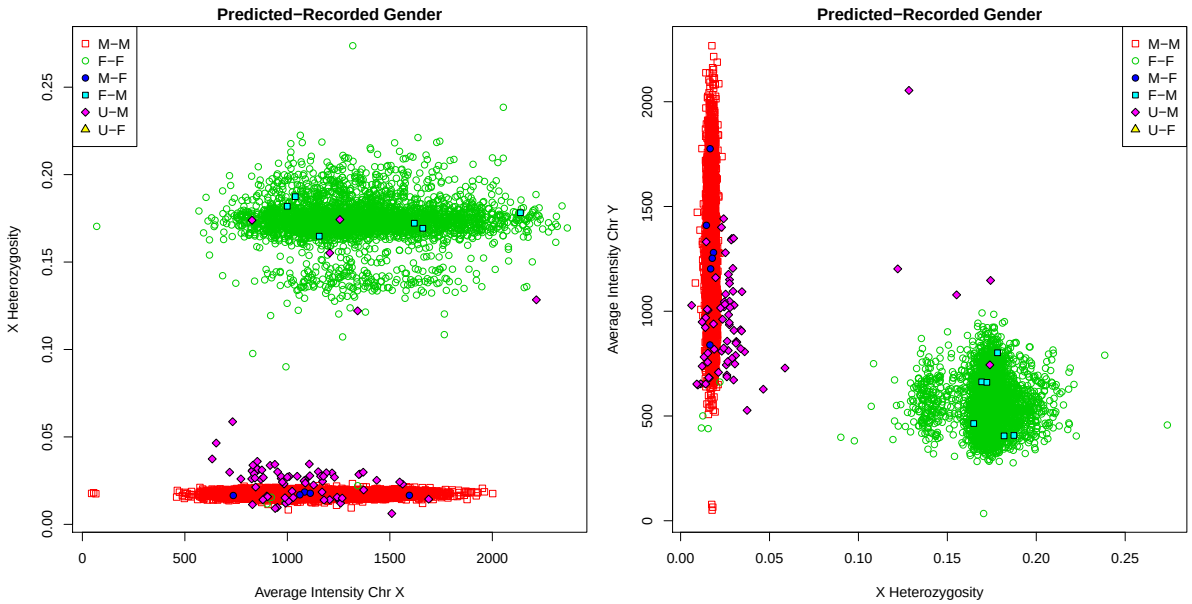


Figure 3.9 Comparison of predicted and recorded gender. Legend lists predicted gender first with M = male, F = female and U = unknown.

samples passing the 0.97% threshold. Although these 4 samples have low average intensities for the gender computation probe set, the gender predictions were correct and they passed the initial QC steps so they were not marked for removal.

Seventy-three samples were predicted as having an unknown gender. Further gender analysis was done by looking at chromosome X heterozygosity versus both chromosome X and Y average intensities (Fig. 3.10). Five samples that were recorded as male but predicted as unknown had chromosome



(a) **(b)**

Figure 3.10 Chromosome X heterozygosity versus average intensity of chromosomes X (**a**) and Y (**b**). Legend lists predicted gender first with M = male, F = female and U = unknown.

X heterozygosity similar to the female samples and were removed. There were 4 samples correctly predicted as female that had chromosome X heterozygosity similar to the male samples. However, three of those were HapMap samples so only the remaining female sample was removed. The four samples with low intensity values have chromosome X heterozygosity in line with their respective

genders, which strengthens the decision to keep these samples in the dataset. In total, 18 samples were marked for removal due to gender issues: 12 for incorrect computed genders, 5 male samples with predicted unknown gender and high chromosome X heterozygosity and 1 female sample with low chromosome X heterozygosity.

3.1.2.3.2 Percent Autosomal Heterozygosity

Relatively high heterozygosity of autosomal chromosomes may indicate sample contamination. A higher autosomal heterozygosity could occur if two samples are combined where each has opposite homozygosity at a large number of alleles resulting in an increase in the number of heterozygous genotype calls in the contaminated sample. The autosomal heterozygosity of the sample set ranged from 10.84% to 17.52% (Fig. 3.11). There were 9 samples with autosomal heterozygosity < 12.5% and

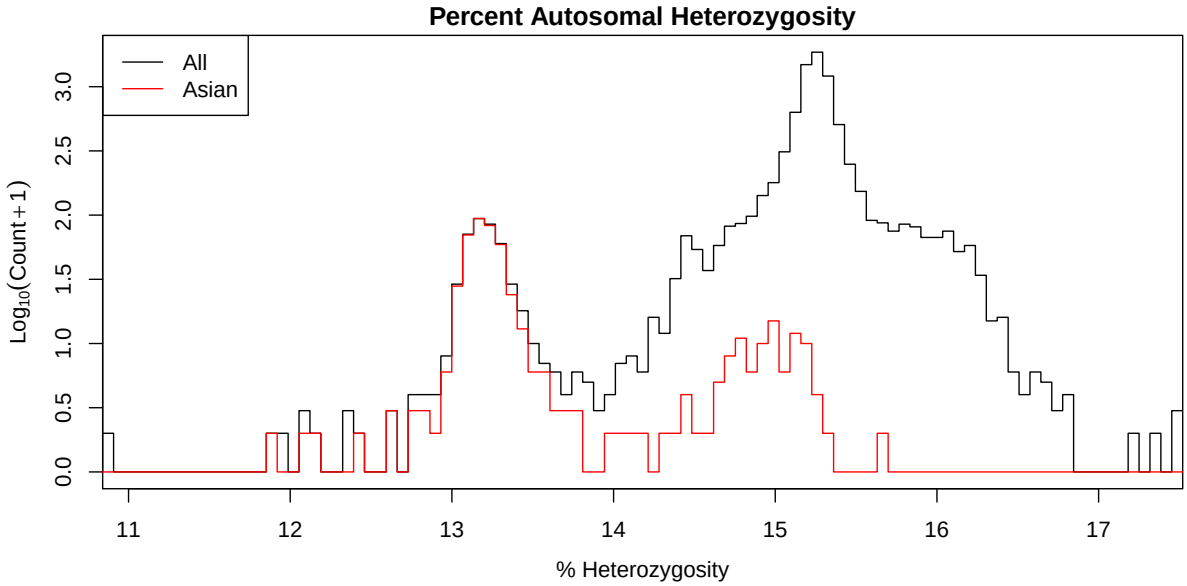


Figure 3.11 Distribution of percent autosomal heterozygosity for all samples and self-reported Asian samples.

4 samples > 17%. Two of the latter samples are from the same individual, which was duplicated on plates 19 and 89 with autosomal heterozygosity of 17.52% and 17.46%, respectively. The distribution of autosomal heterozygosity is bimodal with a small peak slightly above 13% and a larger peak around 15%. This bimodal heterozygosity is also observed in chromosome X (Fig. 3.10). Roughly 90% of samples with autosomal heterozygosity below 14% are Asian and ∼ 82% of the Asian samples have autosomal heterozygosity < 14%. The bimodal distribution of Asian samples may be due to sample being obtained from two distinct subpopulations from East and West Asia. No samples were marked for removal.

3.1.2.3.3 Sample Duplication

The experimental design of the plate layout resulted in 270 duplicate samples for computing concordance within and between plates. There were an additional 4 samples that were unintentionally duplicated. The sample with the highest average call rate was retained in each case, resulting in 274 samples being identified for removal.

3.1.2.4 Probes

Metrics can be used to identify any probes that have poor performance, which could be the result of issues such as the probe not been able to anneal to the target genetic location well or incorrect prediction from the genotype calling algorithm.

3.1.2.4.1 Concordance

Duplicate and HapMap samples were used to compute concordance values for each probe as shown in Fig. 3.12. Concordance was computed as n_{match}/n_{call}, where n_{call} is the total number of sample pairs where both samples had genotype calls and n_{match} is the number of sample pairs with matching calls. A no-call in either sample resulted in the sample pair being removed from the concordance computation. The concordance for duplicate samples were computed for the entire filtered probe

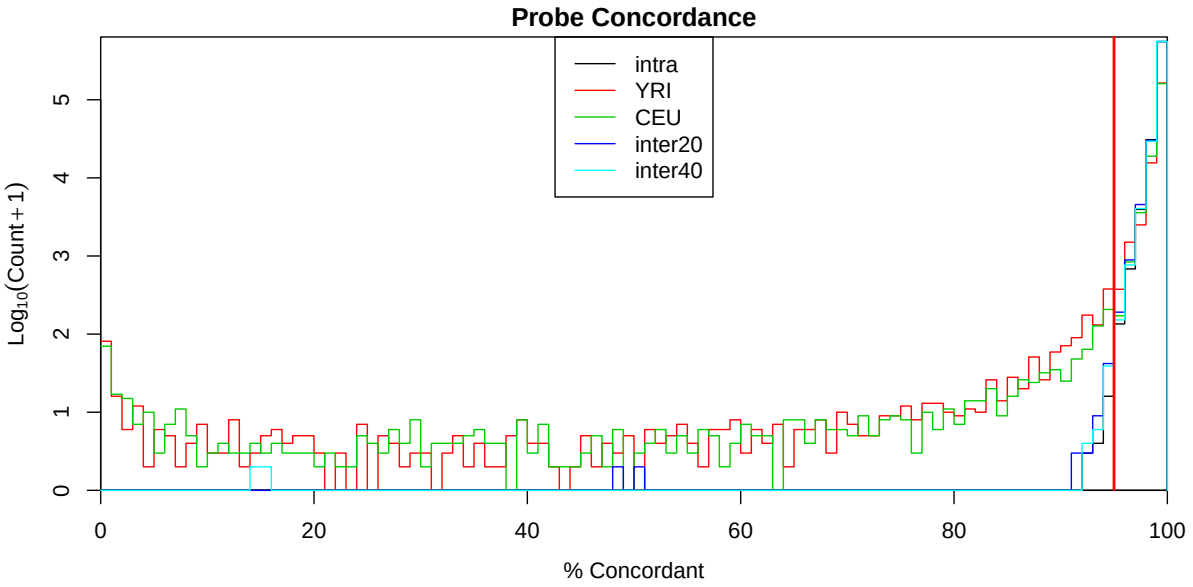


Figure 3.12 Probe concordance measurements for duplicate and HapMap samples. Duplicates samples were placed on the same plate (intra), 20 plates apart (inter20) and 40 plates apart (inter40). Any probe with concordance < 95% (red line) for intra, inter20 or inter40 sample pairs was removed from the dataset.

set while those for HapMap samples were computed for a subset of probes annotated with dbSNP rsIDs that matched those in the reference genotypes. Concordance of HapMap samples was slightly lower than the duplicate samples and this presumably reflects the difference in genotyping platforms. HapMap concordance was not used to remove any probes from the dataset. Probes were removed if any of the intra, inter20 or inter40 concordances were < 95%. The number of probes below this threshold were 20, 55 and 48 probes for intra, inter20 and inter40 duplicate samples, respectively, with slight overlap resulting in a set of 118 unique probes flagged for removal.

3.1.2.4.2 Minor Allele Frequencies

Minor allele frequencies of the filtered probe set were computed for all samples (excluding HapMap samples) and separately for each of the four main ethnic backgrounds and are recorded in Table 3.1. Each MAF was computed as $\min(p,q)$, where $p=(2n_{AA}+n_{AB})/2n$ and $q=(n_{AB}+2n_{BB})/2n$ are the

Table 3.1 Minor Allele Frequency Distributions for Filtered Probe Set

Variant Type	MAF Range (%)	All	White	Black	Asian	Hispanic
Homozygous	0	88587	136218	159939	238533	180808
Rare	(0,1]	163073	146556	83750	74682	67080
Low Frequency	(1,5]	61730	33056	77200	50788	71702
Common	(5,50]	278583	276143	271084	227970	272383

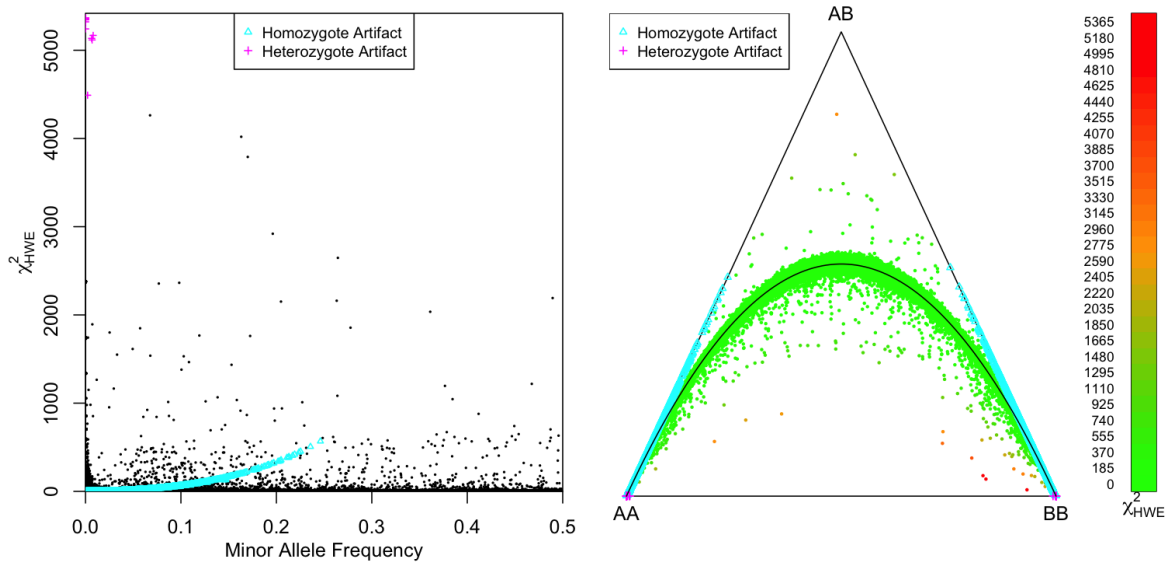
frequencies of alleles A and B, respectively, while n_g, $g \in \{AA, AB, BB\}$, is the number of genotype g calls and n is the total number of genotype calls. No probes were removed based on the MAF.

3.1.2.4.3 Hardy-Weinberg Equilibrium

A probe is in Hardy-Weinberg Equilibrium (HWE) if the observed genotype frequencies are the same as the expected frequencies computed using the observed allele frequencies. Deviations from HWE were tested using Pearson's chi-squared test:

$$\chi^2_{HWE} = \sum_{g \in \{AA,AB,BB\}} \frac{(n_g - E_g)^2}{E_g},$$

where n_g and E_g are the observed and expected number of samples with genotype g. The expected number of samples were computed as $E_{AA} = p^2 n$, $E_{AB} = 2pqn$ and $E_{BB} = q^2 n$ using the variable definitions in the MAF section above. Fig. 3.13a depicts the χ^2_{HWE} statistics as a function of MAF. Genotyping artifacts can be seen when the number of observed genotypes is $< 10\%$ of the expected. Deviations from HWE are graphically visualized using a ternary plot [Graffelman & Camarena, 2007] in Fig. 3.13b. Probes in HWE would be located along the arc between genotypes AA and BB. Deviation from HWE was not used to remove probes from the dataset. However, probes that did not follow HWE were flagged and did not contribute to the imputation process.



(a) **(b)**

Figure 3.13 Hardy-Weinberg Equilibrium visualization for autosomal chromosomes of white samples. (**a**) χ^2 values from asymptotic test plotted versus MAF. (**b**) Ternary plot color coded for χ^2 values, black arc represents ideal HWE curve. Artifacts indicate when the observed number of genotype calls is below 10% of the expected number.

3.1.2.4.4 Mendelian Trio Error

Concordance with Mendelian inheritance was computed using the 90 CEU samples, which consisted of 30 family trios. The trio containing the sample with low concordance on plate 58 (see Fig. 3.7) was removed before computing Mendelian trio error. A Mendelian error occurs when the child of the trio has a genotype incompatible with the parent genotypes. For example, if the mother is called as AA and the father called as BB, then the child being called as either AA or BB will result in an error. The distribution of number of errors is shown in Fig. 3.14. A total of 7380 probes with ≥ 2 Mendelian errors were flagged for removal.

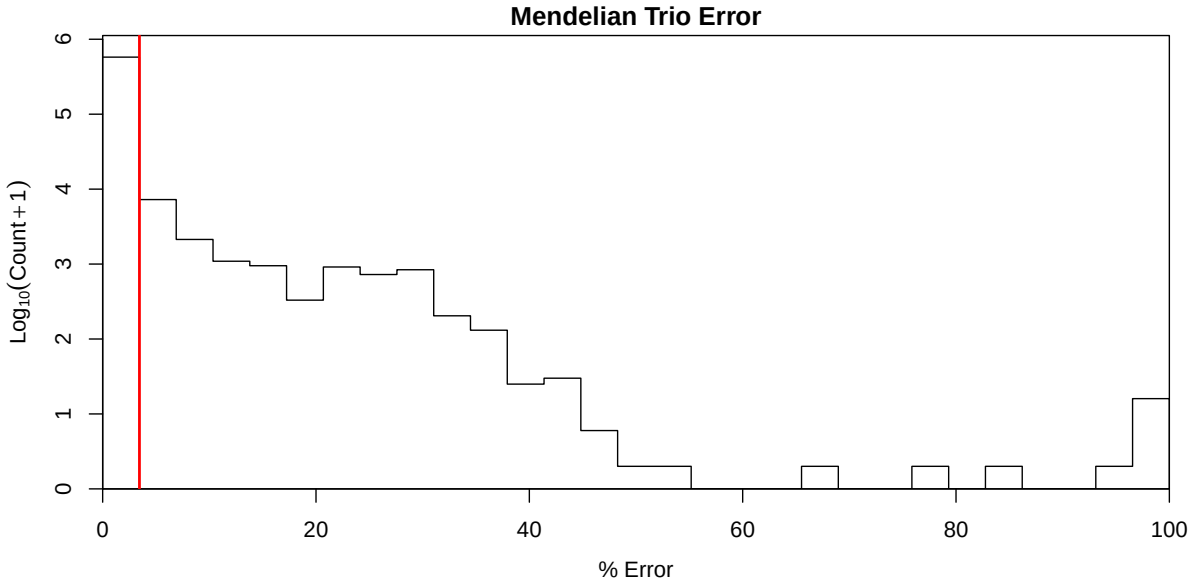


Figure 3.14 Distribution of Mendelian inheritance errors. Any probe with ≥ 2 errors (red line) was removed from the dataset.

3.1.2.4.5 Probe Duplication

There were a total of 1737 probes in the filtered probe set with non-unique chromosome locations consisting of 867 duplicate pairs and 1 triplicate set. The triplicate set and 583 of the duplicate pairs had identical alleles. The remaining 284 duplicate pairs have the same locations but different allele targets. In all cases, the probe with the highest call rate was retained, resulting the 869 probes being flagged for removal.

3.1.3 Summary

The 90 plates sent for processing consisted of 8628 samples and were genotyped using 628,679 probes. Pre-genotyping QC removed 3 samples for failure to produce CEL files, 12 samples for low signal quality and 87 samples for low QC call rates. Post-genotyping QC removed 18 samples for gender issues and 274 duplicate samples. Post-genotyping QC also removed 223 probes on the mitochondria or Y chromosome, 36,483 probes for failing either CR, FLD, HetSO or HomRO metrics, 7491 probes due to either poor concordance or high Mendelian trio error rates and 869 duplicate probes. The final dataset contains 8234 samples, of which 180 are HapMap samples, and 583,613 probes. The final genotype dataset was stored in PLINK binary format. Table 3.2 shows the samples sizes for each trial arm after the QC pipeline. The genotype quality was very good with only 1.5% of the ACCORD samples ($1-\frac{8234-180}{8628-274-180}$) and 7.1% ($1-\frac{583613}{628679-223}$) of the probes failing the QC filter metrics.

3.2 Imputation Pipeline

Data imputation was done using a two-step approach where the genotype calls were first pre-phased using SHAPEIT2 [Delaneau et al., 2012, 2013] (v2.r778) and then imputation was done using IMPUTE2 [Howie et al., 2009, 2011] (v2.3.0). Both steps used the 1000 Genomes Phase1 integrated haplotypes reference panel (release date Dec 2013) from the IMPUTE2 website. All steps were automated using custom shell scripts.

Table 3.2 Post-QC Sample Sizes

Glycemia Trial	Blood Pressure Trial		Lipid Trial		Total
	SBP < 120	SBP < 140	Fibrate	Placebo	
HbA_{1c} < 6.0%	571 179 60 **905** 42 53	533 217 59 **932** 51 72	743 137 83 **1108** 76 69	719 150 74 **1082** 70 69	2566 683 276 **4027** 239 263
HbA_{1c} 7.0%-7.9%	554 196 54 **916** 48 64	568 180 59 **912** 47 58	756 129 90 **1121** 73 73	720 151 74 **1078** 66 67	2598 656 277 **4027** 234 262
	1125 375 114 **1821** 90 117	1101 397 118 **1844** 98 130	1499 266 173 **2229** 149 142	1439 301 148 **2160** 136 136	
Total	2226 772 232 **3665** 188 247		2938 567 321 **4389** 285 278		5164 1339 553 **8054** 473 525

SBP = systolic blood pressure (mm Hg); HbA_{1c} = glycosylated hemoglobin; subgroup sample sizes ordered as white, black, Hispanic, Asian and other.

3.2.1 Probe exclusion

An additional QC step was applied for the genotype data going into the imputation pipeline. The imputation process makes use of assuming variants are in Hardy-Weinberg equilibrium (HWE) in order to convert predicted haplotypes into genotypes. Therefore, probes that do not follow the HWE assumption were excluded from the imputation pipeline. HWE chi-squared values were computed separately for self-reported white, black, Hispanic and Asian subgroups. Any probe with $\chi^2 > 19.51$ (p-value $< 10^{-5}$) in at least two of the four main ethnic subgroups was excluded, reducing the number of probes by 1145. HWE values of the cleaned probe set for white samples are shown in Fig. 3.15 and can be compared with those in Fig. 3.13.

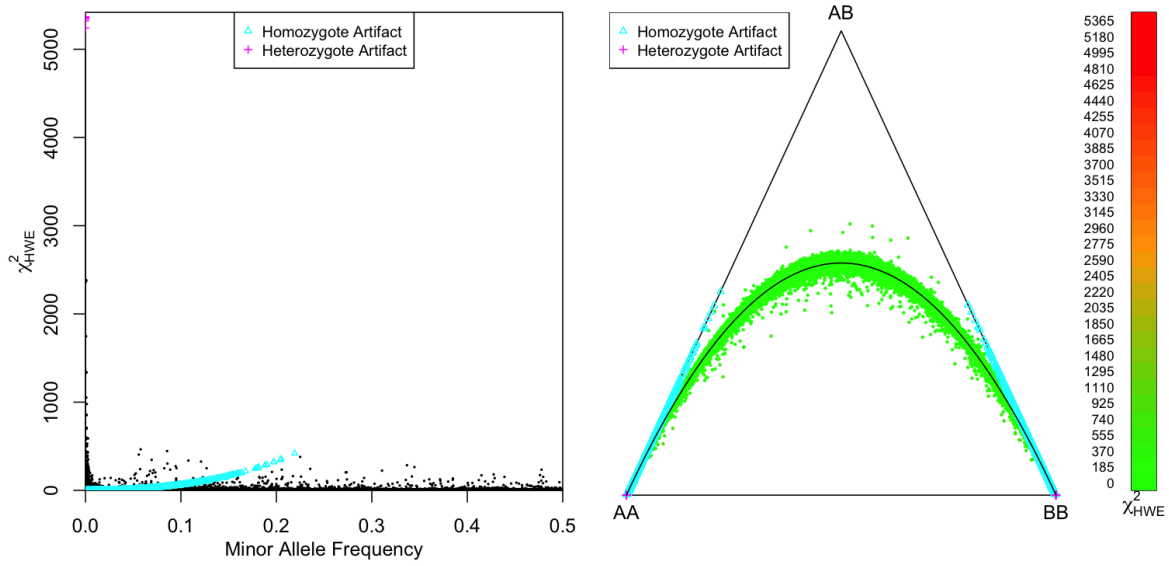


Figure 3.15 Hardy-Weinberg Equilibrium visualization for autosomal chromosomes of white samples for probes used in the imputation process. (**a**) χ^2 values from asymptotic test plotted versus MAF. (**b**) Ternary plot color coded for χ^2 values, black arc represents ideal HWE curve. Artifacts indicate when the observed number of genotype calls is below 10% of the expected number.

3.2.2 Pre–phasing

Prephasing genotype data reduces the computational cost of imputation. The benefits of prephasing are highly desirable when considering that datasets will need to be reimputed as larger haplotype reference panels become available and prephasing only needs to be carried out once. In the approach used by SHAPEIT2, known haplotypes are collapsed into a graph structure and represented in a compact hidden Markov model (HMM). An individual's genotype vector is split into disjoint segments and the unknown haplotypes in each segment are represented as nodes of a graph, where each path represents a compatible haplotype. Transition probabilities between segments are computed using a forward-backward algorithm on the HMM and plausible haplotypes are sampled for each genotype using a forward sampling procedure. This method scales linearly in the number of SNPs, samples and conditioning haplotypes.

Genotype data was prepared for prephasing by first splitting it into separate chromosome files using PLINK [*PLINK (v1.07)*; Purcell et al., 2007] (v1.07) after excluding probes failing the HWE filter. Any probe that was not in the reference panel as determined by chromosome location was excluded, n=143,331. In addition, probes that were homozygous or had alleles that were different between the genotype data and reference panel were excluded, n=16,110 and n=2,600, respectively. All samples were prephased together since this approach was previously found to improve phasing in a cohort with diverse ancestries compared to phasing groups separately [Delaneau et al., 2013].

3.2.3 Imputation

The purpose of data imputation is to increase both the power to detect associated genetic variants and the resolution of genetic association studies [Howie et al., 2011]. The strategy of IMPUTE2 is to phase haplotypes and impute unknown genotypes in alternating steps of a Markov chain Monte Carlo (MCMC) algorithm. In the first step, an individual's genotype is phased by sampling from the distribution of haplotype pairs conditioned on the observed genotypes, estimates of other individual's

haplotypes, the known reference panel haplotypes and a fine-scale recombination map, where the conditional distribution is represented using a HMM. In the second step, new alleles are imputed conditioned on the individual's sampled haplotypes, the known reference panel haplotypes and the recombination map. Marginal posterior probabilities for each allele are analytically determined independently and are converted into genotypes under the assumption of HWE. The final genotype posterior probabilities are the normalized sum across iterations. The computational burden of the first step grows quadratically with the number of haplotypes while the second step is linear. Prephasing the study genotypes eliminates the need for the first (and most expensive) step of the IMPUTE2 algorithm.

The imputation process began by splitting the prephased haplotype data into 5 Mbp non-overlapping segments for each chromosome. The same reference panel from the prephasing step was used during the imputation computations. Only those variant with an "info" metric > 0.5 were retained for association testing, resulting in a total of 26,330,649 imputed variants ($\sim 71.7\%$ of total imputed variants).

3.3 Rare Variant Dataset

One primary interest of the genetic association studies for the ACCORD analysis is the identification of functional variants. To this end, a rare variant dataset was created by constructing groups of functional variants for each gene. First, the posterior probabilities of imputed variants were converted into genotype calls using a threshold of 0.9. The call was considered to be missing if none of the genotype posterior probabilities exceeded the threshold value. Second, genotyped and imputed variants were combined and grouped based on location and annotated function. Variants were grouped into gene sets using the GRCh37.p13 assembly GFF3 files from the NCBI FTP site. If the gene entry contained no exons in the GFF3 file, then any variant between the gene start and stop locations were included in the gene group. If exon entries did exist, then any variant between the exon start-2 and stop+2 locations were included. Annotation information was obtained from the NCBI SNPContigLocusId file for dbSNP

build 141 mapped to genome assembly GRCh37.p13. Variants were retained if they were classified as one of the following: 1) stop-gain (nonsense), 2) missense, 3) stop-loss (readthrough), 4) frameshift, 5) cds-indel (indel with length being a multiple of 3 bp), 6) splice-3 (3' acceptor dinucleotide), 7) splice-5 (5' donor dinucleotide).

Chapter

4

ACCORD Analysis

This chapter describes the analysis pipeline that was established to accelerate future analyses using the ACCORD dataset. The pipeline was written to automate much of the standard procedures for common and rare variant analyses. The pipeline includes steps that extract phenotype data from the clinical trial records, detect and remove cryptically related individuals, compute principal components, adjust p-values for genomic inflation, conduct common and rare variant analyses, and produce a variety of diagnosis and result plots. The pipeline was tested by analyzing the baseline lipid levels of all ACCORD samples after the quality control process. Significant variants and genes were compared to those previously reported in the literature.

4.1 Analysis Pipeline

The ACCORD clinical trial has provided a rich data source that will be used in genetic association studies for many years to come. A streamlined analysis pipeline is desirable to accelerate the execution of analyses of new phenotypes and to ensure consistency of analysis quality. This has been accomplished by developing a semi-automated analysis pipeline, which has been used initially to study genetic association of baseline lipid levels. Investigating baseline lipid levels allows for replication of findings previously reported in the literature in order to validate both the quality control and analysis pipelines, while also contributing results from a novel dataset consisting exclusively of type 2 diabetic patients. An overview of the analysis pipeline is shown in Figure Fig. 4.1. Development of the analysis pipeline requires knowledge of advantages and disadvantages of many methods in order to appropriately select and properly apply them to this particular problem.

4.1.1 Dataset Generation

Each analysis begins by constructing a complete dataset. These datasets include the phenotypes of interest (e.g. total cholesterol, LDL, HDL and triglycerides), a list of covariates (e.g. age, gender) and an appropriate group of unrelated samples. The first step in the dataset generation process requires

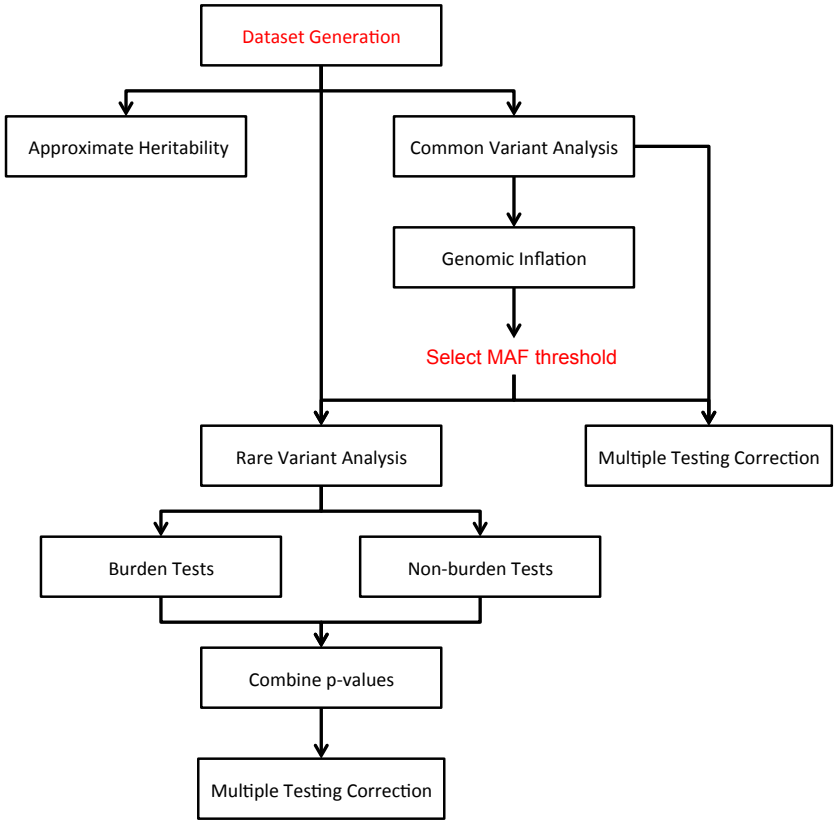


Figure 4.1 Analysis pipeline. Manual steps in red, automated in black.

input from three text files, which consist of 1) a list of phenotypes, 2) a list of all candidate covariates and 3) a list of forced covariates, which is a subset of the candidate covariate list. There is also an option to provide a list of sample IDs, which would be useful if a specific sample subset is of interest (e.g. only samples in the phenofibrate trial arm). A flowchart of the dataset generation process is shown in Figure Fig. 4.2.

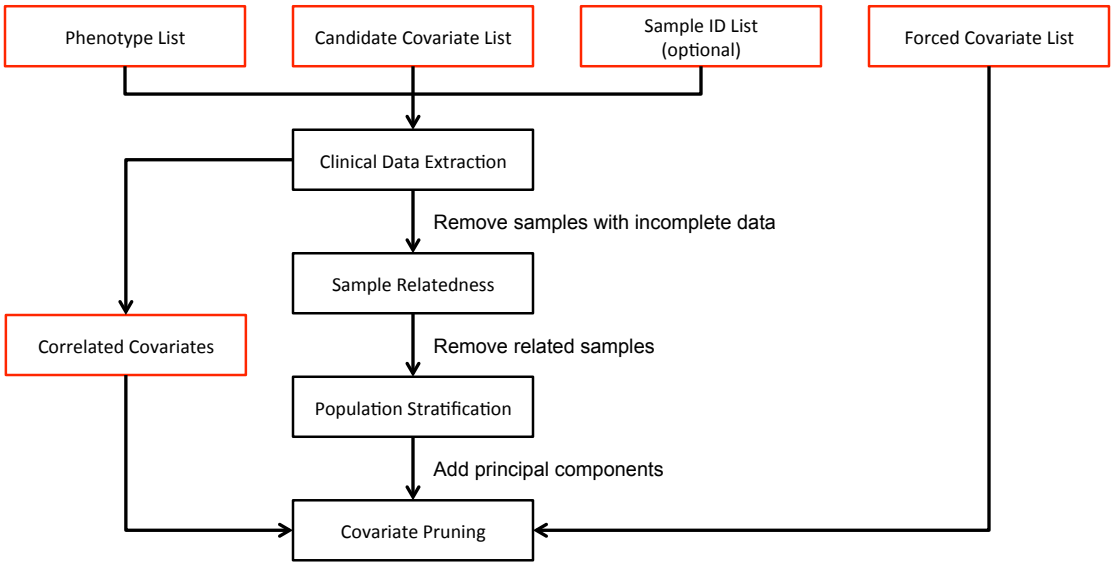


Figure 4.2 Dataset generation steps. Manual steps in red, automated in black.

4.1.1.1 Clinical Data Extraction

The analysis pipeline uses the input lists to extract clinical data from the appropriate clinical forms. The clinical data is fragmented among 34 files where certain fields may be reported in multiple locations, which are not always in agreement. Careful consideration of the accompanying clinical trial forms and notes is essential before adding new entries to the extraction script. Only those samples with complete clinical data for the phenotype and candidate covariate lists are retained. A correlation matrix is computed for all covariates and any pair with a correlation absolute value > 0.5 is reported.

The highly correlated pairs are then use to manually create a list of variables to be remove from the dataset to prevent issues due to collinearity during regression analysis.

4.1.1.2 Sample Relatedness

The samples selected at the beginning of the analysis pipeline need to be checked for cryptic relatedness because the downstream analysis methods assume independence of each sample. While the ACCORD trial did not knowingly collect related individuals, the genome-wide data allows us to detect cryptic (unknown) relatedness and to remove samples if the independence assumption does not hold. It is known that the dependence created by cryptic relatedness can cause an inflated Type I error in association analysis [Voight & Pritchard, 2005]. Related individuals are removed during the analysis pipeline rather than in the quality control pipeline to maximize sample sizes when testing associations for sample subsets. For example, if there are several pairs of related individuals with one individual from each pair in the blood pressure trial and the other in the lipid trial, then it would be unnecessary to have removed any of these individuals when conducting association tests limited to one of those trial arms.

Methods that predict relationship inference assuming a homogenous population are based on estimating relatedness as a function of allele frequencies, which are in turn estimated from the genotype frequencies of the entire sample under the assumption of Hardy-Weinberg equilibrium (HWE). However, allele frequencies are not necessarily the same across different population subgroups and this approach can lead to inflated degrees of relationship between pairs of individuals both within and between population subgroups [Manichaikul et al., 2010].

This problem is avoided by using KING (v1.4) [Manichaikul et al., 2010] to estimate kinship coefficients for each pair of samples, which is based on a method that is robust to heterogeneity of sample ethnicity. KING can be directly applied to the ACCORD dataset, while other methods, such as that used by PLINK, would require an initial step of grouping samples in more homogenous subpopulations before estimating cryptic relatedness between sample pairs in each group. The estimated

kinship coefficients reflect the probability that alleles sampled at random from two individuals are identical by descent. In this approach, allele frequencies are treated as random variables and kinship coefficients are estimated as function of genetic distance, represented in terms of shared genotype counts, and genome-wide average heterozygosity for the individuals in each pair. This method assumes HWE among SNPs with the same underlying allele frequencies and kinship coefficients can be overestimated if HWE is violated in the direction of too little homozygosity. This potential estimation inflation is guarded against by using the smaller of the observed heterozygosity rates between the two individuals.

Relationship information is visualized by plotting the estimated kinship coefficients vs. proportion of alleles with zero identity by state. One individual in each pair with an estimated kinship greater than the separation between full and half siblings (estimated kinship $> 1/2^{5/2} = 0.1768$) is randomly removed.

4.1.1.3 Population Stratification

Allele frequency differences due to population stratification can result in frequent false positive results or reduced power in genetic studies [Tian et al., 2008]. One well established way to compensate for these differences is to include principal components (PCs) as covariates when testing for associations. Principal component analysis reduces genotype data into a small number of orthogonal dimensions that describe as much variability as possible. The axes of variation can often be interpreted as describing differences in ethnic background or geographical location, although linear combinations of the top PCs may not be sufficient to capture this information [Mathieson & McVean, 2012]. EIGENSTRAT (v4.2) [Price et al., 2006] is used to compute the first ten PCs from a subset of genotyped variants. This subset consists of the variants that remain after filtering based on minor allele frequency (MAF) and linkage disequilibrium (LD) pruning using a variance inflation factor (VIF) threshold. The VIF threshold corresponds to $1/(1-R^2)$ where R^2 is the multiple correlation coefficient for a variant being regressed on all other variants simultaneously within a sliding window. PLINK is used to create a

variant set with MAF > 0.01 and VIF < 1.5. The top ten PCs computed from the subset of variants are automatically added to the candidate covariate list.

4.1.1.4 Covariate Pruning

Covariate selection is an important component of a robust analysis plan. Including confounding variables can reduce the heterogeneity in the data that can result in more power, but over-parameterization of a model can result in reduced power [Mefford & Witte, 2012]. We take a combined approach to variable selection to address these concerns. Some covariates can be forced into the model based on the results of previous studies or on expert knowledge related to the phenotype, while variable selection is performed on candidate variables to identify covariates specific to the ACCORD dataset.

Backwards selection is used to prune the list of candidate covariates using Bayes Information Criterion (BIC) and is applied separately to each phenotype. The pruning algorithm begins by normalizing all continuous variables and splitting the covariate list into two groups: 1) the covariates forced into the model, $\mathbf{X}_{\text{forced}}$, and those remaining after removing the forced and selected correlated covariates from the candidate list, $\mathbf{X}_{\text{remaining}}$, which consists of $p_{\text{remaining}}$ covariates. Pseudocode is given in Algorithm 1, where M represent the model, $\mathbf{y}$ is the phenotype, $\boldsymbol{\alpha}$ are the regression parameters, $\boldsymbol{\epsilon}$ is the error, $L(M)$ is the likelihood of the model, p is the total number of regression parameters, n is the number of samples and $\mathbf{X}^{(i)}_{\text{remaining}}$ and $\boldsymbol{\alpha}^{(i)}_{\text{remaining}}$ are the remaining covariates and regression coefficients after removing covariate i.

4.1.2 Heritability Approximation

A rough estimate of a phenotype's heritability provides information that can be used to determine if the results from association tests seem plausible. This is useful when determining if it is worthwhile to investigate the potential genetic associations with a new phenotype that hasn't been studied before. If the approximate heritability is very low it may be the case that the trait is not even mappable. The

Algorithm 1 Covariate pruning via backwards selection based on BIC

loop
  Fit $M:\ \mathbf{y} = \alpha_0 + \mathbf{X}_{\text{forced}}\boldsymbol{\alpha}_{\text{forced}} + \mathbf{X}_{\text{remaining}}\boldsymbol{\alpha}_{\text{remaining}} + \boldsymbol{\epsilon}$
  $\text{BIC} = -2\ln L(M) + p\ln n$
  for $i = 1, \ldots, p_{\text{remaining}}$ **do**
    Fit $M_i:\ \mathbf{y} = \alpha_0 + \mathbf{X}_{\text{forced}}\boldsymbol{\alpha}_{\text{forced}} + \mathbf{X}^{(i)}_{\text{remaining}}\boldsymbol{\alpha}^{(i)}_{\text{remaining}} + \boldsymbol{\epsilon}$
    $\text{BIC}_i = -2\ln L(M_i) + p\ln n$
  end for
  Let j be the index of $\min \text{BIC}_i$
  if $\text{BIC}_j < \text{BIC}$ **then**
    $\mathbf{X}_{\text{remaining}} = \mathbf{X}^{(j)}_{\text{remaining}}$
    if $p_{\text{remaining}} == 0$ **then**
      return M_j; **exit loop**
    end if
  else
    return M; **exit loop**
  end if
end loop

inability to map a trait could occur if causal loci were not genotyped, the trait is complex and involves rare variants requiring a larger sample size to detect, or environmental factors are largely responsible for phenotype variation. The phenotypic variation explained by genome-wide variants is estimated using the software tool genome-wide complex trait analysis (GCTA; v1.22) [Yang et al., 2010, 2011]. The approach of GCTA is to fit the effects of all genetic variants as random effects by a linear mixed model,

$$\mathbf{y} = \alpha_0 + \mathbf{X}\boldsymbol{\alpha} + \mathbf{W}\mathbf{u} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is the phenotype, α_0 is the intercept, $\mathbf{X}$ are covariates, $\boldsymbol{\alpha}$ are fixed effects, $\mathbf{W}$ is a standardized genotype matrix, $\mathbf{u}$ are random effects with $\mathbf{u} \sim \mathcal{N}(0, \mathbb{I}\sigma^2_u)$, $\mathbb{I}$ is an identity matrix and $\boldsymbol{\epsilon}$ are the residual effects with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbb{I}\sigma^2_\epsilon)$. The ij^{th} element of $\mathbf{W}$ is $w_{ij} = (x_{ij} - 2p_j)/\sqrt{2p_j(1-p_j)}$, where x_{ij} is the number of copies of the reference allele for the j^{th} variant of the i^{th} individual and p_j is the frequency of the reference allele.

The phenotypic variance is then $\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{W}\mathbf{W}'\sigma_u^2 + \mathbb{I}\sigma_\epsilon^2$. By defining the genetic relationship matrix (GRM) between individuals as $\mathbf{A} = \mathbf{W}\mathbf{W}'/N$, where N is the total number of variants, the phenotypic variance is then $\mathbf{V} = \mathbf{A}\sigma_g^2 + \mathbb{I}\sigma_\epsilon^2$, where σ_g^2 is the variance explained by all variants and is estimated using a restricted maximum likelihood approach. This results in an estimate of variance explained by genome-wide variants rather than trait heritability, which is the variance explained by all causal variants. The total genetic variation is a biased estimate of heritability due to incomplete LD between causal and genotyped variants and has a prediction error of $c + 1/N$, where c depends on the distribution of MAF of causal variants. An unbiased estimate is obtained by adjusting the genomic relationship between individuals i and i' as

$$A^*_{ii'} = \begin{cases} 1 + \gamma(A_{ii} - 1), & i = i' \\ \gamma A_{ii'}, & i \neq i' \end{cases}$$

where $\gamma = 1 - (c + 1/N)/\text{var}(A_{ii'})$.

4.1.3 Common Variant Analysis

Association between a phenotype and single common variant is tested using the linear regression model

$$\mathbf{y} = \alpha_0 + \mathbf{X}\boldsymbol{\alpha} + \beta_g \mathbf{g} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is the phenotype, α_0 is the intercept, $\mathbf{X}$ are the covariates, $\boldsymbol{\alpha}$ are the covariate regression parameters, β_g is the regression parameter for the variant, $\mathbf{g}$ is the additively coded genotype and $\boldsymbol{\epsilon}$ is the error term. Genotyped variants are tested using PLINK, where $g_i \in \{0, 1, 2\}$ is the number of minor alleles for the i^{th} individual. Imputed variants are tested in a custom R script, where $g_i = p_i(Aa) + 2p_i(aa)$ is the dosage score computed from the posterior probabilities for genotypes Aa and aa. The dosage score was tested for association because using a threshold or best-guess genotype can lead to both false positives and loss of power [Marchini & Howie, 2010]. Association p-values are

computed under the null hypothesis $H_0: \beta_g = 0$.

4.1.3.1 Genomic Inflation

Genomic inflation can be used to accommodate for the overdispersion of test statistics due to population heterogeneity and cryptic relatedness [Devlin & Roeder, 1999]. Earlier in the pipeline, PCs were added to the statistical model to account for these issues. However, inflation of test statistics has been shown to be different for rare and common variants with rare variants showing greater inflation for sharply defined nongenetic risk in spatially structured populations [Mathieson & McVean, 2012]. To adjust for the inflation of rare variants, genomic inflation values are used as a guide in selecting a MAF threshold for common variant analysis. The concept behind genomic inflation is that a set of independent test statistics, t_j, from non-causal loci will have a χ^2_1 distribution when no population structure is present. If the effects of heterogeneity and cryptic relatedness are constant across the genome, then the set of test statistics will follow a $\lambda\chi^2_1$ distribution, where λ is the variance inflation factor. A robust estimate of the variance inflation factor is $\widehat{\lambda} = \text{median}(t_j)/F(0.5)$, where $F(0.5)$ is the inverse cumulative distribution function of χ^2_1 evaluated at 0.5.

A MAF threshold to separate common and rare variant analyses is determined based on estimated variance inflation factors. Genotyped probes with MAF > 0.0001 are pruned to a subset with LD VIF $<$ 1.5 using PLINK. Estimates of λ are computed for various groups of probes based on MAF binning and/or cutoffs to determine an appropriate MAF threshold. A final variance inflation factor, $\widehat{\lambda}_{\text{gc}}$, is computed for the variants in the pruned probe set with MAF above the chosen threshold. Any variant, both genotyped and imputed, with a MAF above the chosen threshold is included in the common variant analysis and has its test statistics adjusted as $t_{\text{adj}} = t/\widehat{\lambda}_{\text{gc}}$.

4.1.3.2 Multiple-Hypothesis Testing Control

The Bonferroni procedure [Bonferroni, 1936] is used to control the family-wise error rate (FWER) for the common variant analysis. Given a set of m hypotheses, the FWER is the probability that at least

one hypothesis is falsely rejected, resulting in at least one type 1 error. If a single-test significance level is set as α, then the FWER for independent tests is $< 1-(1-\alpha)^m$, where the bound quickly approaches 1 as m increases. Controlling the FWER to be $< \alpha$ can be done using the Bonferroni procedure, where a hypothesis will be rejected if the p-value is below α/m. This is a conservative controlling procedure that does not require the tests to be independent and is simple to implement.

4.1.4 Rare Variant Analysis

Missing heritability unaccounted for by traditional genome wide association studies has motivated the investigation of rare variants [Manolio et al., 2009]. A traditional single-locus analysis approach applied to rare variants suffers from low power [Gorlov et al., 2008] and new methods have been developed to address this problem. There now exists a wide range of rare variant tests and they have been summarized in various ways [Asimit & Zeggini, 2010; Bansal et al., 2010; Dering et al., 2011; Lee et al., 2014b], yet power analyses have indicated no one method is best. Therefore, the approach used in this analysis pipeline is to apply a suite of tests that are representative of some of the more popular methods in the literature and cover both burden and non-burden approaches.

4.1.4.1 Burden Tests

The basic idea behind burden tests is to collapse a set of rare variants from a region of interest (e.g. a gene) into a single variable, which is then tested for association with a phenotype. Simple collapsing methods use indicator, proportion and weighted approaches chosen *a priori* that can easily be computed from genotype data, however, they are unable to take into account the possible direction of a rare variant effect. More sophisticated approaches can incorporate variant-specific information and variable thresholds for collapsing. Many burden tests were originally developed for case-control data, but the basic concepts from each method can be readily adapted to a regression framework.

Three burden tests are used in the rare variant analysis pipeline and are based on simple collapsing approaches. These methods work by first creating a collapse score, **c**, and then testing for association

using the linear regression model

$$\mathbf{y} = \alpha_0 + \mathbf{X}\boldsymbol{\alpha} + \beta_c \mathbf{c} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is the phenotype, α_0 is the intercept, $\mathbf{X}$ are covariates, $\boldsymbol{\alpha}$ are the covariate regression parameters, β_c is the regression parameter for the collapse score and $\boldsymbol{\epsilon}$ is the error term. Association p-values are computed under the null hypothesis $H_0 : \beta_c = 0$. The selected burden tests differ in how $\mathbf{c}$ is computed.

The first two tests are based on RVT1 and RVT2 originally proposed by Morris and Zeggini [Morris & Zeggini, 2010]. In RVT1, $c_i = r_i/n_i$, where n_i is the number of rare variants successfully genotyped for the i^{th} individual and r_i are the number of variants at which that person carries at least one copy of the minor allele. In RVT2, c_i is simply an indicator function $c_i = I(r_i > 0)$. RVT2 is implemented in the analysis pipeline as described, but a slight change to RVT1 is adapted where $c_i = \left(\sum_j g_{ij}\right)/2n_i$, which counts the total number of rare alleles rather than the number of variants with rare alleles. Similar collapsing methods are utilized in case-control methods such as the cohort allelic sums test (CAST) [Morgenthaler & Thilly, 2007] and the combined multivariate and collapsing (CMC) method [Li & Leal, 2008].

The third test uses a weighting scheme similar to the one proposed in the weighted sum statistic (WSS) case-control method [Madsen & Browning, 2009]. This method reflects the idea that rarer variants will have larger effect sizes and are weighted accordingly. The collapse score is computed as $c_i = \sum_j g_{ij}/w_j$, where $w_j = \sqrt{p_j(1-p_j)}$ is the estimated standard deviation of the minor allele frequency of variant j.

4.1.4.2 Non-Burden Tests

A major limitation of simple burden tests is that they cannot account for the possible direction (positive or negative association) of a rare variant effect [Wu et al., 2011]. One very popular non-burden rare variant test that allows for different directions and magnitudes of effects for each variant

is the sequence kernel association test (SKAT) [Wu et al., 2011]. While burden tests use simple linear regression, SKAT uses the linear mixed model

$$\mathbf{y} = \alpha_0 + \mathbf{X}\boldsymbol{\alpha} + \mathbf{G}\mathbf{u} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is the phenotype, α_0 is the intercept, $\mathbf{X}$ are covariates, $\boldsymbol{\alpha}$ are fixed effects of the covariates, $\mathbf{G}$ is the genotype matrix, u_j are random effects following an arbitrary distribution with mean 0 and variance $w_j\tau$ with variance component τ and prespecified weight w_j for the j^{th} variant, and $\boldsymbol{\epsilon}$ is an error term with a mean of zero and a variance of σ^2. Association p-values are computed under the null hypothesis $H_0 : \tau = 0$, which is tested with a variance-component score test.

The variance-component score statistic is

$$Q = (\mathbf{y} - \widehat{\boldsymbol{\mu}})'\mathbf{K}(\mathbf{y} - \widehat{\boldsymbol{\mu}}),$$

where $\widehat{\boldsymbol{\mu}} = \widehat{\alpha}_0 + \mathbf{X}\widehat{\boldsymbol{\alpha}}$ with $\widehat{\alpha}_0$ and $\widehat{\boldsymbol{\alpha}}$ estimated under the null and $\mathbf{K}$ is a kernel function. Various different kernel functions can be used to model certain genetic characteristics (e.g. epistatic effects), but the basic kernel function $\mathbf{K} = \mathbf{G}\mathbf{W}\mathbf{G}' = \sum_j w_j G_{ij} G_{i'j}$, where $\mathbf{W} = \text{diag}(w_1, \ldots, w_J)$, reflects the assumption that the trait depends on the variants in a linear fashion. Under the null hypothesis Q follows a mixture of chi-squared distributions and p-values are computed using the Davies exact method [Davies, 1980].

While SKAT has been found to be more powerful than burden tests when the variants have different directions of effect, it is less powerful when all variant effects are in the same direction. The balance between SKAT and burden tests was addressed by the optimal test, SKAT-O, where a combination of the two approaches is optimized [Lee et al., 2012]. In SKAT it is assumed that the u_js are independent, but SKAT-O allows $\mathbf{u}$ follow a multivariate distribution with a correlation structure $\mathbf{R}_\rho = (1-\rho)\mathbb{I} + \rho\mathbb{1}\mathbb{1}'$, where $\mathbb{I}$ is an identity matrix and $\mathbb{1}$ is a vector of ones. The information from both SKAT and a weighted burden test can be captured using a new kernel function $\mathbf{K} = \mathbf{G}\mathbf{W}^{1/2}\mathbf{R}_\rho\mathbf{W}^{1/2}\mathbf{G}'$.

The new variance-component score can be seen as the linear combination of SKAT and a weighted burden test $Q_\rho = (1-\rho)Q + \rho Q_{\text{burden}}$. The SKAT-O test statistic is determined by scanning various values of ρ using a simple grid search, computing Q_ρ at each point, and then choosing the minimum p-value from the set as the statistic. The null distribution of the statistic can be approximated by a weighted sum of independent χ^2_1 random variables and the association test p-value is computed numerically.

The analysis pipeline uses both SKAT and SKAT-O in addition to the three burden tests mentioned in the previous section. Only those variants with MAF above the cutoff determined using genome inflation are included the in the rare variant approaches, therefore preventing any contribution from common variants. The non-burden approaches were implemented using the R package "SKAT" (v0.95) [Lee et al., 2014a].

4.1.4.3 Multiple-Hypothesis Testing Control

The use of multiple rare variant tests for each gene compounds the problem of multiple-hypothesis testing. Two questions arise when considering how to correct for multiple testing in this situation: 1) how should the type 1 error be controlled and 2) how can the desired control scheme be applied to the set of p-values.

Controlling the FWER for rare variant testing may be too conservative of an approach because a substantial proportion of genes may be associated with a specific phenotype, especially if the phenotype is highly heritable. Instead, controlling the false discovery rate (FDR), which is the expected proportion of type 1 errors among the rejected hypotheses, is more desirable. Controlling the FDR rather than the FWER was first popularized by Benjamini and Hochberg [Benjamini & Hochberg, 1995], who argued that the FWER is too conservative in situations where rejecting a large number of hypotheses is desirable. Later, Storey proposed a method for computing q-values, which is the FDR analog of the p-value, for each hypothesis [Storey, 2002].

Application of an FDR controlling procedure is not straightforward due to each gene having a set

of p-values, one for each rare variant test. This complication can be resolved by combining the set of p-values into a single p-value for each gene. There are several methods for combining p-values, however, most methods assume the p-values are independent. Intuitively the p-values from two rare variant tests would not be completely indepenent, e.g. if a gene is found to be significant in one test, then it may be more likely to be significant in one of the other tests as well. The dependence of p-values is accounted for by using the "correlated Lancaster procedure" described by Dai et al. to combine the p-values from each test into a single p-value for each gene [Dai et al., 2014]. This procedure is based on Lancaster's method [Lancaster, 1961] but corrects for dependence between tests using the correlation structure.

Application of the suite of rare variant tests results in a set of p-values, p_{jk}, corresponding to the j^{th} gene and k^{th} rare variant test. Following the correlated Lancaster procedure, these p-values are converted to chi-squared values $t_{jk} = \gamma^{-1}_{(w_k/2,2)}(1-p_{jk}) \sim \chi^2_{w_k}$, where w_k is the relative weight of the k^{th} test and $\gamma^{-1}_{(w_k/2,2)}$ is the inverse cumulative distribution function of the Gamma distribution with shape parameter $w_k/2$ and scale parameter 2. The chi-squared values are combined to create a set of Lancaster test statistics $T_j = \sum_k t_{jk}$. Independent p-values would result in T_j having chi-squared distributions with $\sum_k w_k$ degrees of freedom. Dependent p-values are adjusted for by using a Satterthwaite approximation for a scaled chi-squared distribution $cT_j \sim \chi^2_v$, where $v = 2[E(T_j)]^2/\text{var}(T_j)$ and $c = v/E(T_j)$. Here $E[T_j] = \sum_k w_k$ and $\text{var}(T_j) = 2\sum_k w_k + 2\sum_{k<k'} \rho_{kk'}$, where $\rho_{kk'} = \text{cov}(t_{jk}, t_{jk'})$ is used to adjust for correlation among p-values. The scaled Lancaster test statistics, cT_j, are converted to p-values, p_j, using the χ^2_v distribution.

Combining the set of p-values for each rare variant test into a single p-value for each gene allows straightforward application of FDR controlling procedures. Computing q-values was done using the R package "qvalue" (v1.36.0) [*qvalue: Q-value estimation for false discovery rate control*]. To begin, the proportion of truly null hypothesis is estimated from the distribution of p-values. A step-up procedure is then employed to estimate q-values.

4.2 Baseline Lipid Results

An initial application of the analysis pipeline tested for genetic variants associated with baseline lipid levels. These phenotypes have been heavily studied and replication of previous findings can be used as a measure of the validity of the quality control and analysis pipelines.

4.2.1 Model Setup

The phenotype for this application of the analysis pipeline were total cholesterol (TC, mg/dL), low density lipoprotein (LDL, mg/dL), high density lipoprotein (HDL, mg/dL) and triglycerides (TG, mg/dL). Raw data was extracted from the clinical forms and then log transformed as shown in Fig. 4.3.

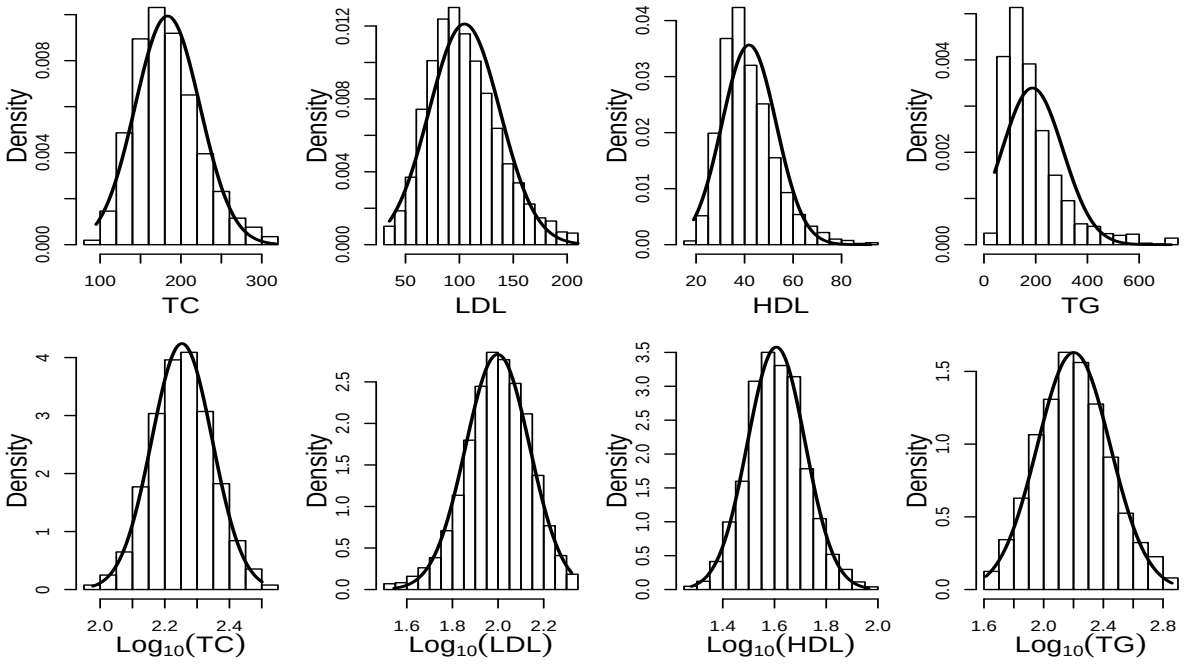


Figure 4.3 Histograms of raw and $\log_{10}$ transformed baseline lipid levels. Normal distributions computed using the baseline means and standard deviations are overlaid.

Lists of forced and candidate covariates were created with the aid from a panel of experts. Covariate names and descriptions can be found in Table 4.1 and Table 4.2. Yrsdiab and yrslipi were mean-imputed due to large numbers of patients missing those records. Only those samples with complete phenotype and covariate data were retained, reducing the sample size from 8054 to 7929. A correlation matrix for all covariates is shown in Fig. 4.4. Three pairs of covariates were flagged for high correlation:

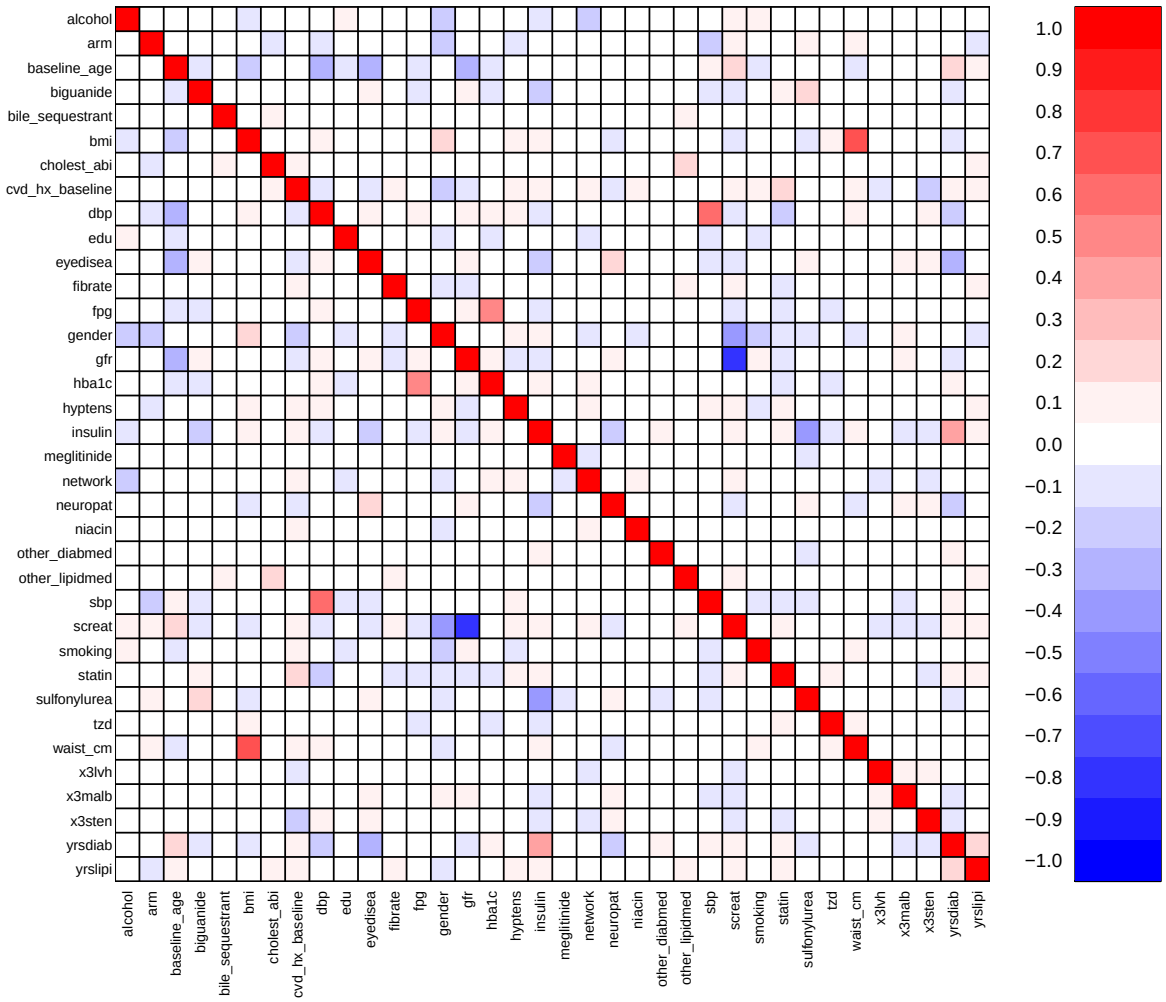


Figure 4.4 Covariate correlation matrix. Screat, dbp and waist_cm were removed from the lists of covariates due to correlations with gfr, sbp and bmi, respectively.

glomerular filtration rate (gfr) and serum creatinine (screat) were negatively correlated with $r = -0.77$, systolic and diastolic blood pressure (sbp and dbp, respectively) were positively correlated with $r = 0.53$ and body mass index (bmi) and waist circumference (waist_cm) were positively correlated $r = 0.63$. Screat, dbp and waist_cm were excluded from the statistical models.

The estimated relationship between pairs of individuals is shown in Fig. 4.5. The horizontal dashed lines divide kinship values into ranges corresponding to the degree of relationship. Duplicate samples or monozygote twins would have a kinship of $\frac{1}{2}^1 = 0.5$, first degree relationships would have a kinship of $\frac{1}{2}^2 = 0.25$, etc. The horizontal lines separate these groups, e.g. the top dashed line is between zero and first degree relationships at $\frac{1}{2}^{3/2} = 0.3536$. One individual from a pair with kinship between first and second degree ($\frac{1}{2}^{5/2} = 0.1768$) was removed, reducing the sample size by 85 to 7844.

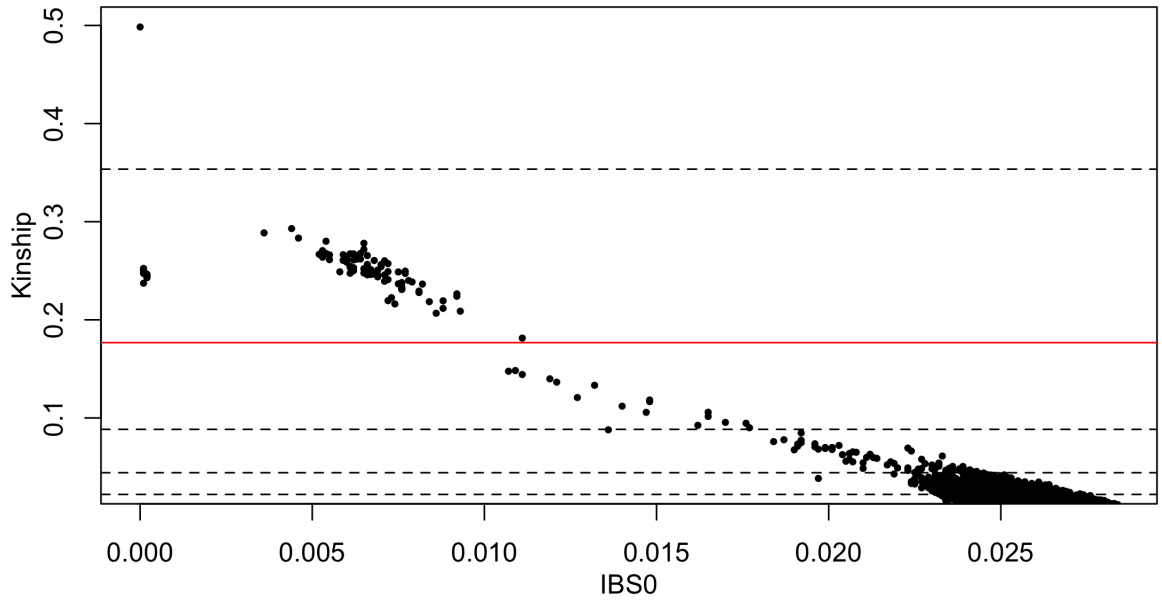


Figure 4.5 Estimated kinship coefficients for all pairs of samples. Dashed horizontal lines depict separation between degrees of relatedness. The red line separates first and second degree relationships, which was the threshold used to remove individuals from the analysis.

Principal components (PCs) were computed for the remaining 7844 samples. The first two PCs are shown in Fig. 4.6 where the marker colors and labels represent the self-reported ethnic background for each sample. The top 3 PCs were added to the forced covariate list while PCs 4-10 were added to the candidate covariate list.

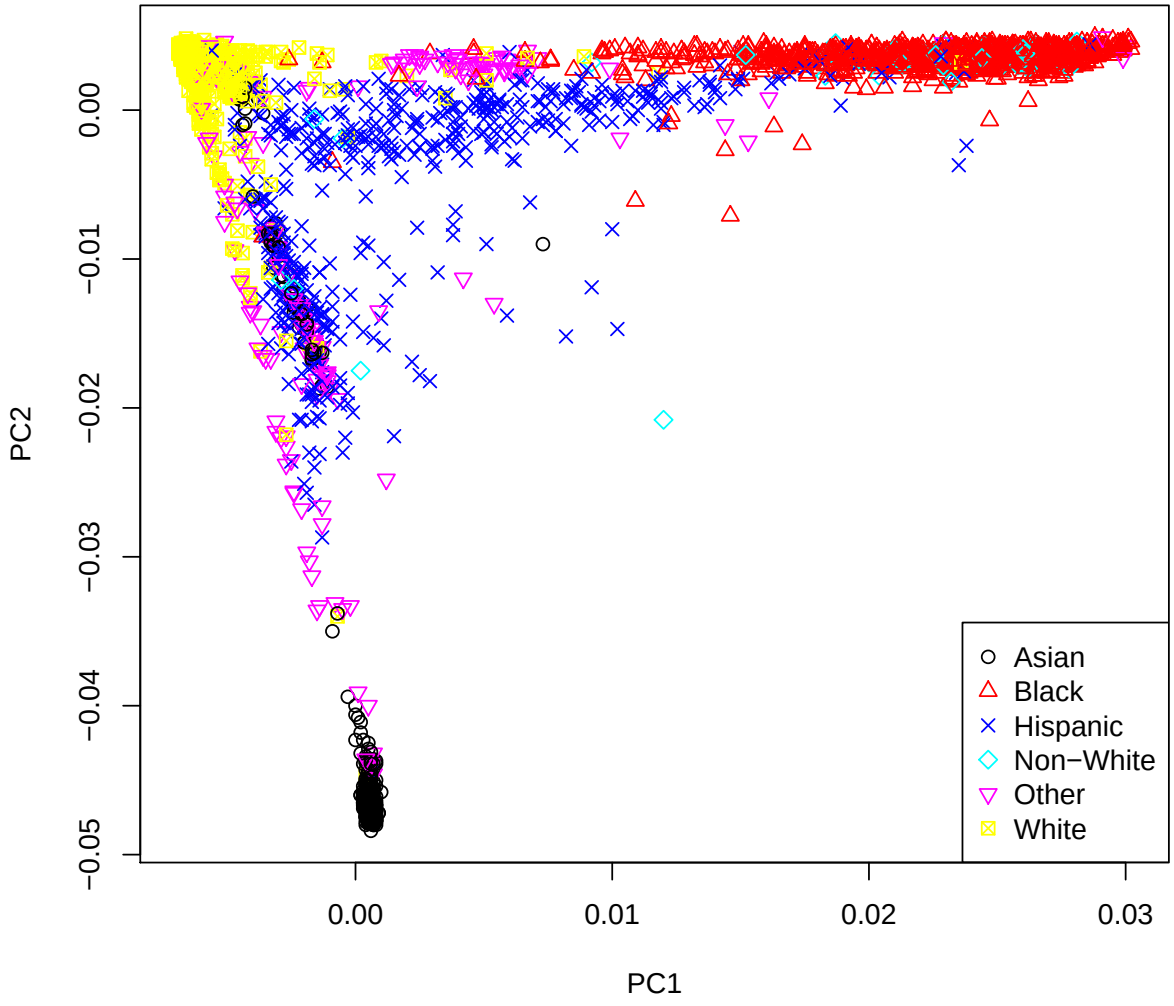


Figure 4.6 First two principal components with markers indicating self-reported ethnic backgrounds.

Table 4.1 Baseline Lipid Model Forced Covariates

Variable name	Phenotype [a] TC	LDL	HDL	TG	Description
arm2	**	**			Randomization arm: 1-8 (1 - Std. Glycemia/Int. BP) 2 - Std. Glycemia/Std. BP
arm3	·	·			3 - Int. Glycemia/Int. BP
arm4	·	*	*		4 - Int. Glycemia/Std. BP
arm5	***	***	***	**	5 - Std. Glycemia/Lipid Fibrate
arm6	***	***	***	*	6 - Std. Glycemia/Lipid Placebo
arm7	***	***	***	·	7 - Int. Glycemia/Lipid Fibrate
arm8	***	***	***		8 - Int. Glycemia/Lipid Placebo
baseline_age	***	***	***	***	Age at baseline (years)
biguanide1	***	***		**	Use of biguanides at baseline: 0=No, 1=Yes
bile_sequestrant1				*	Use of Bile-acid sequestrants at baseline: 0=No, 1=Yes
bmi		*	***	***	Body mass index: weight(kg)/(height(m))^2
cholest_abi1	**	**			Use of cholesterol absorption inhibitors at baseline: 0=No, 1=Yes
cvd_hx_baseline1	*		***	*	CVD history at baseline: 0=No, 1=Yes
fibrate1	**		***	***	Use of fibrates at baseline: 0=No, 1=Yes
fpg	***		***	***	Fasting plasma glucose (mg/dL)
gender1	***	***	***	**	Gender: 0=Male, 1=Female
insulin1	***	*		***	Insulin use at baseline: 0=No, 1=Yes
meglitinide1	·				Use of meglitinides at baseline: 0=No, 1=Yes
niacin1		**	***	***	Use of niacin and nicotinic acid at baseline: 0=No, 1=Yes
other_diabmed1					Other diabetes treatments at baseline: 0=No, 1=Yes
other_lipidmed1		*	**	*	Use of miscellaneous lipid-lowering drugs at baseline: 0=No, 1=Yes
pc1	***	***	***	***	Principal components: 1-3
pc2		*		***	
pc3	**				
statin1	***	***		***	Use of HMG CoA reductase inhibitors (statins) at baseline: 0=No, 1=Yes
sulfonylurea1			***	**	Use of sulfonylureas at baseline: 0=No, 1=Yes
tzd1	***	***	***		Use of thiazolidinediones at baseline: 0=No, 1=Yes
yrsdiab	***	***	***	***	Years of diabetes diagnosis
yrslipi	***	***	***	***	Years of hyperlipidemia diagnosis

[a] significance codes: '·' $p < 0.01$; '*' $p < 0.05$; '**' $p < 0.01$; '***' $p < 0.001$.

Covariate pruning was performed separately for each of the four baseline lipid phenotypes. Table 4.1 and Table 4.2 summarize the significance for the forced and selected covariates, respectively, remaining after pruning. In these tables, variables that end with a number (excluding the principal components) are categorical while the others are continuous.

Residuals from linear regressions of the phenotypes against the pruned covariates are shown in Fig. 4.7. The upper panels are histograms of the residuals with normal distribution fit using the mean and standard deviation overlaid. The lower panels show intensities of the residuals versus fitted values.

Table 4.2 Baseline Lipid Model Selected Covariates

	Phenotype [a]				
Variable name	TC	LDL	HDL	TG	Description
alcohol1	–	–	***	***	Alcohol use in a typical week: 0=No, 1=Yes
dbp	na	na	na	na	Diastolic blood pressure (mmHg)
edu	–	–	–	–	1=Less than HS, 2=HS, 3=Some college, 4=College grad+
eyedisea	–	–	–	–	Eye disease at baseline: 1=Yes, 2=No
gfr	–	–	**	–	eGFR from 4 variable MDRD equation (ml/min/1.73 m^2)
hba1c	***	***	–	–	Glycated hemoglobin (%)
hyptens1	***	***	–	–	History of hypertension: 0=No, 1=Yes
network	–	–	–	–	Network identifier: 1-7
neuropat2	–	–	–	***	Neuropathy/nerve problems at baseline: 1=Yes, 2=No
pc4	–	–	***	***	Principal component 4
sbp	***	***	–	–	Systolic blood pressure (mmHg)
screat	na	na	na	na	Serum creatinine (mg/dL)
smoking1	–	–		–	} Cigarette smoking status: 0=Never, 1=Former, 2=Current
smoking2	–	–	***	–	
waist_cm	na	na	na	na	Waist circumference (cm)
x3lvh	–	–	–	–	Left ventricular hypertrophy by ECG or Echocardiogram within past 2 years: 1=Yes, 2=No/NA
x3malb	–	–	–	***	Micro or macro albuminuria within past 2 years: 1=Yes, 2=No/NA
x3sten	–	–	–	–	>=50% stenosis of coronary, carotid, or lower extremity artery within past 2 years: 1=Yes, 2=No/NA

[a] significance codes: 'na' removed due to collinearity; '–' not included in model; '·' p< 0.01; '*' p< 0.05; '**' p< 0.01; '***' p< 0.001.

Heritability was approximated using the set of genotyped variants to construct the genetic relationship matrix (GRM). The phenotype, pruned covariates and GRM were used by GCTA to compute approximate heritabillities of 11.7%, 11.3%, 43.7% and 14.0% for TC, LDL, HDL and TG, respectively.

4.2.2 Common Variants

All genotyped and imputed variants were tested for association using either PLINK or the custom R script, respectively. A set of LD pruned (VIF < 1.5), genotyped probes was used to compute genomic inflation values for different MAF thresholds, as shown in Fig. 4.8. A MAF threshold of 3% was chosen as that is the point at which the genomic inflation value for HDL begins to level off. The inflation values at that point are 1.00, 1.00, 1.02 and 1.00 for TC, LDL, HDL and TG, respectively. The number of variants passing this MAF threshold and included in the common variant analysis are 292,816 genotyped

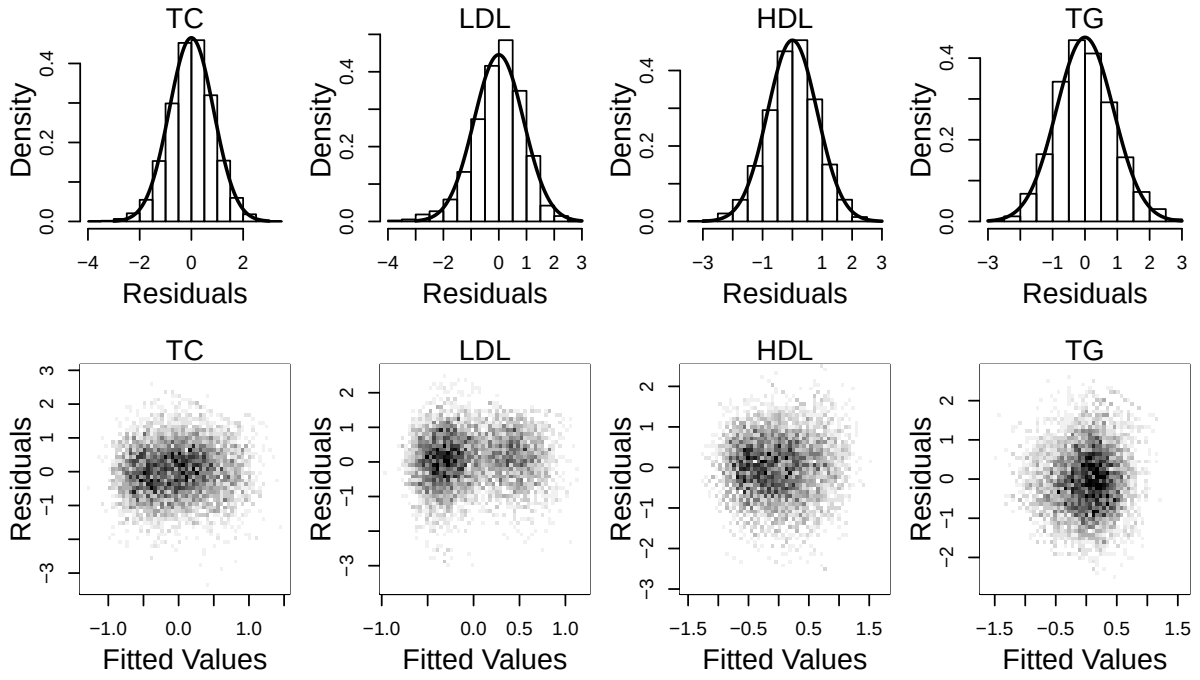


Figure 4.7 Residuals from linear regressions of the phenotypes against the pruned covariates.

and 7,812,348 imputed variants. Quantile-quantile (QQ) plots for the common variants are shown in Fig. 4.9. These plots compare the observed p-values to those expected under the null hypothesis of no genetic associations. Practically no inflation is observed, which is reflected in the genomic control values above. The location of significantly associate loci can be seen by using Manhattan plots. The negative log transformed, genomic control adjusted p-values versus genomic location for each of the phenotypes are shown in Figures 4.10 (TC) 4.11 (LDL) 4.12 (HDL) and 4.13 (TG).

4.2.3 Rare Variants

All variants below the MAF threshold chosen by genomic inflation values were included in the rare variant analysis. A total of 16,538 genes contained at least two variants that were below this MAF threshold in addition to having less than 15% missingness. The suite of 3 burden and 2 non-burden tests were applied and the resulting p-values were combined using the correlated Lancaster

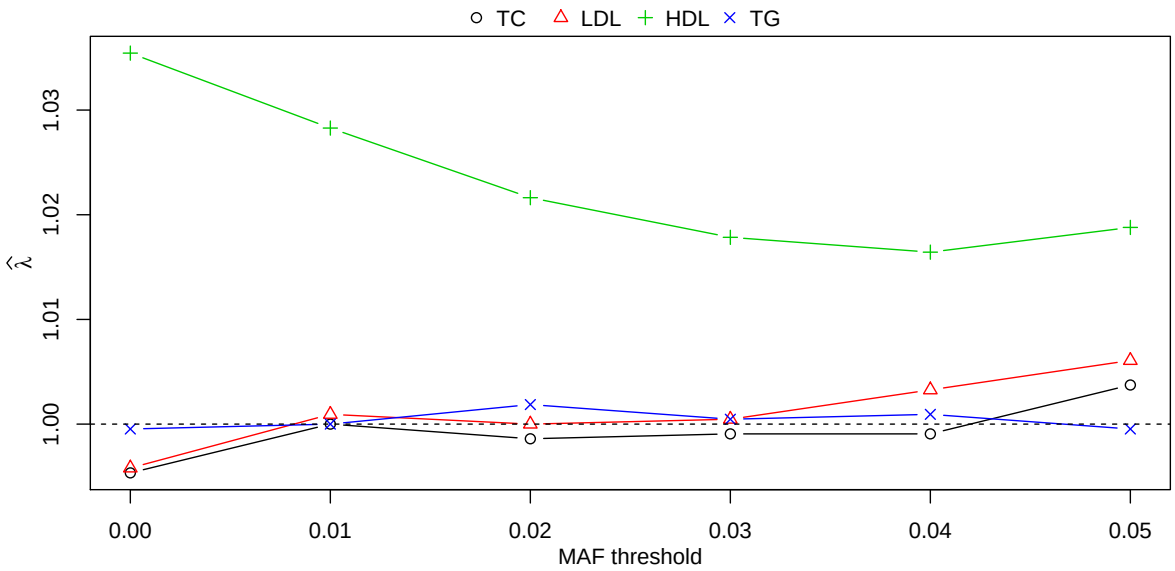


Figure 4.8 Genomic inflation values for different MAF thresholds.

procedure. QQ plots for the phenotypes are shown in Fig. 4.14, where the marker colors indicate q-value thresholds.

4.3 Baseline Lipid Discussion

Associations between lipid plasma concentrations and genetic markers have been of great interest because total cholesterol (TC), low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL) and triglycerides (TG) are heritable risk factors for cardiovascular disease. Lately, meta-analyses of genetic association studies have resulted in greater power to detect new loci containing common variants associated with specific lipid concentrations. Many rare variant studies are also being published as that field of research continues to be developed.

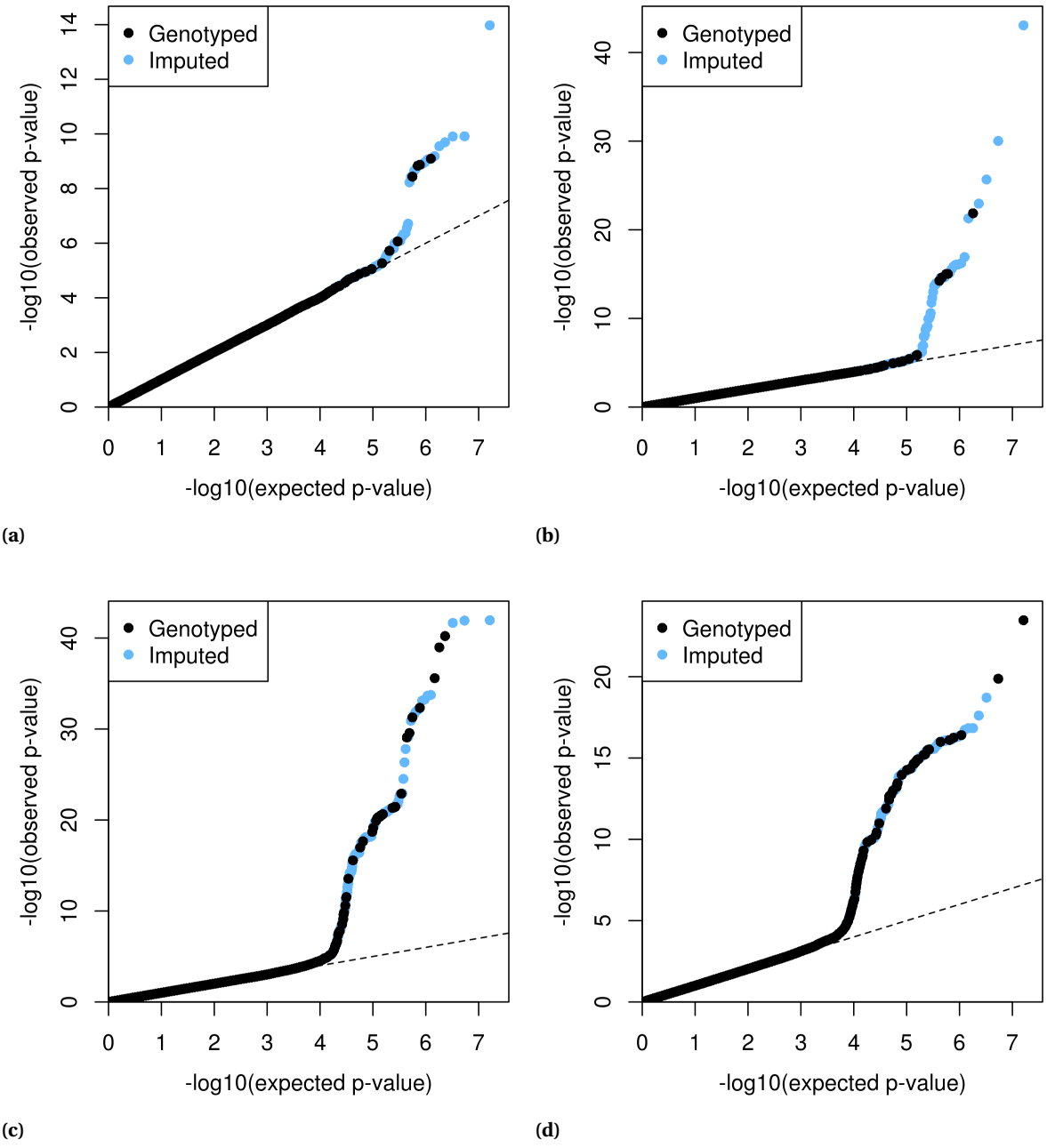


Figure 4.9 Common variant quantile-quantile plots for **a**) TC. **b**) LDL. **c**) HDL. **d**) TG.

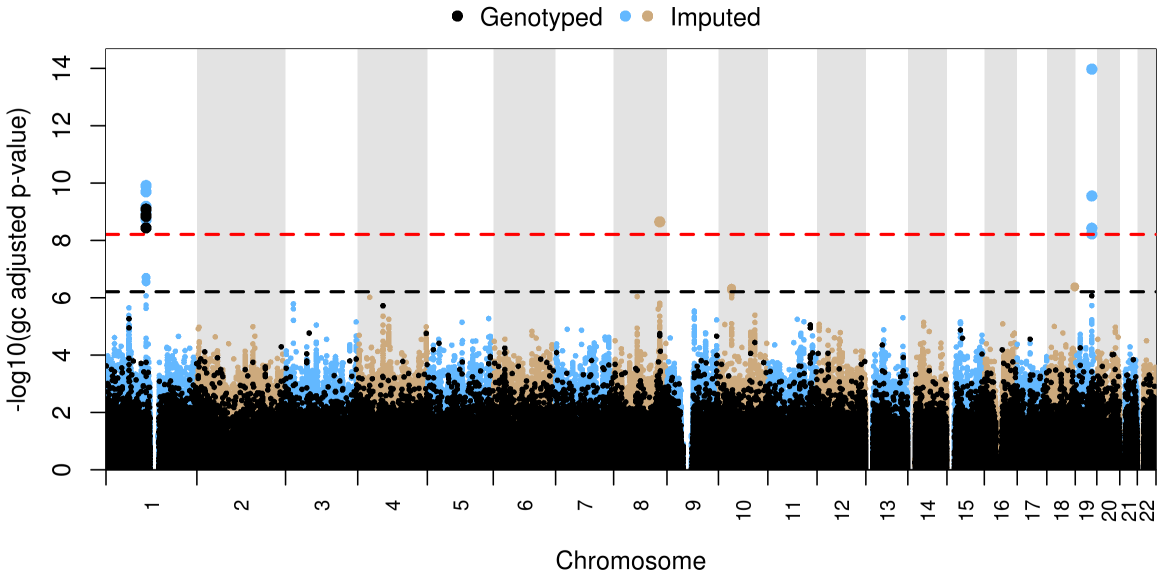


Figure 4.10 Manhattan plot for total cholesterol. Red line indicates significant p-values after Bonferroni adjustment. Black line are suggestive p-values.

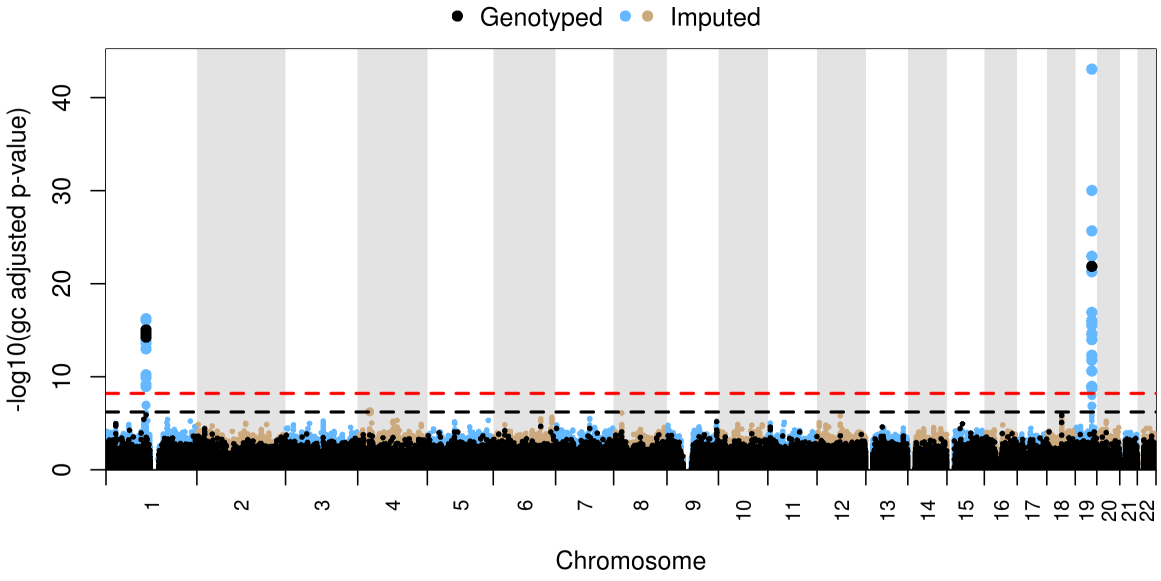


Figure 4.11 Manhattan plot for low density lipoprotein. Red line indicates significant p-values after Bonferroni adjustment. Black line are suggestive p-values.

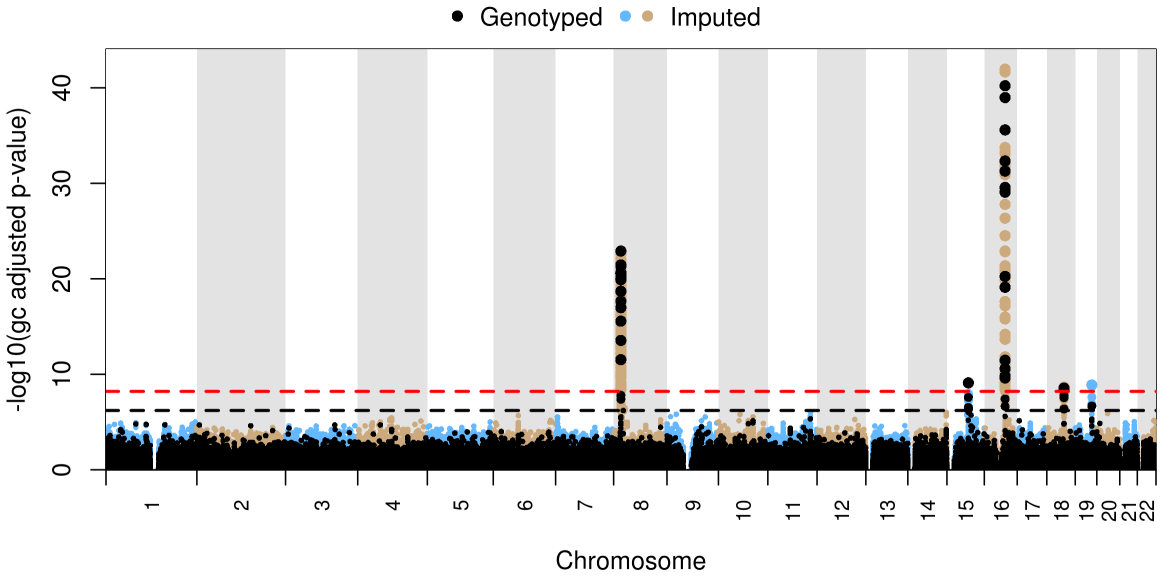


Figure 4.12 Manhattan plot for high density lipoprotein. Red line indicates significant p-values after Bonferroni adjustment. Black line are suggestive p-values.

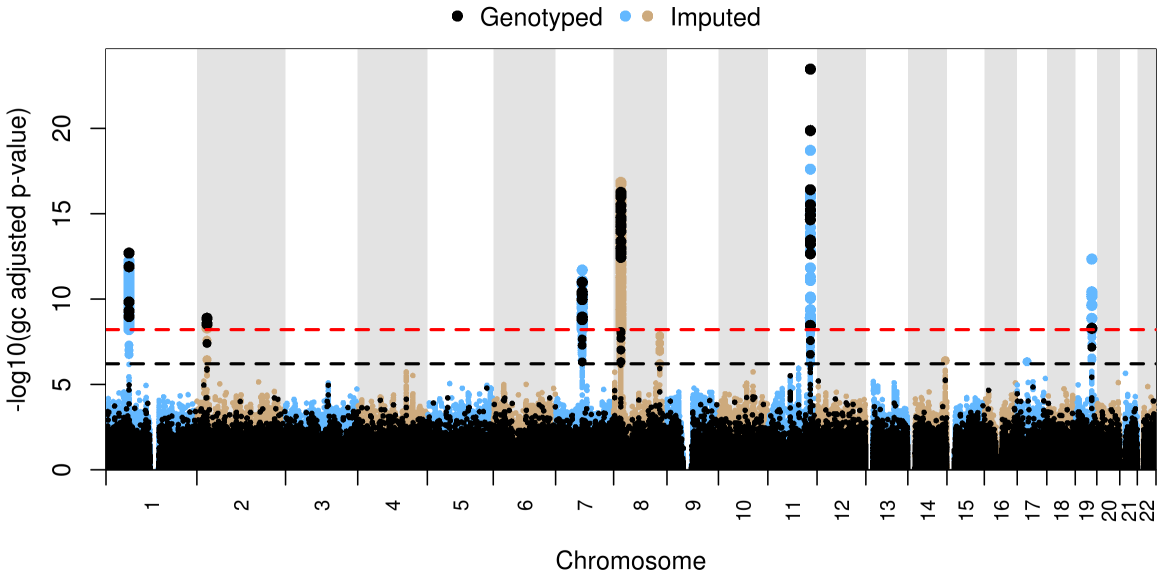


Figure 4.13 Manhattan plot for triglycerides. Red line indicates significant p-values after Bonferroni adjustment. Black line are suggestive p-values.

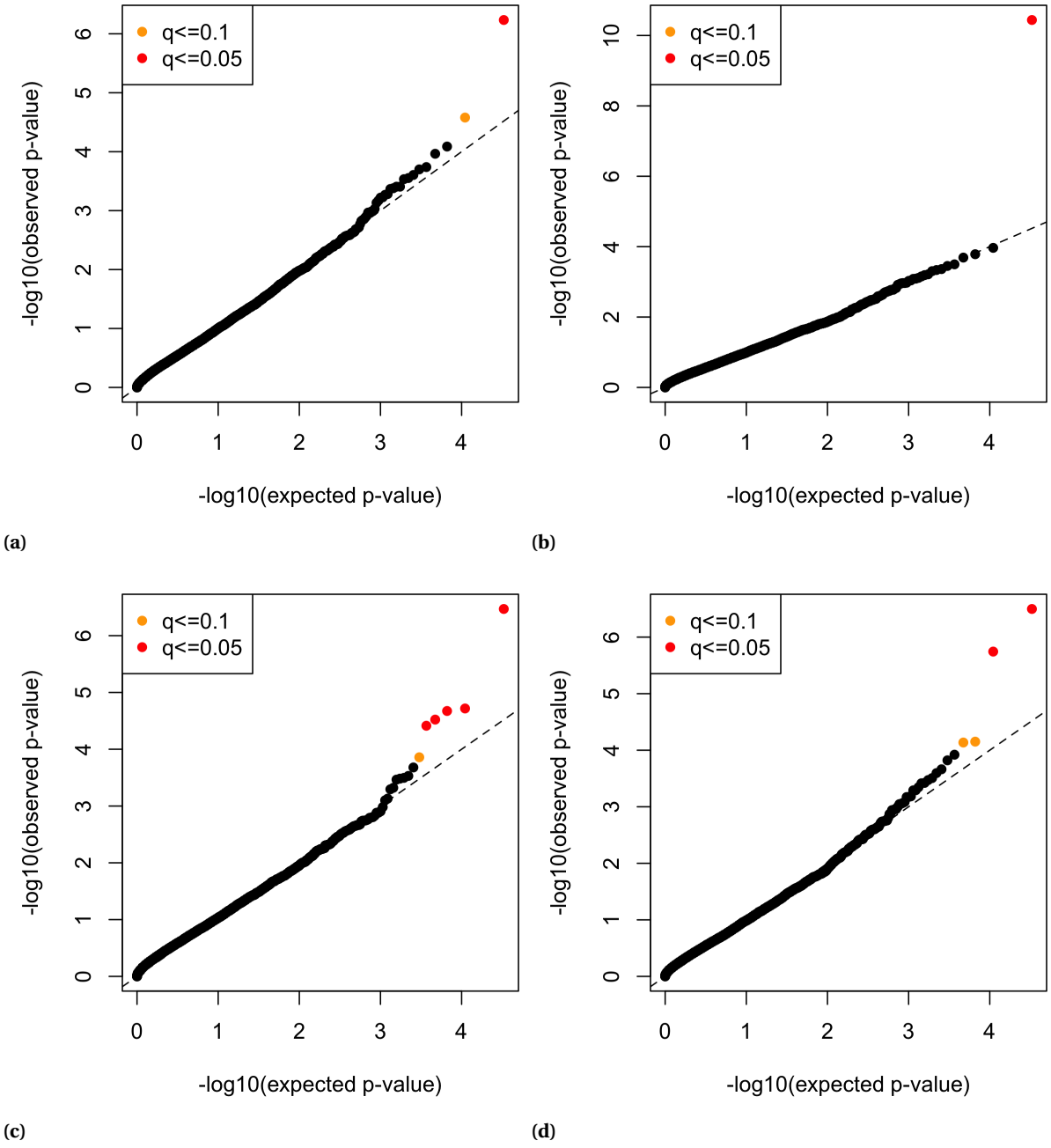


Figure 4.14 Rare variant quantile-quantile plots for **a**) TC. **b**) LDL. **c**) HDL. **d**) TG. Color indicate q-value threshold.

4.3.1 Common Variants

In 2010, Teslovich et al. published the results of a meta-analysis combining 46 lipid genome-wide association studies (GWASs) for all four lipid concentrations [Teslovich et al., 2010]. This study tested variants with a minor allele frequency (MAF) > 1% for associations with lipid levels involving ∼100,000 individuals of European ancestry. Imputation was performed using the Phase II CEU HapMap reference panel and a total of ∼2.6 million SNPs were tested for association with each of the four lipid traits in each study. Covariates included age, age^2, sex and optional study-specific covariates (e.g. PCs, study site). This study found 95 significant loci with p-values < 5x10^{-8}, which included all of the 36 loci previously reported in GWASs at the time and 59 novel loci.

This set of associated loci was expanded in a 2013 meta-analysis by the Global Lipids Genetics Consortium [Consortium, 2013]. This meta-analysis was comprised of 23 GWASs (∼95,000 individuals) and 37 Metabochip studies (∼94,000 individuals) for a total of ∼189,000 individuals, primarily of European ancestry. The 23 GWASs were a subset of those used in the 2010 meta-analysis and the Metabochip was a custom genotype chip including ∼200,000 variants based on previous GWAS findings. All studies used age, age^2 and sex as covariates. In addition, 24 studies used PCs or mixed-model approaches to adjust for population structure. This study identified 157 loci associated with lipid concentrations, of which 62 were novel.

Table 4.3 summarizes the findings of our baseline lipid common variant analysis. The lead maker corresponds to the variant with the smallest p-value for each peak when viewing a manhattan plot. The locus corresponding to each lead marker is either the gene in which the marker is located or the closest gene. Some markers are associated with multiple traits, in which case the p-value corresponds to the primary trait. All but two of these loci, CELSR2 and ZNF259, have been previously reported in the meta-analyses described above. While CELSR2 was not directly reported in the meta-analyses, a supplementary expression quantitative trait loci (eQTL) study in the first meta-analysis did find CELSR2 expression in human liver tissue to be associated with the lead marker found in SORT1 for

Table 4.3 Common Variant Genes of Interest

Locus	Lead marker	Chr	GRCh37 pos. (Mb)	Associated trait(s)[a]	Major/minor allele	MAF	Beta sign	p-value
ANGPTL3	rs67461605	1	63.07	TG	GTTAATGTG/-	0.34	−	2×10^{-13}
CELSR2	rs7528419	1	109.82	LDL, TC	A/G	0.23	−	6×10^{-17}
GCKR	rs1260326	2	27.73	TG	C/T	0.34	+	1×10^{-9}
MLXIPL	rs13240065	7	73.02	TG	G/A	0.11	−	2×10^{-12}
LPL	rs15285	8	19.82	HDL	C/T	0.32	+	5×10^{-24}
	rs75278536		19.82	TG	T/G	0.09	−	1×10^{-17}
TRIB1	rs28601761	8	126.50	TC	C/G	0.38	−	2×10^{-9}
	rs6982502		125.48	TG	T/C	0.41	−	1×10^{-8}
ZNF259	rs964184	11	116.65	TG	C/G	0.16	+	3×10^{-24}
LIPC	rs1532085	15	58.68	HDL	G/A	0.40	+	6×10^{-10}
CETP	rs247617	16	56.99	HDL	C/A	0.30	+	2×10^{-43}
LIPG	rs4939884	18	47.17	HDL	C/T	0.14	−	2×10^{-9}
APOE	rs7412	19	45.41	LDL, TC	C/T	0.08	−	8×10^{-44}
	rs75627662		45.41	TG	C/T	0.18	+	4×10^{-13}
	rs429358		45.41	HDL	T/C	0.14	−	9×10^{-10}

Chr, chromosome; MAF, minor allele frequency; TC, total cholesterol; TG, triglycerides.
[a] Primary trait listed first.

both LDL and TC. However, both of these loci are reported as significant in a meta-analysis done by Waterworth et al. [Waterworth et al., 2010], which investigated genetic associations for LDL, HDL and TG. This meta-analysis used data from 8 study populations involving ∼18,000 individuals of white European descent. Imputation was done using the HapMap II reference panel, resulting in ∼2.2 million SNPs. Covariates included age, sex and geographical or population variables where appropriate and only SNPs with MAF $>$ 1% were tested. While none of the loci in our common variant analysis are novel, the replication of previous results indicates that the QC and analysis pipelines are working well.

4.3.2 Rare Variants

LDL is one of the major risk factors for coronary artery disease and a whole-exome sequencing project was conducted to find associations between LDL and low-frequency and rare coding variants [Lange

et al., 2014]. The study began by sequencing ∼2000 individuals from 7 population-based cohorts, including ∼550 with extremely high and low LDL. This was followed by sequencing an additional ∼1300 individuals from the same 7 studies for 17 promising genes and genotyping ∼52,000 individuals for 15 genes from 12 studies. Single variant analysis was done for variants with MAF $> 1\%$. Rare variant association was done using a suite of burden and non-burden tests. The burden tests used the combined multivariate and collapsing (CMC) test with multiple MAF thresholds ($< 5\%$, 1%, 0.5% and 0.1%) for nonsynonymous and splice variants and MAF $< 5\%$ for loss of function (LoF) variants. The non-burden tests used the optimized sequence kernel association test (SKAT-O) with MAF $< 5\%$ for LoF, LoF + missense predicted as "probably damaging" by PolyPhen-2 and missense + LoF variants. The p-value for each gene was selected as the minimal p-value across all 8 rare variant tests. Covariates for both single variant and rare variant analyses included age, sex, ethnicity, primary cardiovascular phenotype, principle component 1 (PC1) and PC2.

Liu et al. conducted a meta-analysis of rare variant association tests for blood lipid levels (LDL, HDL and TG) involving ∼18,500 individuals genotyped with Illumina exome arrays from 7 studies [Liu et al., 2014]. A total of ∼171,000 variants were polymorphic with ∼125,700 having a MAF $< 1\%$ and were grouped into ∼16,000 genes. Covariates included age, age^2, sex and cohort-specific covariates (e.g. PCs). Five different rare variant tests were used; simple burden tests that sum the number or rare alleles with MAF cutoffs of 5% and 1%, a variable-threshold burden test with MAF cutoff of 5%, and SKAT with MAF cutoffs of 5% and 1%. Meta-analyses were performed for each of the rare variant tests, resulting in 5 p-values for each gene.

Table 4.4 summarizes the findings of our baseline lipid rare variant analysis. Some markers are associated with multiple traits, in which case the p-value corresponds to the primary trait. Potentially novel genes are in bold. PCSK9 was the only gene we found to be associated with LDL in the rare variant analysis. This result replicates the findings of Lange et al., where PCSK9 had the strongest association with LDL out of the 3 genes found to be exome-wide significant in their rare variant analysis. The meta-analysis by Lui et al. also found PCSK9 to be significantly associated with LDL. In

Table 4.4 Rare Variant Genes of Interest

Gene[a]	Chr	Associated trait(s)[b]	Number of Variants[c]	p-value	q-value
PCSK9	1	LDL, TC	22 (12)	4×10^{-11}	1×10^{-7}
LPL	8	HDL	16 (15)	3×10^{-5}	0.0301
APOC3	11	TG	2 (1)	3×10^{-7}	0.0018
PAFAH1B2	11	TG	3 (1)	7×10^{-5}	0.0775
CNOT2	12	HDL	2 (2)	2×10^{-5}	0.0283
CETP	16	HDL	18 (14)	4×10^{-7}	0.0014
HNF1B	17	HDL	5 (3)	1×10^{-4}	0.0848
ANGPTL4	19	TG, HDL	14 (12)	2×10^{-6}	0.0049
HPN-AS1	19	HDL	2 (2)	1×10^{-4}	0.0915
SIRPD	20	TG	5 (3)	8×10^{-5}	0.0775
UBE2L3	22	TG	2 (0)	4×10^{-5}	0.0636

Chr, chromosome; TC, total cholesterol; TG, triglycerides.
[a] Bold indicates potentially novel gene association. [b] Primary trait listed first. [c] Genotyped + imputed variants. Number genotyped reported inside parentheses.

addition, they found significant associations for LPL and ANGPTL4, further supporting the findings of our analysis. UBE2L3 was found to be associated with HDL in the common variant meta-analysis by Teslovich et al. [Teslovich et al., 2010], but our rare variant results demonstrate an association with TG as well.

Peloso et al. conducted a mega-analysis of 13 studies for a total of ∼56,000 individuals, 42k of European ancestry (EA) and 14k of African ancestry (AA), which investigated the association of low-frequency and rare coding variants with LDL, HDL and TG [Peloso et al., 2014]. Low-frequency variants were defined as having MAF $> 0.02\%$ in EA individuals and $> 0.07\%$ in AA individuals and were tested using single variant analysis methods. Gene-based tests only included non-synonymous and splice site variants and again several rare variant tests were used: T1 (CMC with MAF $< 1\%$), T0.1 (CMC with MAF $< 0.1\%$) and SKAT with MAF $< 5\%$. Covariates included age, age^2, sex and up for first 4 PCs. In addition, the allele dosages of top previously identified GWAS variants in each region were used as covariates. Gene-based rare variant associations were found for APOC3 and CETP, which we replicated

in our findings. Peloso et al. also found a novel low-frequency variant association for PAFAH1B2 with HDL and TG, which supports our rare variant association between PAFAH1B2 and TG.

4.3.3 Novel Genes of Interest

Most of our significant associations have been reported in the literature. This is a good indication that the QC/imputation/analysis pipelines are working well and produce results that are replications of previous findings. The RV analysis produced a small list of novel loci for both HDL and triglyceride levels. Fig. 4.15 shows boxplots of the log-transformed phenotype versus genotype for the rare variants of each gene. The boxplots are color coded for each variant and the number of samples for each genotype are indicated below the allele labels. Genotyped variants are indicated by an asterisk following the rsID.

The vast majority of the functional variants are missense. The only exceptions to this are rs145641015 and rs150515110 in SIRPD, which are both STOP-GAIN variants. In addition, there are two annotation complications. In HNF1B, rs8068014 is annotated as both a missense and 3-prime untranslated variant. The missense annotation is from a predicted mRNA model. In UBE2L3, rs186477191 is annotated as missense, non-coding and an intron. The missense annotation is also from a predicted mRNA model.

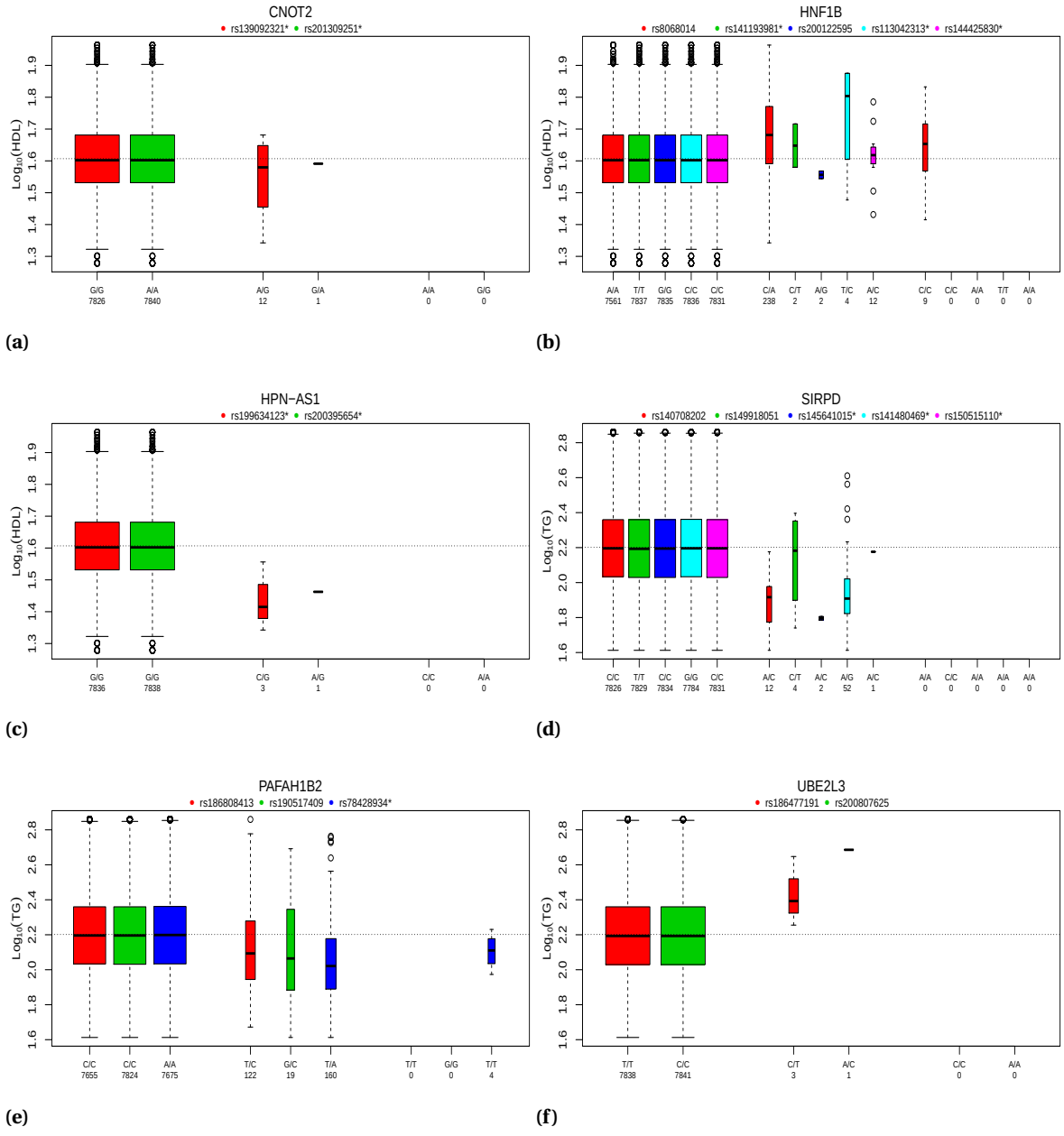


Figure 4.15 Boxplots of log-transformed phenotype versus genotype for variants.

Chapter

5

Grammatical Evolution Support Vector Machine

This chapter covers the development of the grammatical evolution support vector machine (GESVM) algorithm. The ultimate goal of GESVM is to detect epistatic effects in large datasets where brute force methods are infeasible. The approach of GESVM is to use an evolutionary algorithm (GE) to explore the search space of features and architectures that define the supervised learning model (SVM).

5.1 Motivation

The ability to identify genes that predict common, complex human diseases is an intense area of research. Such diseases are often caused by the combination of many genetic and environmental factors, each contributing a small effect [Manolio, 2010]. Identification of genetic factors is made difficult by the interactions between different genes, referred to as epistasis [Cordell, 2002]. Traditional parametric statistical methods used to characterize gene-gene or gene-environment interactions fail when applied to large datasets [Culverhouse et al., 2002], which has stimulated the development of novel computational approaches that are able to extract information from data obtained during this 'omics' era.

One popular approach for detecting disease association involves the use of machine-learning classification methods [Capriotti et al., 2006; Chen et al., 2008; Motsinger-Reif et al., 2008; Wei et al., 2009]. A few of the most common methods are artificial neural networks (ANNs), decision trees (DTs) and support vector machines (SVMs), the later of which has been steadily gaining popularity. Due to the enormous size of the datasets that are being analyzed, feature selection is an extremely important aspect of these classification methods [Saeys et al., 2007]. In addition, properties innate to the classification technique also influence performance, e.g. the architecture of an artificial neural network or the kernel parameter(s) of a support vector machine.

To address these issues, many techniques are being developed that combine machine-learning classification methods with algorithms that select features and classifier architecture [Chen et al., 2008;

Jack & Nandi, 2002; Motsinger-Reif et al., 2008; Samanta et al., 2003]. Genetic programming algorithms are often used for this purpose, however, application of grammatical evolution (GE) has been shown to outperform the genetic programming counterpart for ANNs [Motsinger-Reif et al., 2008]. Motivated by this result and the increasing use of SMVs, this work begins the process of combining GE and SVMs for the purpose of predicting human genetic disease associations.

5.2 Grammatical Evolution

Grammatical evolution (GE) is a technique similar to genetic programming that is capable of exploring vast search spaces involving potential features and classification architectures. Details of GE can be found in O'Neill and Ryan [O'Neill & Ryan, 2003]. GE uses a Backus-Naur form grammar to convert a binary string into a computer program in any language by following the central dogma of biology where a binary string (genotype) is converted to an integer string (mRNA) and then mapped to a program (phenotype). Backus-Naur form consists of sets of non-terminal symbols (N), terminal symbols (T), starting symbols (S), and production rules (P). The use of a grammar allows for unconstrained searching via evolutionary operations that occur at the genetic level, such as point mutations and crossover events, while maintaining a valid phenotypic program. GE has been used successfully in the creation of ANNs and DTs for identifying disease susceptibility genes [Motsinger-Reif et al., 2008]. The use of genetic algorithms to optimize the feature selection and architecture of both ANNs and SVMs has been investigated [Chen et al., 2008; Jack & Nandi, 2002; Samanta et al., 2003], however, to our knowledge GE has not been applied to SVMs.

5.3 Support Vector Machine

Support vector machines (SVMs) are non-probabilistic binary classifiers that can be used to construct a hyperplane to separate data into one of two classes [Vapnik, 2000]. Consider a set of n data points, each consisting of p features, $\mathbf{x} \in \mathbb{R}^p$, and a class label, $y \in [-1,1]$, i.e. $(\mathbf{x}_i, y_i)$ for $i = 1,\ldots,n$. A

hyperplane can be defined by a normal vector, $\mathbf{w}$, and offset, b. In addition, slack variables, ξ_i, can be introduced to represent the degree of misclassification when data points are not linearly separable. The objective function of the SVM is then

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b) \geq 1-\xi_i, \quad \xi_i \geq 0, \end{aligned}$$

where C is a linear misclassification penalty and ϕ is a nonlinear transformation function that projects $\mathbf{x} \in \mathbb{R}^p$ into a higher-dimensional feature space. Using the relationship $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$, this problem can be solved in dual form

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \end{aligned}$$

where $k(\mathbf{x}_i,\mathbf{x}_j) = \phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j)$ is a nonlinear kernel function. One of the most popular kernel functions is the radial basis function (RBF) $k(\mathbf{x}_i,\mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i-\mathbf{x}_j\|^2}$, for $\gamma > 0$ [Chen et al., 2008]. The RBF projects $\mathbf{x}$ into an infinite number of dimensions, which can be seen by expanding the exponential $e^{-\gamma\|\mathbf{x}_i-\mathbf{x}_j\|^2} = \sum_{k=0}^{\infty}\frac{1}{k!}(-\gamma\|\mathbf{x}_i-\mathbf{x}_j\|^2)^k$.

5.4 Grammatical Evolution Support Vector Machine (GESVM)

Grammatical evolution support vector machine (GESVM) is an algorithm that uses GE to select features and parameter values for an SVM classifier that can be applied to case/control data for human diseases. GESVM uses the RBF kernel, which requires a value for the kernel parameter γ in addition to the misclassification penalty C and was designed for use with datasets where the features consist of single nucleotide polymorphisms (SNPs). SNPs are point mutations that show variation

across a population and can result in a variety of effects, such as modifying transcription binding sites or changing a protein's amino acid composition. A typical representation of a SNP is to encode the genotype into integers, for example if the major and minor alleles are represented as A and a, respectively, then the genotype may be coded as $AA = 0,\ Aa = 1,\ aa = 2$.

The grammar used to select features and parameter values for the GESVM can be expressed in Backus-Naur Form as follows:

$$
\begin{aligned}
N &= \{C_1,\ \gamma_1,\ C_2,\ \gamma_2,\ X,\ L,\ E\} \\
T &= \{\text{C values},\ \gamma \text{ values},\ \text{SNP indices}\} \\
S &= \{< C_1 >< \gamma_1 >< X >< L >\}
\end{aligned}
$$

with P defined as

$$
\begin{aligned}
< C_1 >, < C_2 > ::= &\ \text{C values} \\
< \gamma_1 >, < \gamma_2 > ::= &\ \gamma \text{ values} \\
< X > ::= &\ \text{SNP indices} \\
< L > ::= &\ < X >< E > \\
&| < X >< L > \\
&| < X >< L >< L > \\
&| < X >< L >< L >< L > \\
&| < E > \\
< E > ::= &\ < C_2 >< \gamma_2 > .
\end{aligned}
$$

The grammar begins by selecting values for C_1 and γ_1, which are used to determine 70% of the misclassification penalty C and kernel parameter γ, respectively. Next, a value X is chosen, which corresponds to a specific SNP index in the dataset. After selecting the initial SNP, the grammar will replace L with the appropriate line, allowing the possibility for additional SNP loci to be selected.

Once the grammar reaches an E, the remaining proportion of C and γ are selected and the mapping process ends. Duplicate SNP indices are allowed to be selected, however, after the mapping process duplicates are removed and a list of unique SNP indices is returned.

Splitting the values of C and γ into two components to allow the "ripple effect" to influence these parameters. The "ripple effect" is a property of grammatical evolution that occurs when a mutation upstream in the binary string genome results in a different mapping sequence after that point. For example, if a mutation caused the grammar to replace L with XL rather than E, then the codons that would have been used to choose values for C_2 and γ_2 are instead used to select an additional SNP and select another line from L. This "ripple effect" can cause SNPs to be added or removed from the model and may shift the location in the genome where the second component of C and γ are obtained.

The development of GESVM involved several phases. It was initially implemented in MATLAB to quickly test the validity of the method. After promising results of a simulation study in the MATLAB implementation, GESVM was implemented in C++ to avoid requirement of a license and to allow more flexibility in data and memory management. The C++ implementation was then modified to incorporate covariate data and allow for application to continuous phenotype data using SVM regression. Simulation studies were performed for each of the three implementations and the GE parameters are shown in Table 5.1.

The grammar, chromosome size and wrapper count limit the number of SNPs that are selected for the SVM classifier. The distribution for the expected number of SNPs selected in a randomly initialized population using the grammar and GE parameters is shown in Figure 5.1. A target range of two to five SNPs was desirable because models with a large number of variants may result in overfitting and cause difficulty when drawing meaningful biological interpretations.

Table 5.1 Grammatical evolution parameters for simulation studies

Parameter	MATLAB	C++	C++ regression
population size	50	100	100
max generations	20	100	15
crossover rate	0.8	0.9	0.9
mutation rate	0.05	0.05	0.05
codon size	8	8	16
chromosome size	25	25	25
wrapper count	2	2	2
selection	tournament	tournament	tournament
fitness	balanced accuracy	balanced accuracy	mean squared error

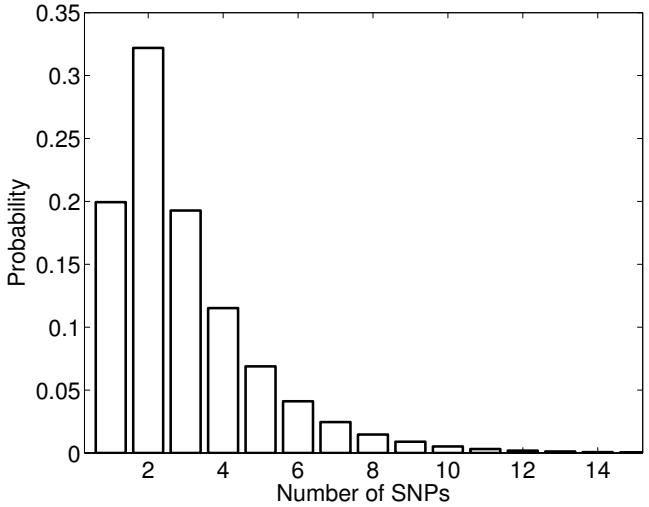


Figure 5.1 Distribution for the number of SNPs selected by the grammar mapping process for a population of 10^6 randomly generated binary strings.

5.5 GESVM (MATLAB Implementation)

The MATLAB implementation of GESVM requires a single input file where each row represents the data from an individual with each column representing the genotype of a SNP, except for the first column which contains the case/control phenotype value (0 or 1). The SVM optimization problem for the training dataset was solved using the quadratic programming routine `quadprog` with the `interior-point-convex` algorithm. Computations were performed using the High Performance Computing hardware at NCSU.

5.5.1 Simulation study

This preliminary study fixed the parameter values C and γ at 5 and 2, respectively, based on an initial parameter sweep of a dataset generated using a purely epistatic model. Initial populations were randomly generated, however, the codons used to select the first SNP value were modified to ensure each SNP was represented in the first generation of the population. GE parameters are reported in Table 5.1. The chromosome size was fixed at 25 codons and a one-point crossover was used. Selection was performed by keeping the proportion of the population that had the best fitness as determined by balanced classification accuracy. Ten-fold cross-validation was used to reduce overfitting by splitting the data into different training and testing sets.

5.5.1.1 Data Generation

For the purposes of this initial work, genetic models that were not purely epistatic were generated. While the eventual goal of this work is to detect gene-gene interaction, initial evaluation of this method is made easier by using models with main effects from single SNPs in addition to epistatic effects. We used penetrance functions, which specify the proportion of individuals with a given genotype that are cases, to define the probability of disease given a particular genotype. Two models were tested, both with similar penetrance functions but different genotype frequencies. The penetrance functions for

the two models are shown in Fig. 5.2a and Fig. 5.2b. Minor allele frequencies were adjusted so that

	BB	Bb	bb
AA	0.11	0.1	0
Aa	0.1	0.11	0
aa	0	0	0

(a)

	BB	Bb	bb
AA	0.3	0.3	0
Aa	0.3	0.3	0
aa	0	0	0

(b)

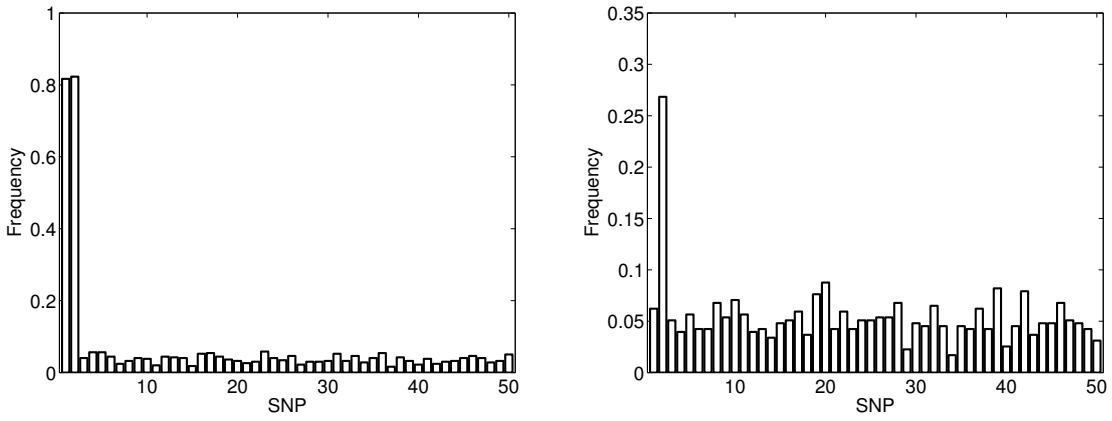
Figure 5.2 Multilocus penetrance functions for models **a**) M_1 and **b**) M_2.

the heritability, or fraction of the trait (disease) variance that is due to genotype, would be modest and were calculated for both models to be ∼0.05 according to Culverhouse et al. [Culverhouse et al., 2002] when the genotypes were generated according to Hardy-Weinberg proportions. The minor allele frequencies for model M_1 were $p(a) = 0.5$ and $p(b) = 0.5$, while for model M_2 they were $p(a) = 0.15$ and $p(b) = 0.32$. Twenty-five datasets consisting of 300 cases and 300 controls with a total of 50 SNPs were generated for each model using the software GenomeSIMLA [Dudek et al., 2006; Edwards et al., 2008] with the disease loci $A, a = \text{SNP}_1$ and $B, b = \text{SNP}_2$.

5.5.1.2 Results

GESVM was applied to the case/control datasets generated from disease model M_1. The architectures of the SVMs for the most fit individuals from each cross-validation step for all 25 datasets were recorded. In this preliminary work C and γ were kept constant so only the number of SNPs and SNP values were different among individual SVM architectures. The accuracy of the best fit individuals ranged from 65.0% to 90.0% with mean 75.6% and standard deviation 4.6%. The expected accuracy of model M_1 was 73.3%. A distribution of which SNP values were included in the SVM models is shown in Figure 5.3a. Of the best fit individuals, 6.6% included SNP_1 but not SNP_2, 7.2% included SNP_2 but not SNP_1, and 75.1% included both SNP_1 and SNP_2. An SVM model containing all 50 SNPs was applied

to the data and the accuracy when using all features ranged form 40.0% to 76.7% with mean 60.4% and standard deviation 6.2%. This analysis was repeated using the datasets generated from disease



(a) **(b)**

Figure 5.3 Frequency of SNPs included in top-performing SVMs for models **a**) M_1 and **b**) M_2.

model M_2. The accuracy of the best fit individuals ranged from 61.7% to 78.3% with mean 69.2% and standard deviation 3.1%. The expected accuracy of model M_2 was 58.3%. A distribution of which SNP values were included in the SVM models shown in Figure 5.3b. Of the best fit individuals, 4.5% included SNP_1 but not SNP_2, 25.1% included SNP_2 but not SNP_1, and 1.7% included both SNP_1 and SNP_2. An SVM model containing all 50 SNPs was applied to the data and the accuracy when using all features ranged form 38.3% to 71.7% with mean 53.3% and standard deviation 6.5%.

These preliminary results show promise for using GESVM as a potential method for identifying human disease susceptibility genes. The results presented in the previous section were obtained using unoptimized GE and SVM parameters, yet the best fit SVMs were still able to identify SNP_1 and SNP_2 for disease model M_1 and SNP_2 for disease model M_2. The overall decrease in ability for GESVM to identify disease associated SNPs in model M_2 compared to M_1 can be attributed to the combination

of smaller minor allele frequencies and larger penetrance in model M_2, which resulted in a lower expected accuracy as well. This combination caused the datasets for disease model M_2 to have a larger proportion of the generated population with haplotypes where the value of the penetrance function is nonzero. Having a larger proportion of the population consisting of incomplete penetrance haplotypes reduced the chance of finding effects from either SNP_1 or SNP_2. In a similar manner, a smaller minor allele frequency for SNP_1 relative to SNP_2 in model M_2 further reduced the ability for GESVM to detect effects from SNP_1, which resulted in GESVM only finding SNP_2.

5.6 GESVM (C++ Implementation)

The C++ implementation of GESVM requires the same input file as the MATLAB implementation. The SVM optimization problem for the training dataset and prediction for the testing dataset were solved using libsvm [Chang & Lin, 2011]. Computations were performed using the High Performance Computing hardware at NCSU.

5.6.1 Simulation Study

The parameter values C and γ were allowed to range from 0.0001 to 2 based on an initial parameter sweep of a dataset generated using a purely epistatic model. GE parameters are reported in Table 5.1. The chromosome size was fixed at 25 codons and a one-point crossover was used. Selection was performed by keeping the proportion of the population that had the best fitness as determined by balanced classification accuracy. Ten-fold cross-validation was used to reduce overfitting by splitting the data into different training and testing sets.

5.6.1.1 Data Generation

A total of 27 disease models were tested corresponding to each combination of MAF = {0.1, 0.25, 0.5}, sample size = {600, 1200, 1800} and H^2 = {2%, 5%, 10%}, where sample size is split equally into cases

Table 5.2 Details for the 27 generated models

	Inputs					Penetrance[a]								
Model	MAF	size	H^2	OR	Ex	AABB	AaBB	aaBB	AABb	AaBb	aaBb	AAbb	Aabb	aabb
1	0.1	300	0.02	4.4973	0.5882	0.1291	0.2297	0.2912	0.2299	0.2734	0.2979	0.2911	0.2978	0.4
2	0.1	300	0.05	4.6881	0.636	0.1245	0.3095	0.3005	0.3095	0.302	0.3001	0.3005	0.3001	0.4
3	0.1	300	0.1	4.9853	0.6821	0.1183	0.4008	0.3029	0.4004	0.3139	0.3006	0.3029	0.3006	0.4
4	0.1	600	0.02	4.4914	0.5884	0.1292	0.2305	0.2907	0.2314	0.2682	0.298	0.2908	0.2975	0.3999
5	0.1	600	0.05	4.6881	0.6342	0.1245	0.3338	0.301	0.2765	0.3047	0.3002	0.3009	0.3003	0.4
6	0.1	600	0.1	5.2901	0.682	0.1182	0.3838	0.3025	0.4149	0.3175	0.3007	0.3047	0.3007	0.4
7	0.1	900	0.02	4.4973	0.5882	0.1291	0.2302	0.2912	0.2296	0.273	0.298	0.2912	0.2979	0.4
8	0.1	900	0.05	4.6881	0.6355	0.1245	0.3236	0.3006	0.2939	0.2996	0.2991	0.301	0.2993	0.4
9	0.1	900	0.1	5.6251	0.6812	0.1177	0.3632	0.3043	0.4287	0.3173	0.301	0.3028	0.3023	0.4002
10	0.25	300	0.02	4.4169	0.5778	0.1301	0.2421	0.276	0.2411	0.25	0.282	0.2762	0.2819	0.3978
11	0.25	300	0.05	4.7899	0.6148	0.1224	0.2988	0.3046	0.3622	0.3396	0.3027	0.3042	0.3027	0.4005
12	0.25	300	0.1	5.9216	0.6713	0.1179	0.1651	0.3114	0.4418	0.4233	0.305	0.3113	0.305	0.4011
13	0.25	600	0.02	4.4111	0.5779	0.1302	0.2425	0.2758	0.2422	0.2489	0.282	0.2754	0.2819	0.3977
14	0.25	600	0.05	4.8078	0.6161	0.122	0.331	0.3055	0.3219	0.365	0.3032	0.3056	0.3032	0.4005
15	0.25	600	0.1	5.9534	0.6638	0.1168	0.4389	0.2952	0.4405	0.1895	0.2981	0.295	0.2979	0.4009
16	0.25	900	0.02	4.4206	0.5777	0.1301	0.2406	0.2772	0.2419	0.2489	0.2834	0.2774	0.2821	0.398
17	0.25	900	0.05	4.7944	0.6119	0.1223	0.2679	0.3068	0.3454	0.3694	0.3043	0.3089	0.3063	0.4005
18	0.25	900	0.1	5.9273	0.6703	0.1178	0.4418	0.3123	0.1677	0.4268	0.3036	0.2881	0.3103	0.4011
19	0.5	300	0.02	4.8789	0.5567	0.124	0.2794	0.2929	0.2743	0.3735	0.3107	0.2961	0.2718	0.4085
20	0.5	300	0.05	5.0508	0.6152	0.1236	0.3833	0.2947	0.2302	0.1609	0.3732	0.3156	0.3728	0.416
21	0.5	300	0.1	6.3709	0.666	0.1221	0.1528	0.2864	0.1528	0.4698	0.1528	0.2205	0.4298	0.4253
22	0.5	600	0.02	4.9541	0.5566	0.124	0.3595	0.3111	0.2961	0.2344	0.2851	0.3007	0.3042	0.4122
23	0.5	600	0.05	5.6123	0.6022	0.1232	0.3149	0.2748	0.1948	0.4409	0.1952	0.2695	0.3214	0.4134
24	0.5	600	0.1	5.587	0.6701	0.1222	0.4375	0.3403	0.1557	0.1258	0.4362	0.2343	0.4375	0.4278
25	0.5	900	0.02	4.9689	0.5586	0.1239	0.2764	0.3019	0.3236	0.2423	0.3624	0.3064	0.2766	0.4127
26	0.5	900	0.05	5.0842	0.6107	0.1236	0.3136	0.2996	0.2409	0.1594	0.3955	0.3017	0.3954	0.4176
27	0.5	900	0.1	6.3497	0.6696	0.1222	0.1565	0.2229	0.1574	0.4692	0.431	0.2198	0.1563	0.4254

MAF - minor allele frequency, H^2 - broad sense heritability, OR - odds ratio, Ex - expected accuracy
[a] proportion of individuals with a given genotype that are cases.

and controls and H^2 is the broad sense heritability. A MATLAB script was written to automatically generate penetrance functions (with penetrance values constrained between 0.1 and 0.5) for each of the 27 models using the equations described by Culverhouse et al. [Culverhouse et al., 2002]. Details of the generated models are shown in Table 5.2. The expected accuracy ranged from 55.6% to 68.2% with odds ratios ranging from 4.4 to 6.4, where the odds ratio was computed from the penetrance table, P, as $OR = \frac{\max P}{1-\max P} / \frac{\min P}{1-\min P}$. GenomeSIMLA [Dudek et al., 2006; Edwards et al., 2008] was used to generate 100 datasets for each model with the disease loci $A, a = \text{SNP}_1$ and $B, b = \text{SNP}_2$.

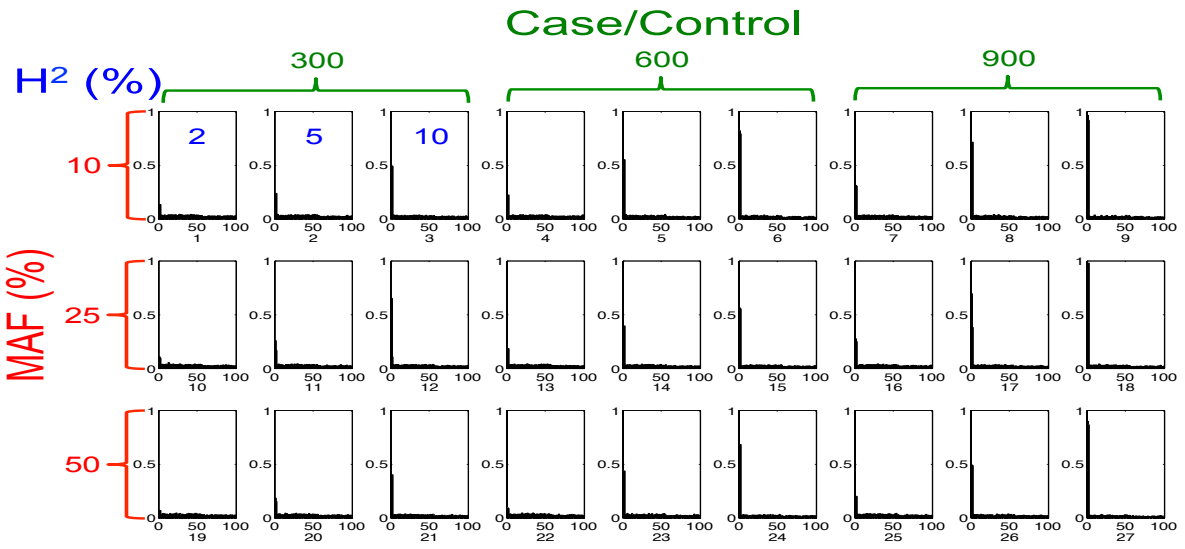
5.6.1.2 Results

The architectures for the 5 most accurate SVMs from each cross-validation step for all 100 datasets were recorded for each model. Fig. 5.4a depicts the frequency for each of the 100 SNPs for each model. The frequency of only the causal SNPs in the 5 most accurate individuals are shown in Fig. 5.4b.

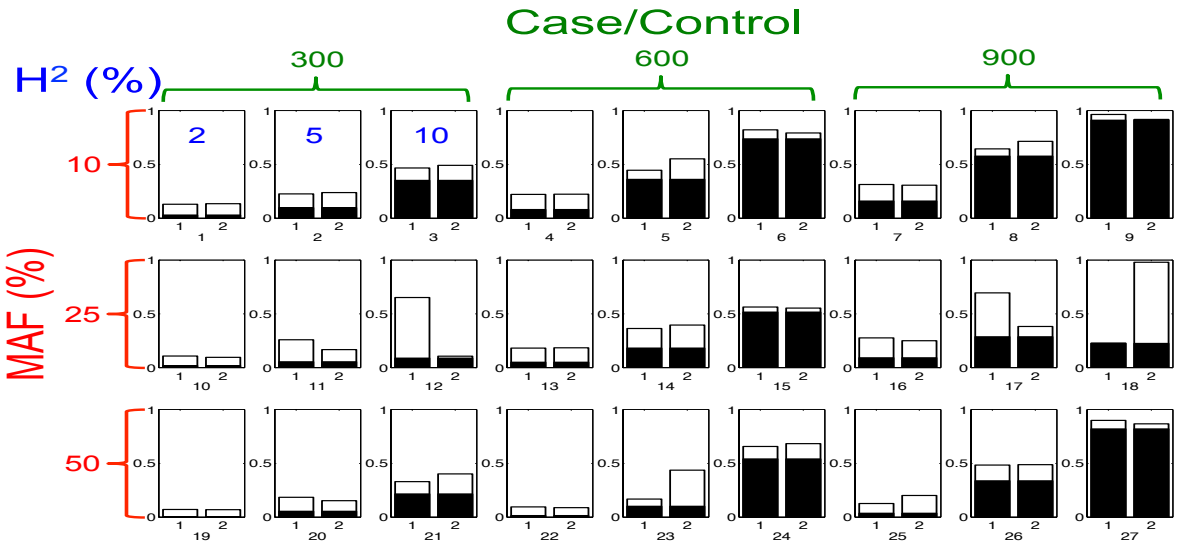
The models were generated using an optimization routine with random initial conditions so the ability of GESVM to identify causal SNPs is not necessarily comparable between models. However, based on the results above there seems to be a general trend for GESVM to more easily identify the causal SNPs as the heritability and sample sizes increase. The disease models with MAF = 0.1 seem to be more consistent when compared to the results for MAF = 0.25 and 0.5. This observation reflects the differences in the expected accuracy, e.g. for a sample size of 600 cases/controls and H^2 of 0.05, the expected accuracy for MAFs of 0.1, 0.25 and 0.5 are 63.42%, 61.61% and 60.22%, respectivly.

5.7 GESVM Regression (C++ Implementation)

The GESVM implementations above were designed for case/control data, utilizing the binary classification power of SVMs. However, the ability to apply GESVM to continuous data greatly extends the applicability of the GESVM approach to a much wider range of phenotypic data. The C++ implementation of GESVM was modified to accommodate continuous phenotype data by using SVM regression rather than binary classification. This required a simple change to the algorithm used in libsvm with regard to computation, but required significant changes to the data handling functions. In the modified version, GESVM requires three files: one for phenotype data, one for covariate data and the third for genotype data. The phenotype and covariates are stored in the program's memory while the genotype data is read in each time an SVM structure is created. The SVM optimization problem for the training dataset and prediction for the testing dataset are again solved using libsvm [Chang & Lin, 2011]. Computations were performed using the Bioinformatics Research Center hardware at NCSU.



(a)



(b)

Figure 5.4 Frequency of SNPs included in the 5 top-performing SVMs for each of the 27 models: **a**) all SNPs, **b**) only causal SNPs. In **b**, solid bars represent frequency of top SVMs containing both causal SNPs while empty bars represent those containing one or the other.

5.7.1 Simulation Study

The parameter values C and γ were allowed to range from 0.0001 to 2 as before. GE parameters are reported in Table 5.1. The chromosome size was fixed at 25 codons and a one-point crossover was used. Selection was performed by keeping the proportion of the population that had the best fitness as determined by minimum mean squared error. Ten-fold cross-validation was used to reduce overfitting by splitting the data into different training and testing sets.

5.7.1.1 Data Generation

Genetic data for 20 probes from 500 samples in the ACCORD genotyped dataset were used to create simulated test data. Two SNPs with MAF > 0.3 and $\chi^2_{HWE} < 0.5$ were selected as causal variants (SNPs 8 and 15). Phenotype data was generated by sampling from a normal $\mathcal{N}(\mu, 1)$ distribution, with μ being a function of the genotypes g_8 and g_{15}, where $g_\bullet \in \{0,1,2\}$. The means were defined as $\mu(0,0) = 0$, $\mu(\{(0,1),(1,0)\}) = 3$, $\mu(\{(0,2),(1,2),(2,0),(2,1)\}) = 6$, $\mu(1,1) = 9$ and $\mu(2,2) = 12$. The resulting set of phenotypes were normalized and histograms of the corresponding phenotype values are shown in Fig. 5.5. Eight covariates were generated for each sample, 4 continuous and 4 binary. The continuous covariates were sampled from $\mathcal{N}(0,1)$ distribution while binary covariates were sampled from a binary distribution with probability $p = 0.5$.

5.7.1.2 Results

Fig. 5.6 depicts the number of individuals containing each of the 20 SNPs as a function of the GE generation. Non-causal SNPs are the faded colors while the two causal SNPs are the sold and dashed bright colors. Each color represents a different cross-validation run. By generation 15 there is a clear separation between the causal and non-causal SNPs. Fig. 5.7 shows the number of individuals containing each of the 20 SNPs in the final GE generation. The boxplots summarize the data across the 10 cross-validation iterations. The causal SNPs are highlighted in red. The simulated dataset for

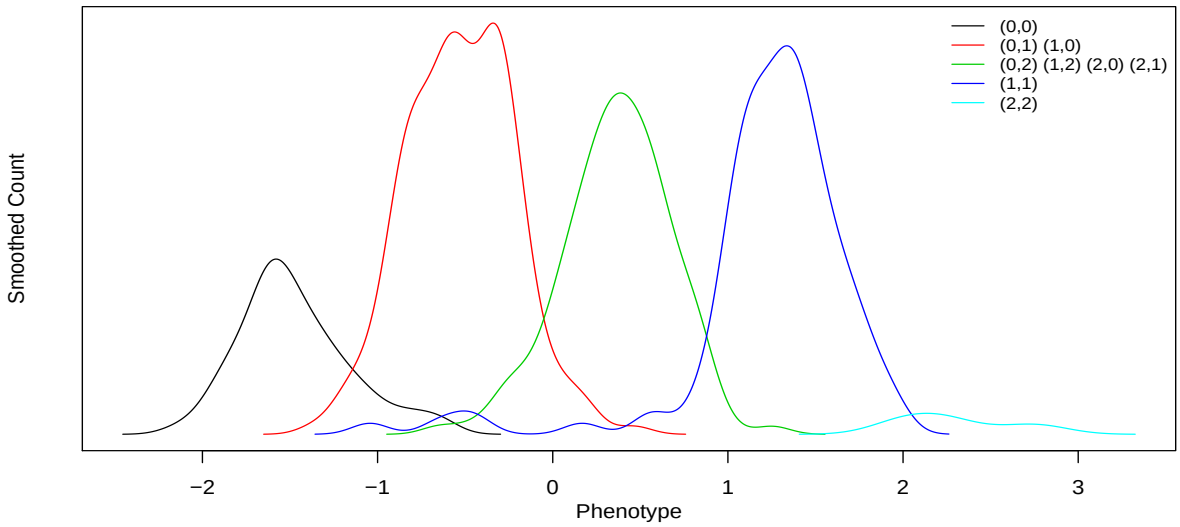


Figure 5.5 Smoothed histograms of normalized, simulated phenotypes for different combinations of causal loci genotypes.

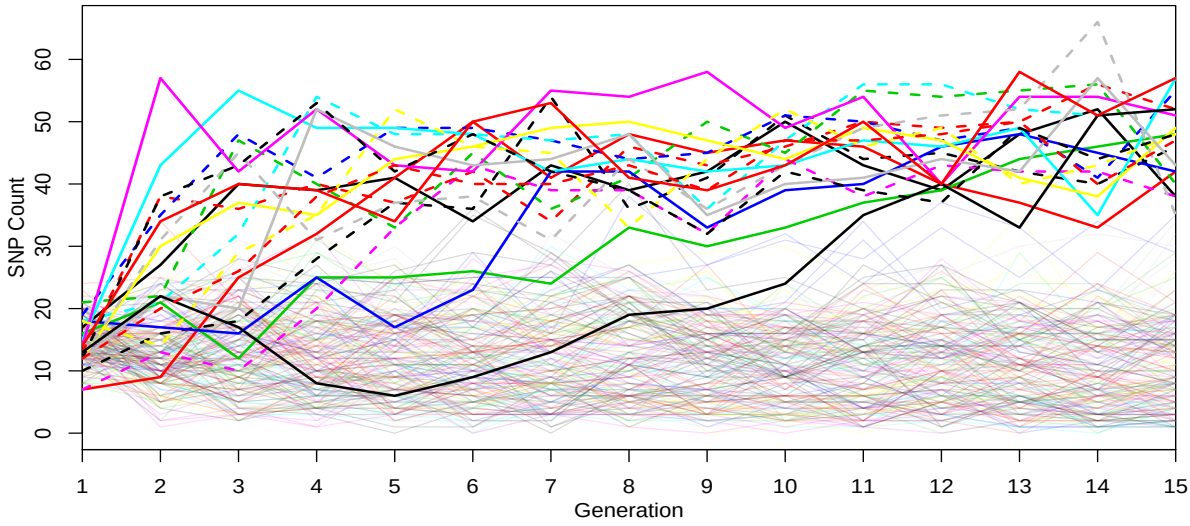


Figure 5.6 Number of individuals containing each of the 20 SNPs. Faded lines are non-causal SNPs. Bright solid and dashed lines indicate the two causal SNPs with the color representing the cross-validation iteration.

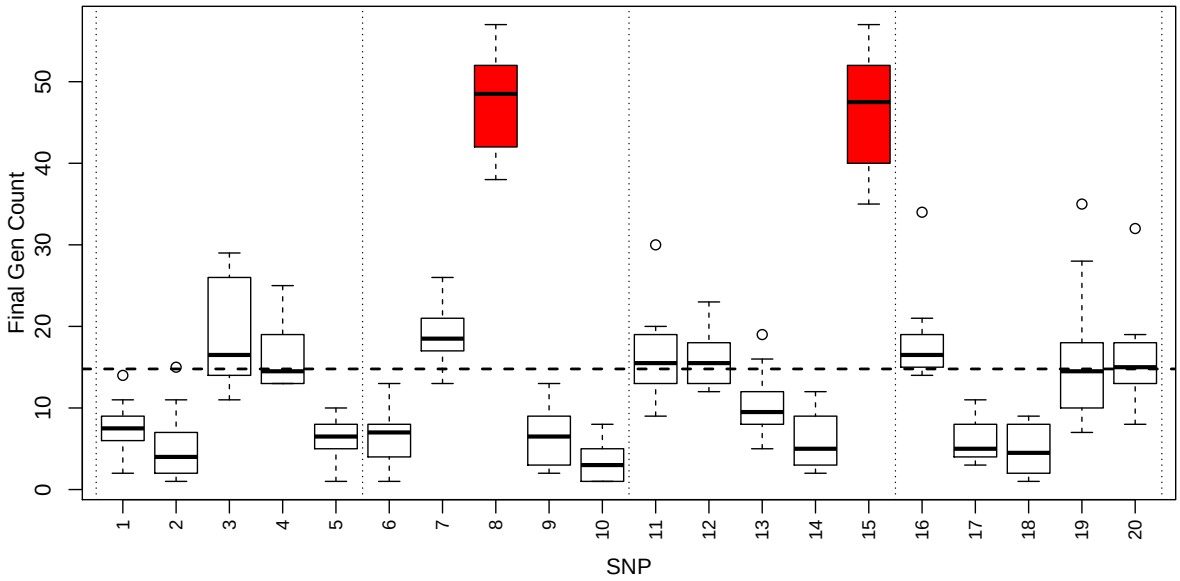


Figure 5.7 Boxplots for the number of individuals containing each of the 20 SNPs in the final GE generation. Each boxplot was created from the 10 values from cross-validation iterations. Causal SNPs are highlighted in red.

this study was created to show that GESVM is able to function as expected after the modifications to enable application to continuous phenotypes. Clearly GESVM is able to perform well with this simple dataset.

5.7.2 Application to ACCORD Dataset

The ACCORD genotype dataset was trimmed in order to reduce computation time for GESVM regression. A total of 10,782 variants with p-values less than 0.01 for any of the four phenotypes and MAF > 0.03 were extracted. These variants were when LD pruned (VIF=1.0526 $\rightarrow$ R^2=0.05) leaving 5,558 variants for GESVM to search. The GE parameters are similar to those for the simulation study in Table 5.1 except the population size was increased to 500 and the number of generations to 25. Initially GESVM was only applied to the HDL phenotype because it had the strongest signals and highest approximate heritability.

5.7.2.1 Results

The probe sets from the top 10 models based on lowest mean-squared error were recorded for each generation. Fig. 5.8 depicts the scaled mean count (averaged across cross validations) for the top models in the final GE generation (black dots). A background of the p-values from the HDL common variant analysis is used for comparison (blue and brown dots). The double layer of p-values are the result of extracting probes based on the p-values from any of the four phenotypes, i.e. the probes with p-values under a y-axis value of 2 were below 0.01 for one of the other three phenotypes.

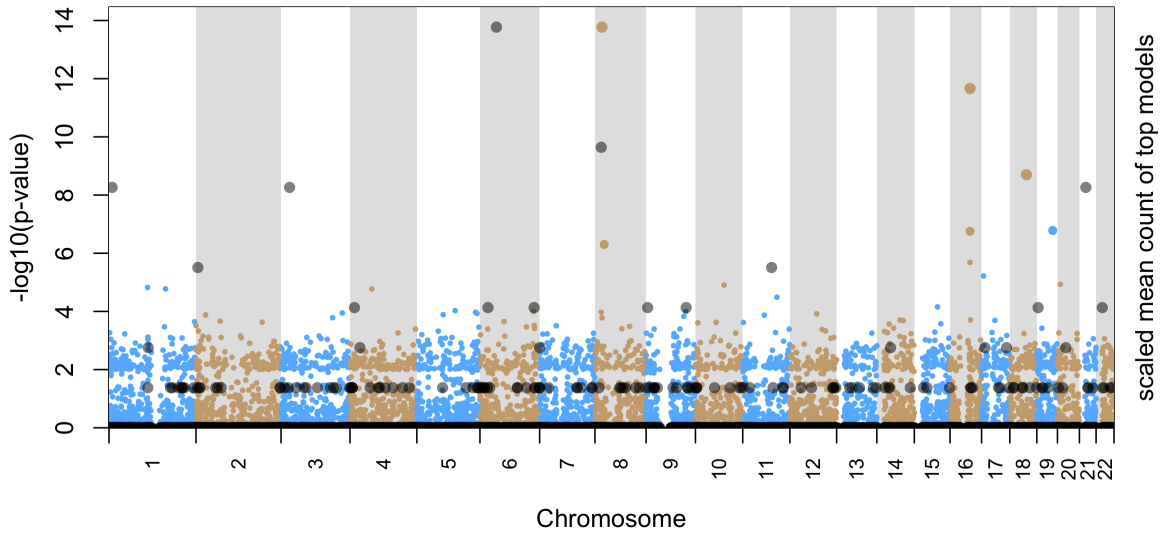


Figure 5.8 Mean count of variants in final generation of GESVM (black dots). P-values from the HDL common variant analysis are used as a background for comparison (blue and brown dots).

5.8 Discussion and Future Work

The development of GESVM is still in its infancy but has shown promising results from the initial proof of concept study with case-control data through the first steps of testing the GESVM regression implementation for continuous phenotypes. Parameter sweeps demonstrated that GESVM is able to correctly include both causal loci in about 50/100 datasets generated using a heritability of 5%, MAF of 10% and sample size of 900 cases and controls. The application of GESVM regression on a simulated dataset based on a subset of the ACCORD genotype data validated that the implementation is working as desired, however, much more work is necessary to fully develop GESVM regression.

When GESVM regression was applied the actual ACCORD data using a pruned genotype dataset a few variants occurred with higher frequency in the top GESVM regression models, however, they were not consistent across cross validations, which suggests each run was getting stuck in local minima. In addition, determining the significance of the more frequent variants would require permutation testing to compute p-values. In this approach, a model is forced to only contain those variables which appear in the top models while phenotype and covariate data is permuted and the model fitness is evaluated. P-values are computed based on the resulting distribution of fitness values.

The local minima problem could be addressed by increasing the number of generations and population size in the GE parameters. Additional modifications could be made, such as attempts to increase the mixing of variant sets by forcing variants into the classifiers in such a way that every variant is in at least one model at some point in each cross validation run. Clearly more work needs to be done before GESVM is applicable to real, large datasets. GE parameter sweeps should be done using simulated datasets based on the ACCORD genotype data to better select parameter values. As the number of generations and population size increases, GESVM would benefit greatly from parallelization where, for example, each GE individual could use a GPU core.

In addition, the fitness metric of GESVM regression should be evaluated for appropriateness. The most significant variants from the common variant analysis should also be the most fit when

GESVM models are limited to a single variant. This idea could be used to evaluate a variety of fitness functions. Another approach may be to allow GE to select a fitness function in addition to the classifier parameters and genetic variants. GE could also be expanded to selecting the SVM kernel rather than always using the RBF. After developmental changes have been implemented, the performance of GESVM will need to be compared to methods currently in the literature using an array of simulated datasets where the performance metrics would include key metrics like computation time and power to detect causal variants.

Chapter

6

Conclusions

The ACCORD clinical trial was designed to be the definitive answer on targeting normal levels of blood glucose, blood pressure and blood lipids in patients with type 2 diabetes. Results of the trial were disappointing, but evidence of heterogeneity of response encouraged the initiation of genetic studies. Samples from those patients who consented to genetic studies were genotyped using a new exome array that consists of roughly 250k genome-wide association markers, 250k exome markers focusing on rare variants, 70k novel loss-of-function markers, 20k eQTLs and 2k pharmacogenomic markers. The amount of data from the ACCORD clinical trial is extensive and will be used in many genome-wide association studies (GWASs) in the years to come. Great care to provide a high-quality genotype dataset is paramount to give the greatest chance of finding true significant associations in these GWASs.

A QC pipeline was implemented following best practices guidelines. This pipeline consists of custom Perl, Bash, and R scripts written to allow for a large amount of automation and east adaptation for future datasets. The pipeline used a variety of pre- and post-genotype calling filters to remove plates, samples and probes. A total of 90 96-well plates were genotyped with each plate containing 3 sets of duplicate samples and 2 HapMap samples. Duplicate samples were used to compute intra- and inter-plate concordance values. The HapMap samples were also used to monitor concordance with known genotypes and allowed for detection of Mendellian trio errors. Only 1.5% of the genotyped ACCORD samples and 7.1% of the genotype probes were removed due to poor quality, resulting in a high-quality dataset consisting of 8054 samples and 583,613 variants. This dataset was pre-phased to accelerate imputation as ever larger reference panels become available. In addition, functional variant sets were created for rare variant analysis.

An analysis plan was streamlined to ensure consistency and reduce the overall time of association studies using the ACCORD genotype dataset. This was accomplished by developing a highly-automated pipeline that accounts for standard issues in association studies such as population structure, cryptic relatedness, covariate selection and multiple-hypothesis testing. The analysis pipeline was applied to baseline lipid levels partially to validate the quality control and imputation pipelines

but also to identify potentially novel variants. All genotyped ACCORD samples were used in the baseline analysis. Of the initial 8054, 7844 remained after removal of those with incomplete phenotype/covariate data and one sample from each pair with a kinship > 0.1768. Both common and rare variant analyses were conducted, with the common variant analysis consisting of all variants with a MAF $> 3\%$ and the rare variant analysis consisting of all functional variants with a MAF $< 3\%$ grouped by gene. Genomic inflation was nearly 1.00 for all four phenotypes. There were 11 loci identified in the common variant analysis and 11 genes identified in the rare variant analysis. All significant common variants and half of the rare variant analysis genes have previously been reported in the literature, which is a good indication that the pipelines are working well. In addition, several novel genes were identified in the rare variant analysis.

A novel computational approach for detecting gene-gene interactions was developed to address the issue of missing heritability in GWASs. This method uses grammatical evolution (GE) to select features and classifier architecture to construct a support vector machine (SVM) classifier. GESVM was initially developed for case-control data and was first implemented in MATLAB as a proof of concept using two simple penetrance functions to create simulated datasets. GESVM was then implemented in C++ and datasets were simulated that consisted of a range of heritability, minor allele frequency of causal variants, and sample sizes. GESVM was then modified to work with continuous phenotype data using SVM regression and tested on a simple dataset with phenotype and covariate data derived based on a subset of ACCORD genotype data. Initial simulation studies look promising but application to a real dataset revealed that additional development is necessary.

Future work for the analysis pipeline is to identify treatment regimens and develop drug scoring schemes. In addition, different types of rare variant sets focusing on other genetic aspects, e.g. transcript binding sites and non-coding RNA, should be studied. Further work on GESVM development would include additional parameter sweeps, testing more complex simulated datasets, parallization, and comparisons to other methods. The work in this dissertation lays the foundation for multiple papers from the ACCORD cohort.

BIBLIOGRAPHY

Asimit, J. & Zeggini, E. (2010). "Rare variant association analysis methods for complex traits". *Annual review of genetics* **44**, pp. 293–308.

Aslibekyan, S. et al. (2013). "Preliminary evidence of genetic determinants of adiponectin response to fenofibrate in the Genetics of Lipid Lowering Drugs and Diet Network". *Nutrition, Metabolism and Cardiovascular Diseases* **23**.10, pp. 987 –994.

Axiom biobank genotyping arrays. White Paper, P/N DNA01801 Rev 4. Affymetrix.

Axiom genotyping solution analysis guide (2011). 702961 Rev 1. Affymetrix.

Bansal, V. et al. (2010). "Statistical analysis strategies for association studies involving rare variants". *Nature Reviews Genetics* **11**.11, pp. 773–785.

Benjamini, Y. & Hochberg, Y. (1995). "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society. Series B (Methodological)* **57**.1, pp. 289–300.

Best practice supplement to axiom genotyping solution data analysis user guide. P/N 703083 Rev 1. Affymetrix.

Bonds, D. E. et al. (2007). "Severe Hypoglycemia Monitoring and Risk Management Procedures in the Action to Control Cardiovascular Risk in Diabetes (ACCORD) Trial". *The American Journal of Cardiology* **99**.12, Supplement, S80 –S89.

Bonferroni, C. E. (1936). "Teoria statistica delle classi e calcolo delle probabilitá". *Pubblicazioni del Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8**, pp. 3–62.

BRLMM-P: a Genotype Calling Method for the SNP 5.0 Array (2007). White Paper. Affymetrix.

Buse, J. B. (2007). "Action to Control Cardiovascular Risk in Diabetes (ACCORD) Trial: Design and Methods". *The American Journal of Cardiology* **99**.12, Supplement, S21 –S33.

Cannon, C. P. et al. (2004). "Intensive versus Moderate Lipid Lowering with Statins after Acute Coronary Syndromes". *New England Journal of Medicine* **350**.15, pp. 1495–1504.

Capriotti, E. et al. (2006). "Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information". *Bioinformatics* **22**.22, pp. 2729–2734.

Chang, C.-C. & Lin, C.-J. (2011). "LIBSVM: A Library for Support Vector Machines". *ACM Trans. Intell. Syst. Technol.* **2**.3, 27:1–27:27.

Chen, S.-H. et al. (2008). "A support vector machine approach for detecting gene-gene interaction". *Genetic Epidemiology* **32**.2, pp. 152–167.

Chew, E. Y. et al. (2007). "Rationale, Design, and Methods of the Action to Control Cardiovascular Risk in Diabetes Eye Study (ACCORD-EYE)". *The American Journal of Cardiology* **99**.12, Supplement, S103 –S111.

Consortium, G. L. G. (2013). "Discovery and refinement of loci associated with lipid levels". *Nature Genetics* **45**.11, pp. 1274–1283.

Cordell, H. J. (2002). "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans". *Human Molecular Genetics* **11**.20, pp. 2463–2468.

Culverhouse, R. et al. (2002). "A Perspective on Epistasis: Limits of Models Displaying No Main Effect". *The American Journal of Human Genetics* **70**.2, pp. 461 –471.

Cushman, W. C. et al. (2007). "Rationale and Design for the Blood Pressure Intervention of the Action to Control Cardiovascular Risk in Diabetes (ACCORD) Trial". *The American Journal of Cardiology* **99**.12, Supplement, S44 –S55.

Cushman, W. C. et al. (2010). "Effects of Intensive Blood-Pressure Control in Type 2 Diabetes Mellitus". *New England Journal of Medicine* **362**.17, pp. 1575–1585.

Dabney, A. & Storey, J. D. *qvalue: Q-value estimation for false discovery rate control*. R package version 1.36.0.

Dai, H. D. et al. (2014). "A Modified Generalized Fisher Method for Combining Probabilities from Dependent Tests". *Frontiers in Genetics* **5**.32.

Davies, R. B. (1980). "Algorithm AS 155: The Distribution of a Linear Combination of χ^2 Random Variables". English. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **29**.3, pp. 323–333.

Delaneau, O. et al. (2012). "A linear complexity phasing method for thousands of genomes". *Nature methods* **9**.2, pp. 179–181.

Delaneau, O. et al. (2013). "Improved whole-chromosome phasing for disease and population genetic studies". *Nature methods* **10**.1, pp. 5–6.

Dering, C. et al. (2011). "Statistical analysis of rare sequence variants: an overview of collapsing methods". *Genetic Epidemiology* **35**.S1, S12–S17.

Devlin, B. & Roeder, K. (1999). "Genomic Control for Association Studies". *Biometrics* **55**.4, pp. 997–1004.

Dudek, S. M. et al. (2006). “Data simulation software for whole-genome association and other studies in human genetics”. *Pacific Symposium on Biocomputing*, pp. 499–510.

Edwards, T. et al. (2008). “Generating Linkage Disequilibrium Patterns in Data Simulations Using genomeSIMLA”. English. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Ed. by Marchiori, E. & Moore, J. Vol. 4973. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 24–35.

Gerstein, H. C. et al. (2007). “Glycemia Treatment Strategies in the Action to Control Cardiovascular Risk in Diabetes (ACCORD) Trial”. *The American Journal of Cardiology* **99**.12, Supplement, S34 –S43.

Gerstein, H. C. et al. (2008). “Effects of Intensive Glucose Lowering in Type 2 Diabetes”. *New England Journal of Medicine* **358**.24, pp. 2545–2559.

Ginsberg, H. N. et al. (2007). “Evolution of the Lipid Trial Protocol of the Action to Control Cardiovascular Risk in Diabetes (ACCORD) Trial”. *The American Journal of Cardiology* **99**.12, Supplement, S56 –S67.

Ginsberg, H. N. et al. (2010). “Effects of Combination Lipid Therapy in Type 2 Diabetes Mellitus”. *New England Journal of Medicine* **362**.17, pp. 1563–1574.

Goff Jr., D. C. 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, Supplement, S4 –S20.

Gorlov, I. P. et al. (2008). “Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms”. *The American Journal of Human Genetics* **82**.1, pp. 100 –112.

Graffelman, J. & Camarena, J. M. (2007). “Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot”. *Human Heredity* **65**.2, pp. 77–84.

Heart Protection Study Collaborative Group (2002). “MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20âĂŁ536 high-risk individuals: a randomised placebocontrolled trial”. *The Lancet* **360**.9326, pp. 7 –22.

Hopewell, J. C. et al. (2013). “Impact of common genetic variation on response to simvastatin therapy among 18 705 participants in the Heart Protection Study”. *European Heart Journal* **34**.13, pp. 982–992.

Howie, B. et al. (2009). “A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies”. *PLoS Genet* **5**.6, e1000529.

Howie, B. et al. (2011). "Genotype Imputation with Thousands of Genomes". *G3: Genes, Genomes, Genetics* **1**.6, pp. 457–470.

Jack, L. & Nandi, A. (2002). "Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms". *Mechanical Systems and Signal Processing* **16**.2-3, pp. 373–390.

Kathiresan, S. et al. (2007). "A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study". *BMC Medical Genetics* **8**.Suppl 1, S17.

Lancaster, H. O. (1961). "The combination of probabilities: an application of orthonormal functions". *Australian Journal of Statistics* **3**.1, pp. 20–33.

Lange, L. 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, pp. 233 –245.

LaRosa, J. C. et al. (2005). "Intensive Lipid Lowering with Atorvastatin in Patients with Stable Coronary Disease". *New England Journal of Medicine* **352**.14, pp. 1425–1435.

Laurie, C. C. et al. (2010). "Quality control and quality assurance in genotypic data for genome-wide association studies". *Genetic Epidemiology* **34**.6, pp. 591–602.

Lee, S. et al. (2012). "Optimal tests for rare variant effects in sequencing association studies". *Biostatistics* **13**.4, pp. 762–775.

Lee, S. et al. (2014a). *SKAT: SNP-set (Sequence) Kernel Association Test*. R package version 0.95.

Lee, S. et al. (2014b). "Rare-Variant Association Analysis: Study Designs and Statistical Tests". *The American Journal of Human Genetics* **95**.1, pp. 5 –23.

Li, B. & 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, pp. 311 –321.

Liu, D. J. et al. (2014). "Meta-analysis of gene-level tests for rare variant association". *Nature genetics* **46**.2, pp. 200–204.

Madsen, B. E. & Browning, S. R. (2009). "A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic". *PLoS Genet* **5**.2, e1000384.

Manichaikul, A. et al. (2010). "Robust relationship inference in genome-wide association studies". *Bioinformatics* **26**.22, pp. 2867–2873.

Manolio, T. A. (2010). "Genomewide Association Studies and Assessment of the Risk of Disease". *New England Journal of Medicine* **363**.2, pp. 166–176.

Manolio, T. A. et al. (2009). "Finding the missing heritability of complex diseases". *Nature* **461**.7265, pp. 747–753.

Marchini, J. & Howie, B. (2010). "Genotype imputation for genome-wide association studies". *Nat Rev Genet* **11**.7, pp. 499–511.

Mathieson, I. & McVean, G. (2012). "Differential confounding of rare and common variants in spatially structured populations". *Nature genetics* **44**.3, pp. 243–246.

Mefford, J. & Witte, J. S. (2012). "The Covariate's Dilemma". *PLoS Genet* **8**.11, e1003096.

Middelberg, R. et al. (2011). "Genetic variants in LPL, OASL and TOMM40/APOE-C1-C2-C4 genes are associated with multiple cardiovascular-related traits". *BMC Medical Genetics* **12**.1, p. 123.

Morgenthaler, S. & Thilly, W. G. (2007). "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST)". *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **615**.1-2, pp. 28–56.

Morris, A. P. & Zeggini, E. (2010). "An evaluation of statistical approaches to rare variant analysis in genetic association studies". *Genetic Epidemiology* **34**.2, pp. 188–193.

Motsinger-Reif, A. A. et al. (2008). "Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology". *Genetic Epidemiology* **32**.4, pp. 325–340.

O'Neill, M. & Ryan, C. (2003). *Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language*. Norwell, MA, USA: Kluwer Academic Publishers.

Peloso, G. et al. (2014). "Association of Low-Frequency and Rare Coding-Sequence Variants with Blood Lipids and Coronary Heart Disease in 56,000 Whites and Blacks". *The American Journal of Human Genetics* **94**.2, pp. 223 –232.

Pluzhnikov, A et al. (2008). "Differential bias in genotype calls between plates due to the effect of a small number of lower DNA quality and/or contaminated samples". *Genetic Epidemiology* **32**.7, pp. 676–676.

Price, A. L. et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies". *Nature genetics* **38**.8, pp. 904–909.

Purcell, S. *PLINK (v1.07)*. `http://pngu.mgh.harvard.edu/purcell/plink/`.

Purcell, S. et al. (2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". *The American Journal of Human Genetics* **81**.3, pp. 559 –575.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Saeys, Y. et al. (2007). "A review of feature selection techniques in bioinformatics". *Bioinformatics* **23**.19, pp. 2507–2517.

Samanta, B. et al. (2003). "Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection". *Engineering Applications of Artificial Intelligence* **16**.7-8, pp. 657–665.

Storey, J. D. (2002). "A direct approach to false discovery rates". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**.3, pp. 479–498.

Teslovich, T. M. et al. (2010). "Biological, clinical and population relevance of 95 loci for blood lipids". *Nature* **466**.7307, pp. 707–713.

The International HapMap Consortium (2005). "A Haplotype Map of the Human Genome". *Nature* **437**, pp. 1299–1320.

Tian, C. et al. (2008). "Accounting for ancestry: population substructure and genome-wide association studies". *Human Molecular Genetics* **17**.R2, R143–R150.

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer-Verlag New York Inc.

Voight, B. F. & Pritchard, J. K. (2005). "Confounding from Cryptic Relatedness in Case-Control Association Studies". *PLoS Genet* **1**, e32.

Waterworth, D. M. et al. (2010). "Genetic Variants Influencing Circulating Lipid Levels and Risk of Coronary Artery Disease". *Arteriosclerosis, Thrombosis, and Vascular Biology* **30**.11, pp. 2264–2276.

Wei, Z. et al. (2009). "From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes". *PLoS Genet* **5**.10, e1000678.

Williamson, J. D. et al. (2007). "The Action to Control Cardiovascular Risk in Diabetes Memory in Diabetes Study (ACCORD-MIND): Rationale, Design, and Methods". *The American Journal of Cardiology* **99**.12, Supplement, S112 –S122.

Wu, M. et al. (2011). "Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test". *The American Journal of Human Genetics* **89**.1, pp. 82 –93.

Yang, J. et al. (2010). "Common SNPs explain a large proportion of the heritability for human height". *Nature genetics* **42**.7, pp. 565–569.

Yang, J. et al. (2011). "GCTA: A Tool for Genome-wide Complex Trait Analysis". *The American Journal of Human Genetics* **88**.1, pp. 76 –82.

Zhou, K. et al. (2011). "Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes". *Nature Genetics* **43**.2, pp. 117–120.