A BIOLOGICALLY INFORMED METHOD FOR DETECTING ASSOCIATIONS WITH RARE VARIANTS

By

Carrie Buchanan Moore

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

December 2013

Nashville, Tennessee

Approved:
Professor Marylyn Ritchie
Professor William Bush
Professor Bingshan Li
Professor Dan Roden
Professor Tricia Thornton-Wells

To three very important educators and mentors,

my mother, grandmother, and sister.

Thanks for your unwavering support and love.

ACKNOWLEDGEMENTS

The work presented in this thesis is the result of much collaboration, guidance, mentoring, and support from members of my thesis committee, coauthors of the manuscripts written along the way, and most importantly my Ph.D. mentor, Marylyn Ritchie. At regular intervals throughout my project, my thesis committee (Dr. Bingshan Li, Dr. Dan Roden, Dr. Tricia Thornton-Wells, Dr. William Bush, and Dr. Marylyn Ritchie) has listened earnestly to my ideas, evaluated my progress, contributed rich discussion, and willingly provided suggestions to improve my work. This project has benefitted immensely from their advice and input.

I would like to especially thank my Ph.D. mentor, Dr. Marylyn Ritchie, for providing a training environment where ideas can be challenged and personal and professional growth are valued. She allowed me opportunities to give presentations at international conferences, encouraged me to form productive scientific collaborations, and has been a steadfast source of support in all of my personal and academic endeavors. I would also like to recognize my co-mentor, Dr. William Bush. I am thankful he readily accepted responsibility for my training when Dr. Ritchie accepted a position at Pennsylvania State University. His brilliant creative thinking and curiosity led me to often seek his advice throughout various stages of this project. Finally, I would like to thank both of my mentors for taking the time a foster a positive mentoring relationship throughout my training. I greatly respect each of their careers and leadership styles. From job searches to promotions, grants and professional peer relationships, I hope to call on them frequently for future counsel.

I would also like to specifically thank several members of Dr. Ritchie's lab. First, I would like to acknowledge John Wallace. He is an outstanding computer scientist and mathematician. He has been critical to the success of this project, provided new insight and ideas, and has been willing to try almost anything as we worked together to build BioBin. I would also like to acknowledge Alex Frase, whose expertise was essential to restructure Biofilter 1.0 to LOKI and Biofilter 2.0. Together, John Wallace and Alex Frase have assisted me with various aspects of scripting and algorithmic concepts. On many occasions, they have helped me resolve errors and improve efficiency. I am incredibly grateful

for their expertise, willingness to chase new ideas, and patience to teach. I would also like to acknowledge Dr. Emily Holzinger, now a Ritchie lab alumni. Throughout our graduate careers, Dr. Holzinger has been willing to enthusiastically debate science and share incredible knowledge to further my projects and ideas. As a friend, we have celebrated many milestones, including my wedding and her graduation. We have developed a friendship that can grow small ideas into big projects and continually challenge and support one another.

I would also thank other key members of the Ritchie lab at Pennsylvania State University, Center for Systems Genomics, and Center for Human Genetics Research at Vanderbilt University. At PSU, I would like to acknowledge Daniel Wolfe, Sarah Pendergrass, Anurag Verma, and Shefali Verma. Dan has helped with many BioBin analyses, particularly contributing to the work described in Chapter V. Sarah tirelessly helped revise the manuscript from which Chapter V is derived. Dan, Anurag, and Shefali have helped me strengthen my skills as a teacher and learner. They have challenged me with questions about scientific ideas and shared many of their own. Finally, I would like to specifically thank the other graduate students in the CHGR, particularly Rafal Sobota, for being helpful, always insightful, and unafraid to enjoy life.

This work was financially supported by the following federal grants: LM010040, NS066638-01, HG004608, HL065962, 5T32GM080178, F30AG041570 from the National Institute on Aging, Public Health Service award T32 GM07347 from the National Institute of General Medical Studies for the Medical-Scientist Training Program, and Pennsylvania Department of Health using Tobacco CURE Funds.

Data accessed in this thesis are part of the NHLBI Grand Opportunity Exome Sequencing Project (GO-ESP). The Next Generation Mendelian Genetics project was provided by NIH grant 1RC2 HG005608-01 to Drs. Debbie Nickerson, Jay Shendure, Michael Bamshad, and Wendy Raskind, and research on Kabuki Syndrome by 5RO1- HD48895 to Michael Bamshad. The dataset(s) used for the analyses described in this manuscript were obtained from the database of Genotype and Phenotype (dbGaP) found at `http://www.ncbi.nlm.nih.gov/gap` through dbGaP accession number phs000295.v1.p1.

Funding for GO-ESP was provided by National Heart, Lung, and Blood Institute (NHLBI) grants RC2 HL103010 (HeartGO), RC2 HL102923 (LungGO) and RC2 HL102924

(Women's Health Initiative Sequencing Project [WHISP]). The exome sequencing was performed through NHLBI grants RC2 HL102925 (BroadGO) and RC2 HL102926 (SeattleGO). The dataset(s) used for the analyses described in this manuscript were obtained from the database of Genotype and Phenotype (dbGaP) found at `http://www.ncbi.nlm.nih.gov/gap` through dbGaP accession number phs000254.v2.p1.

Sincere thanks to the anonymous individual participants in these dbGaP studies for their contribution and willingness to participate in research.

LIST OF FIGURES

Figure		Page
1	Rate of approach to linkage equilibrium with regard to the recombination frequency between the genes	18
2	Pipeline for BioBin analysis	27
3	BioBin resource requirements with varying the population sizes and numbers of study variants	28
4	Example binning strategies using biological knowledge	36
5	P-value distribution under three different MAF binning thresholds to test RCC option	39
6	Bin dependency screen output	43
7	Quantile-quantile plots for type I error simulation studies in continuous regions	52
8	Quantile-quantile plots for type I error simulation studies using grouped regions	55
9	Correlation between bin p-value and number of variants in the bin	57
10	Estimate graphs of variance (w_i) and locus weight ($\frac{1}{w_i}$) for varying allele frequencies	59
11	Power estimates for multiple binning strategies on simulated data with 100 cases and 300 controls where only 25% of variants are functional	62
12	Power estimates for multiple binning strategies on simulated data with 100 cases and 300 controls where only 50% of variants are functional	63
13	Power estimates for multiple binning strategies on simulated data with 100 cases and 300 controls where 100% of variants are functional	64
14	Minor allele frequency distribution on chromosome 1 in 1000 Genomes Project Phase I populations	70
15	Investigating differential bias in 1000 Genomes Project data using principal components analysis	73
16	IBD estimates using variants with MAF $> 10\%$ within and between ancestral groups	75
17	Within population identity-by-state (IBS) estimations: A) before and B) after removing individuals with cryptic relatedness	77
18	Pairwise IBS calculations for low frequency variants (MAF $< 3\%$) within continental groups	79
19	Pairwise IBS calculations for common variants (MAF $> 25\%$) within continental groups	80
20	Proportion of significantly different bins in: A) gene exon, B) gene intron, and C) intergenic regions	83
21	Proportion of significantly different bins for gene exon filters: A) nonsynonymous and B) predicted deleterious variants	84
22	Proportion of significantly different bins in: A) ORegAnno regulatory and B) pathway feature analysis	85
23	Proportion of significantly different bins for the pathway-exon feature analysis	87
24	Proportion of significantly different bins in evolutionary conserved region feature analysis: A) conserved with primates, B) conserved with mammals, and C) conserved with vertebrates	89

25 Proportion of significantly different bins in natural selection analysis by region of identification: A) AFR continental group, B) ASN continental group, and C) EUR continental group 91
26 Proportion of loci in top bins in high LD with other variants in the same bin 96
27 Investigation of pathway significant correlation with bin size using untransformed pathway variables 98
28 Investigation of pathway significant correlation with bin size using $\log_{10}$ transformed pathway variables 100
29 Pathway characteristics presented by LOKI source 101
30 Principal component analysis (PCA) using merged samples from the CF analysis and 1000 Genomes Project Phase I data to identify ancestry 114
31 Correlation matrix for all variables considered in PA analysis 117
32 Correlation matrix for highly correlated variables in the PA analysis 118
33 Quantile-quantile plots for pulmonary function (PF) analysis illustrating effect of perfect separation in PF data 123
34 Quantile-quantile plots for *Pseudomonas aeruginosa* infection (PA) analysis 126
35 Gene set network for one of the top models 132

LIST OF TABLES

Table		Page
1	Example of linkage disequilibrium	16
2	Examples of r^2 and age of mutation adapted from Hartl and Clark [1].	19
3	Maximum r^2 and expected odds ratios (OR) between a rare causal variant and common genotyped SNP with an odds ratio of 1.1	20
4	Example phenotype input file	31
5	Example bins report output file	32
6	Example locus report output file	33
7	Custom region feature file	35
8	Custom region feature file	37
9	Allele selection and variant binning using rare-case-control (RCC) and overall-major-allele (OMA) parameters	42
10	Type I error simulation results from continuous region simulation studies	50
11	Type I error simulation results from group simulation studies	53
12	Power analysis using simulations with 4000 replicates for each of four sample sizes (N=2000, 1000, 500, and 250)	60
13	Excerpt of custom region file containing regions with signatures of natural selection	67
14	1000 Genomes Project Phase I data characteristics	69
15	Phase I 1000 Genomes Project sequence technology data characteristics	71
16	Analyses performed for each population comparison	90
17	Genes identified in Barreiro study between CEU/CHB and CEU/YRI population comparisons with an F_{ST} value > 0.65 that were also found in the regions identified by Pritchard, Stoneking, and Grossman	93
18	Genes identified in Barreiro study in YRI-CHB comparison with an F_{ST} value > 0.65 that overlapped with the regions identified by Pritchard, Stoneking, and Grossman	94
19	Gene results for specific genes of interest with known allele frequency differences between ancestral populations	95
20	Data characteristics for cystic fibrosis study sample for 416 European descent individuals	113
21	dbGaP cystic fibrosis clinical and demographic characteristics for two studies: recurrent pseudomonas infection (PA) and mild/severe pulmonary phenotype (PF)	115
22	Top results from PF analysis using no variant weights and binning only non-synonymous variants	124
23	Top results from PA analysis using no variant weights and binning only non-synonymous variants	128

TABLE OF CONTENTS

Page

DEDICATION iii

ACKNOWLEDGEMENTS iv

LIST OF FIGURES vii

LIST OF TABLES ix

Chapter

I INTRODUCTION **1**
- Terminology 5

II BACKGROUND **9**
- Contribution of low frequency variants to disease heritability 9
- Evolution of low frequency variants 11
  - Population expansion 12
  - Effects of selection 13
  - Evolutionary models 14
  - Characteristics of low frequency variants 15
- Current methods to analyze low frequency variation 20
  - Burden tests 23
  - Nonburden tests 24
- Conclusions 25

III BIOBIN SOFTWARE **26**
- BioBin resource requirements 27
- Library of Knowledge Integration database (LOKI) 29
- BioBin software overview 30
  - Input files 30
  - Output files 31
- BioBin software features 34
  - Customized knowledge 34
  - Multi-level feature binning 35
  - Filtering strategies 37
  - Locus selection 37
  - Optional inheritance patterns 41
  - Bin dependency 41
  - Variant weighting 43
- Statistical tests 45
- Summary 45

IV SIMULATION STUDIES **47**
Type I error assessment . . . 48
Continuous region simulation . . . 48
Pathway simulation studies . . . 51
Correlation between significance and bin size . . . 54
Non-bias variant weighting comparison . . . 56
Power simulation assessment . . . 60
Varying sample size . . . 60
Unequal sample size and comparison with other methods . . . 61
Summary . . . 65

V 1000 GENOMES PROJECT DATA: POPULATION COMPARISON OF LOW FREQUENCY BURDEN **66**
Methods and results . . . 66
Binning approach . . . 66
Statistical analysis . . . 67
1000 Genomes Project data . . . 67
Investigation of allele sharing . . . 72
Genomic feature exploration . . . 78
Discussion . . . 102
1000 Genomes Project data . . . 102
Investigation of allele sharing . . . 102
Genomic feature exploration . . . 103
Conclusion . . . 107

VI BIOBIN ANALYSES IN NATURAL DATA **109**
Kabuki analysis . . . 109
Study sample . . . 109
Methods and results . . . 110
Conclusions . . . 111
Cystic fibrosis analysis . . . 112
Study sample . . . 112
Methods . . . 116
Results and discussion . . . 121
Pathway and elastic net analyses . . . 130
Conclusions . . . 131

VII CONCLUSIONS **135**
Evaluation of the analyses presented . . . 136
Strengths of this approach . . . 136
Limitations of this approach . . . 137
Binning considerations . . . 138
Study design . . . 138
Replication in rare variant analyses . . . 140
Future improvements to BioBin . . . 143
Future of rare variant analyses . . . 144
Data integration and complex modeling . . . 144
Summary . . . 146

BIBLIOGRAPHY **147**

CHAPTER I

INTRODUCTION

In the past five years, the genomics field has generated a prolific amount of sequence data. Groups such as the 1000 Genomes Project and Complete Genomics have pioneered next-generation data pipelines, including study design, data generation, variant calling, and quality control of output data. These technological advances and collaborative projects have made the study of low frequency variants increasingly achievable. It has become possible to look beyond common variant polymorphisms typical of genome-wide association studies (GWAS) and potentially explain additional trait variance using rare or low frequency variants (defined here as minor allele frequencies less than 1% and less than 5%, respectively).

Yet, even with increased data availability, progress toward understanding genomic variation and its association to common human disease lags behind. Scientists are hindered in exploiting these laboratory advances because strategies for analyzing these data to utilize their maximal potential are underdeveloped. In fact, the wealth of available data has made distinguishing true scientific discoveries from the thousands of false discoveries even more challenging. The growing disparity in rapidly advancing data collection vs. slowly developing data analysis methods mandates a more concerted research effort to develop the necessary analytical tools to successfully interpret the genotypic and biologic data. Successful analyses will ultimately improve the prevention, diagnosis, and treatment of common disease. Research that meets this critical challenge will include developing methods to analyze the data and developing pipelines to integrate low frequency data from sequencing with other "-omic" measures.

The study of low frequency variation on a genome-wide scale has been minimal prior to the next-generation sequencing era. Due to the infancy of this research, none of the currently available analytical methods are accepted as the "gold standard." Previously developed pipelines and tools used in GWAS are largely ineffective because rare variants have low

r^2 values and cannot be detected using a tag-SNP approach. According to the literature, low frequency variants have larger effect sizes than common variant associations, are much more prevalent than common variants, and have a higher proportion of nonsynonymous variation. Therefore, they require special consideration when developing analytical tools to study disease association [2, 3, 4].

Since low frequency variants are individually uncommon, large sample sizes are needed to ensure that multiple copies of a variant of interest can be sampled [5, 6]. The study design and cost of sequencing can make the required sample size prohibitive, particularly as the minor allele frequency decreases below 1%. To increase the composite allele frequency and analyze smaller sample sizes, collapsing methods can be utilized. Commonly referred to as burden tests, variants in a specific genetic region can be binned into a single genetic variable, which is then used for analysis [5]. An alternate collapsing strategy to the burden test compares distributions of variants across the trait of interest. Nonburden tests do not assume all variants binned together are causal or have the same direction of effect and can model rare variant epistasis. Variations of burden and nonburden tests are described in Chapter II. While nonburden tests are more powerful in cases with both protective and deleterious variants or many noncausal variants, they are less powerful than burden tests if a large proportion of binned variants in the same direction are truly causal or if the sample size is relatively small [7].

Previous collapsing strategies have focused on a particular statistical test in a predefined region rather than how to best group variants in informative regions. Agnostic or uninformed binning approaches can often lead to a decrease in power when there are variants with different directions of effect or too many neutral variants that mitigate the signal. The most successful collapsing method groups variants likely to have an impact on the function of a specific gene or genomic unit and compares the variant distribution or composite genetic score distribution across the trait of interest.

The goal of this project is to address this major limitation of current rare variant association tests, uninformed binning. BioBin is a novel knowledge-guided collapsing method which focuses on bin generation rather than association testing. By generating meaningful bins with biologically related variants, the power of any statistical association method

increases. In addition, BioBin provides the framework to create interesting and complex hypotheses by allowing multi-level bin generation using prior biological knowledge. While the implementation facilitates burden tests most easily, BioBin is not coupled to any statistical test. Users are able to use a variety of association tests, burden or nonburden methods, and permutation strategies appropriate for the hypothesis and data being tested.

This thesis introduces the functionality of BioBin. First, the software is described in detail with an explanation of many novel parameters and options specific to BioBin. Second, extensive testing is presented using a variety of simulation parameters and a method comparison. Third, application of BioBin to natural data sets is described to identify rare variant burden differences between cases and controls.

Chapter II describes the genetic architecture of low frequency variants and the contribution of low frequency variants to Mendelian and common complex disease. Chapter II also discusses the evolution of low frequency variants and specifically why GWAS analysis pipelines fail in low frequency variant association tests. Lastly, the current state of low frequency variant analysis is examined by reviewing available computation tools and algorithms.

Chapter III details the development of BioBin and available options in the BioBin package. This chapter presents an introduction and overview of resource requirements for both BioBin and Library of Knowledge Integration (LOKI). Six software features are characterized: custom knowledge input, multilevel feature binning, filtering strategies, loci selection, optional inheritance patterns, and variant weighting. Lastly, statistical tests commonly used with BioBin are briefly described. Portions of this chapter were derived from "A Biologically Informed Method for Detecting Associations with Rare Variants" [8].

Chapter IV includes comprehensive type I error and power simulations under various conditions. Described evaluations include different allele frequency weights, statistical methods, and power comparisons between BioBin and other methods. First, type I error was evaluated using two continuous region simulations (simulating genes and pathways). Second, simulations to study type I error for pathway-type analyses which have the potential for dependent bins were performed. Lastly, correlation results between bin size and bin significance and a few simulated power assessments were addressed. This chapter was partially

adapted from a peer-reviewed manuscript, "BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge" [9].

Chapter V describes the application of BioBin to 1000 Genomes Project data. A manuscript describing this work is in preparation at PLOS genetics, "Low Frequency Variants, Collapsed Based on Biological Knowledge, Uncover Complexity of Population Stratification in 1000 Genomes Project Data." In addition, some of the text from "Using BioBin to Explore Rare Variant Population Stratification" was adapted for this chapter [10]. In order to reveal the magnitude of low frequency population stratification, Chapter V describes how pairwise population comparisons using the 1000 Genomes Project Phase I data were performed to investigate differences in low frequency variant burden across multiple biological features. Low frequency variant confounding is much more prevalent than one might expect, even within continental groups. The proportion of significant differences in low frequency variant burden is also dependent on the region of interest; for example, annotated regulatory regions showed fewer low frequency burden differences between populations than intergenic regions.

Chapter VI consists of two applications of BioBin on natural whole-exome data. BioBin was applied to two data sets available from dbGaP: Kabuki syndrome (10 individuals) and cystic fibrosis with chronic *Pseudomonas aeruginosa* infection (431 individuals). The Kabuki analysis description and results were adapted from "BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge" [9]. The Kabuki sample is too small for adequate power, but the cystic fibrosis data analysis identified several interesting genes for follow-up analyses.

In Chapter VII, the benefits and limitations of this approach are discussed. Chapter VII also includes a discussion of the future of low frequency variant binning analyses and how BioBin will be amenable to future improvements of genomic analyses. Finally, the fundamental considerations to advance genomic research are considered, particularly at the level of sequence data analysis.

Terminology

Allele

Humans inherit one allele from each parent at every locus, resulting in a diploid state. Each pair of alleles form the genotype at a particular locus. At the population level, allelic variation at a locus is measurable as the number of different alleles (polymorphism) present. If 99% of population is homozygous for TT at a genetic locus and 1% of population is heterozygous with genotype TG, there are two alleles at this locus in the population, the major allele is T and the minor allele is G.

Bin

Bins contain combinations of variants. Bins can be generated based on any number of similarities. In this thesis, bins are often generated based on genomic location, e.g. gene bins are formed based on the start and stop position of the gene. All of the variant loci present in this gene are binned together and analyzed as a single unit.

Burden test

In this thesis, collapsing methods that use burden-type statistical analysis are referred to as burden tests. Instead of retaining characteristics of each variant in the bin, the statistical tests use a single summative value for all variants in the bin for each individual. The simplest of these would be dichotomizing low-frequency variants for each individual (independent variable designated as 0 if individual does not contain a low frequency variant and 1 if the individual contains at least one low frequency variant).

Collapsing method

Collapsing methods refer to a general type of low frequency analysis, also called "binning" methods, where low frequency variants are combined together based on some similarity (genomic coordinates, functional significances, etc.) for analysis. The description of a collapsing method refers to how the low frequency variants are combined. In the literature

some collapsing methods also have a statistical test associated with the software. BioBin is a collapsing method with multiple algorithmic components but without any incorporated statistical test.

Gene

A molecular unit of heredity of a living organism. A gene is a defined region of DNA with transcription start and stop sites coding for messenger RNA.

Linkage

Genetic linkage describes the way in which two loci located in close proximity on a chromosome are often inherited together. Loci in close proximity are more likely to be inherited together. In contrast, loci located farther away from each other on the same chromosome are more likely to be separated during recombination, the process that recombines DNA during meiosis. The strength of linkage between two loci depends upon the distance, rates of recombination, and functional interaction (might affect viability of offspring) in that region of the chromosome. If two loci are in linkage equilibrium, they are inherited independently in each generation. If two loci are in linkage disequilibrium, alleles at each locus are inherited together more often than would be expected by random chance.

Locus

The location of a gene or particular sequence on a chromosome. The locus of a gene refers to the start and stop positions of that gene. The locus of a variant is the specific genomic coordinates of that variant in the genome (variant site). In this thesis, the term locus refers to a single genomic coordinate where a variant occurs. Therefore, a single locus can refer to the site of multiple allelic variants (polymorphic locus).

Mutation[11]

Change in genetic sequence that affects function. There are some that argue that the term mutation should refer to any change in sequence below 1% and all other changes should be called polymorphisms. However, this thesis only uses the term to describe variants with

known effects on function, regardless of allele frequency. This is more consistent with the definition of mutant from biology (wild-type versus mutant). The process of "mutation" (verb) introduces novel variants to the population, but for a variant itself to be called a mutation, it has to have known functional effects (harmful or protective). Literature from the field of human genetics typically focuses on mutations that have harmful effects due to the bias in research, which more often studies low frequency disease-causing variation rather than variation that leads to improved health or longevity.

Nonburden test

Collapsing methods that use statistical analyses that retain characteristics of each variant in the bin. These statistical tests often use vector-based methods to compare distributions of variants across the trait of interest rather than using a single summative value. Nonburden tests do not assume all variants binned together are causal or have the same direction of effect and can model rare variant epistasis.

Polymorphism

A relatively benign change in genetic sequence. The term single nucleotide polymorphism (SNP) is very popular in the literature and refers to more than one allele at a locus in a population. Polymorphisms can cause variable phenotypes, but these changes are unlikely to contribute to phenotypes that decrease fitness and have an allele frequency of an arbitrary threshold of at least 1%.

Variant

Any change in genetic sequence at a particular locus. The term variant is used as a very general and inclusive term relating to all genetic changes (single nucleotide changes, insertions, deletions, copy number variation changes, etc.) of any frequency and of any functional consequence (neutral, protective, damaging, or unknown). A single locus can have multiple variants, e.g., if G is the referent allele, $G \rightarrow C$ represents one variant and $G \rightarrow T$ represents a second variant at the same locus. In this thesis, the term variant typically refers to a single nucleotide allelic change of any frequency or functional consequence. Unless otherwise

specified, in the following chapters, "rare", "low frequency" and "common" variants refer to variants with minor allele frequencies $\leq 1\%$, $\leq 5\%$, and $> 5\%$ respectively.

CHAPTER II

BACKGROUND

Since pioneering observations of genetic and evolutionary properties were reported by Darwin and Mendel in 1859 and 1866, respectively, geneticists have been interested in uncovering the secrets of inheritance, patterns of selection, distinguishing genetic and non-genetic causes of traits or diseases, etiology of genetic diseases, and determining risk profiles in individuals harboring variation [12, 13]. Later, the discovery of linkage and epistasis by Bateson and autosomal recessive inheritance patterns of alcaptonuria by Garrod further evolved the field of genomics [14, 15]. Present day genomics research utilizes unprecedented technology and computational power, but the goal of uncovering disease associations and understanding inheritance is much the same. This task is complicated by the numerous types of genetic variation and genetic architecture, and interpretations of results are contingent on understanding the landscape of genetic variation. This thesis focuses on the analysis of low frequency variants; however, first characteristics of low frequency variants are reviewed. It is important to consider the potential ways low frequency variants contribute to complex disease, the evolution of low frequency variants, and how to translate these properties into tools for studying low frequency variation.

Contribution of low frequency variants to disease heritability

Genome-wide association studies (GWAS) focus on common variants that often miss valuable information about epistatic (gene-gene, GxG) and gene-environment (GxE) interactions, structural variants, and rare variants (RV) [16]. While researchers have been able to attribute almost 11,000 variants from 1657 publications to over 80 diseases and traits [17, 18], the estimated odds ratios for these variants are predominantly less than

1.5 and a variable but small fraction of the estimated heritability has been explained. For example, in a recent study of metabolic traits, Vattikuti et al. found previously published common variants to explain between 25% (HDL trait) and 80% (systolic blood pressure measurement) of estimated narrow sense heritability [19]. In the case of HDL and most other traits, large proportions of heritability have yet to be explained. In an effort to elucidate additional heritability and to take advantage of the new sequencing technology, many researchers are investigating, in particular, the effects of rare variants. Either because of sheer number of rare variants or because of the effects of weak selection, rare variants are thought to be more likely to be disease predisposing than common variants [20, 21]. It is believed that rare variants can act alone, in concert with other rare variants, or together with common variants. Bansal et al. describes many reasons rare variants likely influence disease susceptibility [22]:

1. The recent population expansion resulted in a large number of segregating and potentially functionally relevant rare variants.

2. Rare variants have been shown to be functional mutations in tumorigenesis.

3. There are many published examples of allelic heterogeneity (breast cancer: *BRCA1*, cystic fibrosis: *CFTR*).

4. Functional assays have been performed in vivo for multiple rare variants and have been shown to influence clinical phenotypes.

5. Rare variants have been associated with phenotypes in candidate gene studies.

One of the earliest and best characterized causal rare variant identifications occurred in the study of cystic fibrosis (CF) in 1989 [23]. Kerem et al. performed an extensive linkage analysis in CF patients with restriction fragment length polymorphisms to identify a single locus, chromosome 7q31, corresponding to the *CFTR* gene. In the original paper, the disease prevalence was stated to be 1/2000 live births in Caucasian populations, with a mutant allele frequency of 2.2% [24]. Since this original publication and resulting Δ508 mutation identification, over 1000 other causative mutations have been identified to cause

cystic fibrosis in the *CFTR* gene. Low frequency variants have been identified for several other Mendelian traits using linkage studies and most recently, next-generation sequencing. Approximately 1/2 to 1/3 of all known or suspected Mendelian diseases (approximately 7000) have been associated with a particular locus [25]. In the past, it has been difficult to fully resolve missing heritability because linkage studies failed when the disease was too rare, when too few family members were affected, when disease decreased reproductive fitness, when the disease exhibited reduced penetrance, or when locus heterogeneity was present. In addition, spontaneous instead of inherited mutations causing monogenic disorders were impossible to study using linkage analysis [25, 26]. Next generation sequencing provides resolution to a single base pair change and can be applied within pedigrees, across unrelated individuals, in trios, and from sampling individuals from phenotype extremes. The user can apply series of filters and deduce a list of potential mutations in a relatively short period of time. In fact, in less than two years (2010-2011) over 27 studies were published identifying rare variant loci for Mendelian traits/disorders [25].

Although dominant rare variants with large effect sizes (OR $>$ 5) generally correspond to Mendelian diseases with close to 100% penetrance, there is increasing evidence to support a role for rare variants to contribute to risk of common, complex disease. Recent studies have implicated rare variants with moderate effect sizes using phenotypes such as obesity, autism, schizophrenia, hypertriglyceridemia, hearing loss, complex I deficiency, type-1 diabetes, sporadic mental retardation, inflammatory bowel disease, sick sinus syndrome, celiac disease, prostate cancer, Alzheimer's, and overall cognition in the elderly [25, 27, 28, 29, 30, 31, 32, 33, 34].

Evolution of low frequency variants

The use of indirect association tests popular in GWAS for rare variants is unlikely to be powerful. In order to better understand why, it is important to consider the evolution of rare variants. There is no defined allele frequency threshold to distinguish which variants are considered rare and which variants are considered common. Dickson et al. identify rare

variants using a minor allele threshold between 0.005 and 0.02 [35]. Gibson labels variants as rare if the minor allele frequency is less than 1% [36]. Alternatively, others consider variants rare that have minor allele frequencies less than 0.01-0.05 [7, 37]. In this text, rare variants refer to variants with MAF $\leq$ 1%, low frequency variants refer to variants with MAF $\leq$ 5%, and common variants refer to variants with MAF $>$ 5%.

Rare alleles are observed for virtually every gene; Gorlov et al. estimate at least 2-3 rare variants per gene on average [20]. The expected number and distribution of disease alleles in the population depend on mutation rate, selection and population ancestry. Rare variants represent a considerable proportion of genome variation; Gorlov estimates up to 60% of SNPs in the genome are SNPs $<$ 5% [20]. Many rare alleles are deleterious and presumably persist in the population by recurrent mutation [1].

Population expansion

Demographic scenarios such as population subdivision with a change in migration rates over time and admixture with archaic humans might have affected patterns of sequence variation and linkage disequilbrium (LD). African populations fit a model of continuous population growth, but other populations show a clear signature of a population bottleneck at about the time of emergence from Africa [38]. A severe bottleneck has multiple genetic effects on the population genetic structure. It changes the allelic frequency (genetic drift), increases the average level of homozygosity (inbreeding), and causes correlations of allele frequencies between multiple alleles at the same locus (Hardy-Weinberg Equilibrium (HWE)) and among variants at different positions (LD). Individuals of African descent have patterns of genetic variation consistent with a larger long-term effective population size than populations of non-African ancestry. The large effective population size is reflected in elevated levels of diversity, elevated haplotype diversity, and reduced levels of linkage disequilibrium [1]. Rapid population growth and weak purifying selection have allowed ancestral populations to accumulate an excess of low frequency variants across the genome. This affects genomic analyses in two ways: it alters proportion of deleterious versus neutral variation expected

in low frequency variants and population stratification.

Low frequency variants exhibit extreme population stratification [39]. Demonstrating the magnitude of low frequency population stratification between two populations, Tennessen et al. identified more than 500,000 single nucleotide polymorphisms (SNPs) using 15,585 protein-coding genes from 2,440 individuals. Of these SNPs, 86% had a MAF $< 0.5\%$ and 82% were population specific between European Americans and African Americans [39]. Low frequency allele sharing between populations on the same continent were between 70% and 80%. In contrast, low frequency allele sharing between populations on different continents were lower than 30% and variants were often unique to a single population. In genomic analyses, this extreme population stratification can lead to higher false positives and difficulty in replicating associations across genetic studies when not considered as part of the experimental design for low frequency SNP analyses [40].

Common variants are often identified in more than one continental group, while rare variants are often specific to one population. Common variants shared between African and non-African populations are older and likely existed before the migration out of Africa. Non-African populations tend to have less rare variants (more positive Tajima's D test statistic) than African populations. This can possibly be explained by a population size reduction (bottleneck) during which the rare variants were lost more quickly than the common variants [38].

Effects of selection

The calculated intronic ratio suggested by Gorlov is calculated as the number of SNPs in specific categories (nonsynonymous, possibly damaging, probably damaging, etc.) divided by the absolute number of intronic SNPs and can be used as an approximate measure of selection. Purifying selection drives variants to lower frequency, and positive selection promotes high-frequency derived alleles. For variants with MAF <10% and particularly <5%, the intronic ratio increases sharply suggesting a strong effect of purifying selection. Comparing the different categories, the intronic ratio for probably and possibly damaging

variants is even more increased, suggesting a stronger purifying selection against these categories [20]. Therefore, the excess of rare variants in the human genome leads one to conclude that many low frequency variants are functional and under the effect of purifying selection. For example, cancer suppressors and oncogenes are under the pressure of purifying selection. As a result, protein-damaging mutations in these genes have a lower frequency in the population.

The actual distribution of allele frequencies in populations suggests that many segregating amino acid polymorphisms present at low frequency are mildly deleterious, less likely to be eliminated by weak purifying selective pressure, and likely major players in common disease susceptibility [1, 20]. For example, Nelson et al. found that in 202 drug target genes, 2/3 of the low frequency variants were nonsynonymous mutations, a much higher ratio than found for common variants. This ratio reflects the expected proportion given random mutation and degenerate coding and also supports the theory that low frequency variants are only weakly filtered by selection [4, 41]. Due to weak selection, low frequency variants appear to be enriched for functional variation, including protein coding changes and altered function [40]. In addition, low frequency variants represent a considerable proportion of the variation in the genome due to recent explosive population growth [39]. Since the allele frequency distribution is skewed towards more low frequency variants and many of these are functional, a higher number of low frequency deleterious variants are expected. These evolutionary conditions explain the prominence of disease-promoting variants at low frequencies and reflect the balance between mutation and selection.

Evolutionary models

Human variation tends to fit the expectations of neutrality reasonably well, except that human genes generally show an excess of rare alleles [1]. A common approach to testing the standard neutral model is based on the Tajima's D test. Under the standard neutral model, the expectations of θW and of nucleotide diversity, π, are equal [38]. Certain types of selection (selective sweep, where a rare variant was quickly favored and fixed in the population)

or recent exponential population growth result in an excess of rare alleles, and Tajima's D statistic is negative [1, 38]. Alternatively, a positive value of D reveals a relative excess of intermediate frequency alleles. This is expected under a model of population subdivision or balanced polymorphism. This pattern suggests either some type of balancing selection, in which heterozygous genotypes are favored, or some type of diversifying selection, in which genotypes carrying the less common alleles are favored. This situation may also happen if the sample population was formed from a recent admixture of two different populations [1].

Characteristics of low frequency variants

Calculating age of variants

Beyond theory, there are a few ways to estimate the age of a variant. As mentioned previously, low r^2 values refer to more recent mutations. More precisely, allelic age can be estimated from genetic variation among different copies (intra-allelic variation) and from its frequency. Intra-allelic variation estimates follow the decay of LD. One must know the recombination rate and expected frequency of the mutation at similar loci to calculate the suspected generation time. Kimura and Ohta were the first to consider the relationship between age and frequency (see Equation 1). Time is measured in $2N$ generations and p is the observed allele frequency. For example, if the MAF is 2%, the estimated age is 32,000 years. One can use these methods together or contrast them to show evidence of natural selection [42].

$$E(t_1) = \frac{-2p}{1-p} \ln p \tag{1}$$

Linkage disequilibrium in low frequency variants

Linkage disequilibrium (LD) between single nucleotide polymorphisms (SNPs) allows for SNP tagging and indirect association testing. Linkage disequilibrium is defined as particular combinations of alleles at closely linked loci which occur more or less often than the individual allele frequencies would predict. For example: imagine a locus with two alleles,

Table 1. Example of linkage disequilibrium. Adapted from Hartl et al. [1]

STEPS	**ALLELE (SITE 1)**	**ALLELE (SITE 2)**	**POSSIBLE GENOTYPES**
1. Ancient monomorphic alleles	A	B	AB
2. A mutates to a	Aa	B	AB, aB
3. B mutates to b in chr carrying aB	Aa	Bb	AB, aB, ab

A and a. The Hardy-Weinberg Equilibrium (HWE) principle states (assuming all assumptions are met) that the genotypes AA, Aa, and aa are expected to be p_A^2, $2p_Aq_a$, and q_a^2, respectively (where p and q represent the major and minor allele frequencies). Thus, the A allele is in random association with the a allele. The same could be said for another locus containing B and b alleles, where the B allele is in random association with the b allele (see Table 1). When the alleles of these two loci are not linked, the frequency of a particular combination of alleles equals the product of their respective allele frequencies. This is called linkage equilibrium. When the alleles of the A locus are linked with alleles of the B locus, the loci are in linkage disequilibrium.

The frequency of recombination (r) between loci is important because it determines the rate towards linkage equilibrium. The frequency of recombination is necessarily $r = 0.5$ when two loci are on different chromosomes and $r = 0$ when the two loci are too close together for a break to occur between them. The farther apart two loci are, the more likely recombination between the loci becomes. The genotypic frequencies are related to the allelic frequencies in the previous generation and D is the difference between the frequency of given haplotype (observed) in the previous generation minus the frequencies of the A and B alleles in the previous generation (expected). See Equation 2 for one example haplotype.

$$D = P_{AB} - p_A p_B \tag{2}$$

$$D_n = (1 - r)^n D_0 \tag{3}$$

D_n is the value of D in the n^{th} generation, thus it shows the decay of LD over generations (time) due to recombination. The term $(1 - r)^n$ goes to zero as n becomes large (see

Equation 3); the smaller the value of r the slower the rate towards equilibrium. D_n will go to zero unless there are other factors to offset the decrease to linkage equilibrium (e.g. nonrandom mating or other violations of HWE). The decay of LD, which can be estimated by exponential decay, is shown in Figure 1 for different recombination frequencies [1].

Three things can affect linkage disequilibrium:

1. Recombination. Recombination occurs at hotspots across the genome and breaks up LD. The rate of approach to linkage equilibrium depends on the rate of recombination in genotypes heterozygous for both loci (see Equation 3). Inbreeding reduces the frequency of heterozygous genotypes so that LD is maintained and recombination is minimal.

2. Gene conversion may replace a small integral part of a conserved segment, producing localized breakdown of LD, whereas markers on each side continue to show LD.

3. Population history. The older the population, the shorter the conserved segments. LD is more extensive and of longer range in populations derived from recent founders. LD can result from mixing subpopulations with different allele frequencies. If subpopulations permanently mix and undergo random mating, LD is expected to decrease according to the recombination rate, r, per generation. Similarly, inbreeding reduces recombination because it reduces the frequency of the double heterozygotes, which are essential for recombination to take place.

D depends on the allele frequencies, so it is often normalized by dividing D by the theoretical maximum for observed allele frequencies. To calculate D' and r^2, see Equation 4, Equation 5, and Equation 6 [1].

$$D' = \frac{D}{D_{max}} \tag{4}$$

where

$$D_{max} = \begin{cases} min\left(p_A q_b, q_a p_B\right) & when\ D > 0 \\ min\left(p_A p_B, q_a q_b\right) & when\ D < 0 \end{cases} \tag{5}$$

Figure 1. Linkage disequilibrium between genes gradually disappears when mating is random, providing no other processes are present. The rate of approach to linkage equilibrium depends on the recombination frequency between the genes. Adapted from Hartl and Clark 2006 [1].

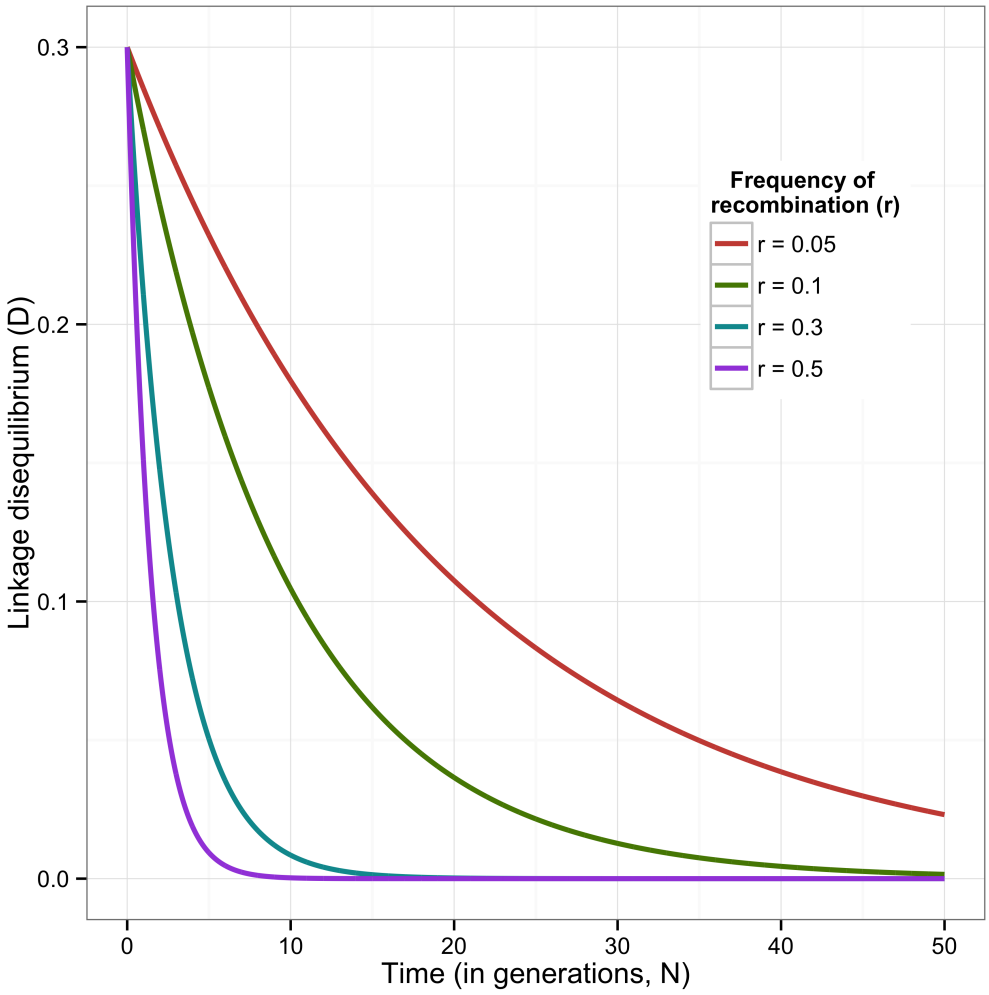


Table 2. Examples of r^2 and age of mutation adapted from Hartl and Clark [1].

AGE OF MUTATION	**GENOTYPE FREQUENCIES**		**CALCULATED** r^2
b arose early in a lineage	$P_{AB} = 0.5$	$P_{Ab} = 0$	$r^2 = 0.96$
	$P_{aB} = 0.01$	$P_{ab} = 0.49$	
b arose much later in a lineage (more recent)	$P_{AB} = 0.5$	$P_{Ab} = 0$	$r^2 = 0.01$
	$P_{aB} = 0.49$	$P_{ab} = 0.01$	

$$r^2 = \frac{D^2}{p_A q_a p_B q_b} \tag{6}$$

D' is a normalized measure of LD and it is mostly influenced by recombination. For example, if $D' = 0.50$, the amount of disequilibrium between the SNPs in the two loci is about 50% of its theoretical maximum. For any given D', r^2 can take any value between 0 and D'^2. The range of r^2 is due to the fact that it also depends on the allele frequencies. r^2 is a measure of linkage disequilibrium, but it captures when/where in the genealogy of the haplotypes the mutation occurred. For example, there are two ancient monomorphic alleles A and B at two sites. Over time, the A allele mutates to a and B mutates to b. The possible genotypes following each of these steps is shown in Table 1 [1]. After the third step (B mutation), notice there is no Ab genotype. It will remain at 0% frequency in the absence of recombination or recurrent mutation. Because of the missing haplotype, $D' = 1$ and the value of r^2 depends on the timing of the $B \rightarrow b$ mutation (see Table 2). The r^2 value relates to the age of the mutation, lower values refer to more recent mutations.

According to simulations performed by Kruglyak, the rapid decay of LD with distance is a consequence of the relatively ancient origin of most common variants. Variants observed at 10% frequency tend to be of a more recent origin, but also date almost exclusively to the time of expansion or earlier (assuming this is a neutral variant). Rare variants often make only three haplotypes with common SNPs, in this case, r^2 can be close to zero (depends on age of variant) while D' is 1 [1, 43]. Therefore, r^2 is the more reliable measure of LD when considering rare variants. The value of r^2 depends on the allele frequency difference between the two loci. The genotyped SNP that tags the most variants from the causal SNP (has the highest r^2) is the SNP with the lowest MAF in which the minor allele is coupled

Table 3. Maximum r^2 and expected odds ratios (OR) between a rare causal variant and common genotyped SNP with an odds ratio of 1.1. Adapted from Wray et al. [43]

Freq. of causal variant	**Estimate**	**Freq. of genotyped SNP**					
		0.05	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**
0.005	r^2	0.10	0.05	0.02	0.01	0.01	0.01
	OR	2.0	3.0	5.0	7.0	9.0	11.0
0.01	r^2	0.19	0.09	0.04	0.02	0.02	0.01
	OR	1.5	2.0	3.0	4.0	5.0	6.0
0.02	r^2	0.39	0.18	0.08	0.05	0.03	0.02
	OR	1.3	1.5	2.0	2.5	3.0	3.5

with the rare causal variant. The possible LD structure between rare and common variants is detailed in Table 3 [43].

Linkage disequilibrium is a result of history; it reflects shared ancestry of haplotypes present in any population [1]. The presence of LD can be explained by LD in a founding population that has not had time to dissipate due to low frequency of recombination or because of natural selection [1, 38]. Random mating reduces haplotype blocks (LD) [1].The International HapMap Consortium defined the ancestral chromosome segments in four human populations and catalogued markers that could be used in GWAS as tag SNPs [20, 44]. The commonly employed strategy of indirect association testing relies on the association between disease and SNPs near a true causal variant, where the associated SNP and causal variant are in LD. This allows for a dense map of tag SNPs to scan the genome for regions associated with a trait of interest [44]. As mentioned before, this has led to the discovery of many common SNPs associated with disease.

Current methods to analyze low frequency variation

There are three categories of analysis possible for rare variants in a whole-genome study: direct association testing, indirect association testing (utilizing LD), and collapsing methods. Direct association testing is plausible, but unlikely to be effective in current common study sizes where the total sample size is < 1000 individuals because rare variants will be scarce and contribute small numbers to the analysis which necessitates cautious interpre-

tation [45]. It is difficult and extraordinarily expensive to ascertain large enough data sets to acquire sufficient numbers of cases that carry the same causal rare variant and to be able to detect a difference in allele frequency when the MAF is so low [20, 21, 22, 46, 47]. As an example, Nejentsev et al. was able to report a rare variant association with a MAF 0.46% in cases and 0.67% in controls using 17,730 individuals [48]. This is much larger than the current size of most sequencing studies due to cost. A single exome with 50x coverage can cost between $850-$1000 US dollars for direct to consumer pricing (https://www.23andme.com/exome/, http://www.axeq.com/axeq.html; accessed July 17, 2013) and approximately $500 US dollars for in-house rates within an institution (http://vantage.vanderbilt.edu/pubutils/ngscalc.html; accessed July 17, 2013). For a single in-house whole-genome sequence (50x coverage) without analysis, the cost is over $8000 US dollars. The prices are steadily falling for next-generation sequence data, but the costs of a study of any reasonable size can quickly exceed $100,000. The prohibitive costs make single variant association testing unfeasible for variants with low minor allele frequencies. Ignoring this limitation with small sample sizes could lead to unstable estimates of rare variant effects on disease and be uninformative [49].

Performing indirect tests of association with rare variants will lead to dubious interpretation of results. Rare variants can be in LD with other variants; rare haplotypes exist and can be associated with disease [50]. However, indirect association testing assumes low-level allelic heterogeneity and assumes that the variants are common [21, 44]. Rare variants have low MAF and low r^2 values and thus exhibit poor tagging properties with common variants (see Table 3) [21, 51]. Inappropriate indirect SNP association testing runs a high risk of false-negative results because rare functional variants can be inadequately tagged. For variants of lower frequency, the decay of LD is similar to common variants, but the maximal level of LD at zero recombination is lower due to the difference in frequency between the variant and the associated SNP allele. To successfully use tag SNPs for indirect association testing, it is best to match the allele frequency of the variant and associated SNP allele. Common variants can be detected with single markers using a tag SNP approach, whereas lower-frequency variants require haplotype analyses or binning for association testing [44]. Haplotype analyses are very sensitive to population stratification, haplotype structure, and

matching allele frequencies and should be utilized with caution [46].

Recently there has been some interest in synthetic associations. Synthetic associations are a particular type of indirect association, specifically, the association of a genotyped common marker resulting from multiple unobserved low-frequency causal variants [35, 43]. Some authors believe rare variants can cause synthetic associations with real risk effects several-fold stronger than what is credited to a tagged common variant [3]. Therefore, variance explained by the causal variants is much higher than what is seen in the associated SNPs because the genotyped SNPs have not tagged the causal variants with great precision.

Synthetic associations are unlikely responsible for signals found in GWAS. Given the required effect sizes (due to poor tagging), it is unlikely that synthetic associations explain common variant associations [43]. These hypothetical rare variant effect sizes are so large they would have been almost certainly picked up in linkage studies (see Table 3). Although very common associated SNPs are unlikely to be causal, they most likely tag causal SNPs with similar allele frequency and are unlikely to represent synthetic associations [43].

Searching for genes with an enrichment of rare variants even in a low number of sequenced genomes is more productive [3]. Given the restrictions of available sample sizes, binning methods are likely to be the most powerful and effective methods to identify causal rare variants. Most often a single variant is likely too rare to completely explain the observed prevalence of a trait of interest, particularly common, complex traits. However, the high proportion of rare variants across the genome, presence of allelic heterogeneity, and presence of locus heterogeneity can explain additional prevalence of a trait [20].

To date, most sequence analysis tools use standard analytical methods to reduce the search space. One standard method is to use family data which allows the analyst to exploit transmission patterns to filter the data [52]. This strategy is effective but not applicable to data sets without family information. Another technique is to perform a candidate gene study and collapse rare variants into bins in order to combine association signals. Collapsing methods are favorable for the following reasons:

1. Applies to case-control studies

2. Applies to whole-genome sequence data

3. Potentially enriches association signals by combining otherwise underpowered rare variants

4. Reduces the degrees of freedom in the statistical test

Collapsing methods, which test cumulative effects of rare variants in genetic regions, can be classified based on the type of statistical test used, either burden or nonburden tests [7].

Burden tests

Instead of testing each variant independently, variants that fall below a specified MAF threshold can be collapsed into a single comprehensive variable for analysis. The burden tests described below are unique approaches that combine variants' weights or manage bins using different MAF thresholds or genomic boundaries. The first researchers to describe a burden test collapsing approach were Morgenthaler and Thilly in 2007 [53]. Their cohort allelic sums test (CAST) calculates the sums of allelic mutation frequencies in cases versus controls and applies a statistical test to determine if the difference is statistically significant. The CAST method assumes that rare variants have the same magnitude and direction of effect. Because the method uses a chi-square statistic, it is less than ideal because it does not easily incorporate covariates, cannot be used in quantitative phenotypes, and does not measure the direction of association [22]. One year later, Li and Leal developed a similar method, the combined multivariate and collapsing method (CMC). The CMC method uses a multivariate statistical test and permits combined analysis of rare and common variants [21]. The CMC method has improved power over CAST, presumably because functional information (direction of effect) was incorporated and because the method can be implemented in a regression framework [54].

The next group of published collapsing tests introduced the idea of individual variant weighting. Witte describes two approaches to weighting: *a priori* weighting or empirical weighting [5]. *A priori* aggregation methods can be used in many ways; the CAST method applies an a priori weight because it requires "all or nothing" bins. If an individual has

one or more variants, the comprehensive genetic variable is the same. Another sensible way to weight variants includes using properties such as minor allele frequency cutoffs, or nonsynonymous versus synonymous changes. Madsen and Browning proposed a collapsing method using *a priori* weights, each variant is weighted using its allele frequency, a comprehensive genetic score is calculated and then a rank sum test between cases and controls is performed [55].

Other burden tests use empirical weights to aggregate variants, essentially utilizing external information about the potential functionality of variants, such as, variable minor allele frequency cutoffs and directionality of effect [5]. Price et al. propose a method to optimize the grouping of rare variants using a variable-threshold approach based on allele frequency [56]. Similarly, Fang et al. propose a pooling method using a threshold of risk measure instead of allele frequencies to build bins with the most powerful association signal [57]. Hoffman et al. utilize a step-up approach to iteratively add variants to a bin only if it improves the association signal [49]. Several other methods cleverly incorporate functional data to guide collapsing and use a regression framework for statistical association [58].

Burden tests are notoriously less powerful when variants binned together have opposing directions of effect [7, 37]. It is important to employ filtering strategies and attempt to create bins with functional variants with the same direction of effect.

Nonburden tests

Instead of assessing the cumulative effects of variants in a bin by summarizing the genetic score into a single value, nonburden tests investigate the variance distribution of allele frequencies. The first published nonburden test was the C-alpha test. In case-control data, it compares the expected variance to the actual variance of the allele frequency distribution [59]. Nonburden tests are often more powerful in bins where variants have different directions of effect. Kernel based tests, such as SKAT and SKAT-O, improve upon C-alpha because they can be implemented in a regression framework (rather than requiring permutation), allow for easy covariate adjustment (including controlling for population stratifica-

tion), and can be applied to dichotomous and continuous phenotypes [37, 7]. The kernel association test aggregates individuals' variant-score tests statistics with weights when SNP effects are modeled linearly. Then they aggregate associations between variants and phenotypes using kernel matrix. SKAT tests can also incorporate local correlation substructure, weights, and can allow epistatic effects [37].

Nonburden tests are often overly conservative, particularly in small study sizes and when the large majority of variants are truly causal. SKAT-O improves upon SKAT because it allows correlation between variant regression coefficients, which improves power when binned variants are in the same direction of effect [7].

Conclusions

As with linkage and GWAS, the number and penetrance of alleles affecting disease risk, i.e., the genetic architecture of a disease, directly affect the strategy for identifying polymorphisms that modulate disease susceptibility [20]. One must be careful to match cases and controls since overrepresentation of rare variants in a specific ethnic group may complicate the interpretation of association analyses of such variants. Even though there are many available testing strategies, statistically significant mutations, multiple mutations that are functional and co-segregate with disease, de novo mutations, and/or model organisms are required to prove a link between variation in these genes and disease [47].

CHAPTER III

BIOBIN SOFTWARE

Many recent publications detail collapsing approaches for low frequency or rare variant association tests. These methods build bins of multiple rare or low frequency variants across pre-defined regions and use a statistical test to detect an association between the presence/absence, number, or the distribution of low frequency variants and case/control status. BioBin contributes novelty to the field of low frequency variant association testing by focusing on defining regions rather than demonstrating the use of a particular statistical test. Most available software packages include a default statistical test and do not provide any guidance or assistance in defining regions for binning. For a given dataset and hypothesis, there are different and sometimes multiple statistical tests that are appropriate. In addition, novel statistical tests for binning methods are published frequently in the literature, and the freedom to choose a specific test for an analysis is often preferable. There are explicit situations that require the use of regression analysis (logistic, linear, or polytomous), Fisher's exact test, or permutation of unique statistical test, etc. A simple analytical pipeline for low frequency variant analyses using BioBin is shown in Figure 2.

Overall, the most powerful collapsing analysis to detect associations with low frequency variants will use a method to define bin boundaries in a way that will combine low frequency variants with similar functional properties and apply the most appropriate statistical test.

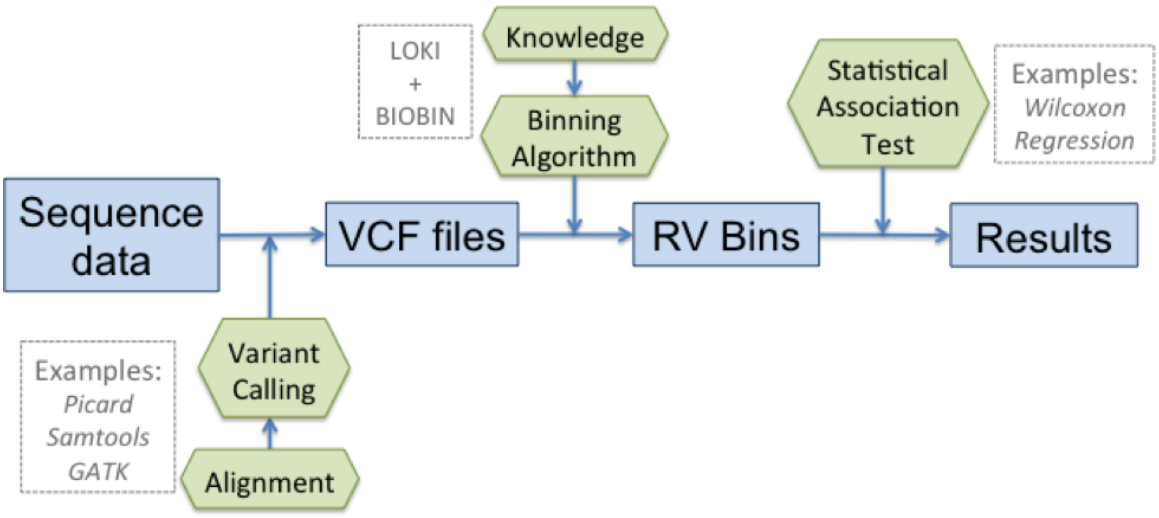


Figure 2. Pipeline for BioBin analysis. Blue squares correspond to data, green hexagons correspond to bioinformatic or statistical method applications.

BioBin resource requirements

BioBin is a standalone command line application written in C++ that relies on a locally built Library of Knowledge Integration (LOKI) database (see later section describing LOKI) to create knowledge based bins. Source distributions are available for Mac and Linux operating systems and require minimal prerequisites to compile. In the BioBin distribution download, included tools allow the user to create and update the LOKI database by downloading information directly from source websites.

BioBin computational requirements scale primarily according to the number of loci in the study. To demonstrate this, the population size and number of loci has been varied in the input variant calling format (VCF) file of the 1000 Genomes Project Phase I low coverage data to assess the resource requirements of BioBin [60]. Over 10 replicates, Figure 3 shows that bin generation is highly correlated to the number of loci in the study and bin generation drives the memory and time usage. The number of individuals in a study does not have a large impact on resource requirements, but does increase the size of the input VCF file and thus time it takes BioBin to read the input VCF file. Even with large datasets, BioBin can be run without access to specialized computer hardware or a computing cluster; however, the number of binnable low frequency variants is the primary

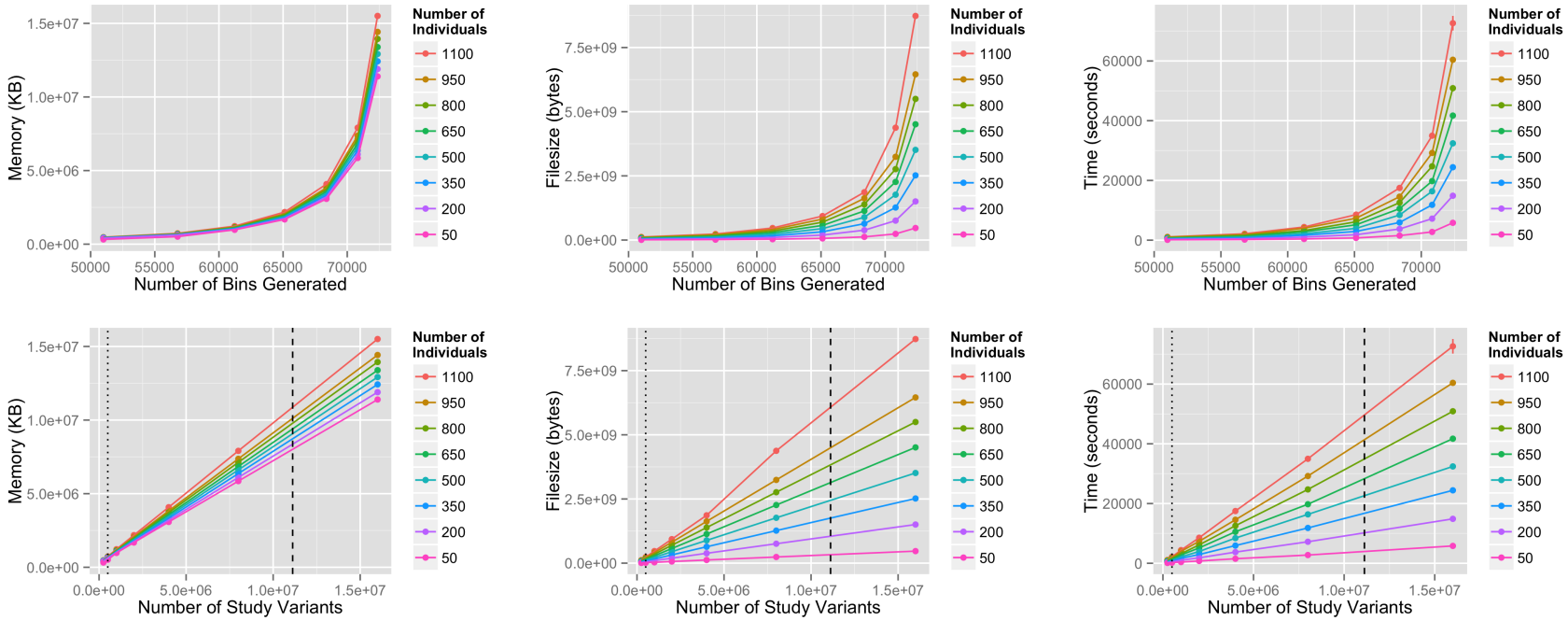


Figure 3. BioBin resource requirements after varying the population size (indicated by legend) and number of study variants (x-axis, bottom row). The top row shows the corresponding number of bins generated by BioBin for each interval of study variants.

driver of memory usage. BioBin is open-source and publicly available on the Ritchie lab website (`http://ritchielab.psu.edu/ritchielab/software/`).

Library of Knowledge Integration database (LOKI)

Harnessing prior biological knowledge is a powerful way to inform collapsing feature boundaries. BioBin relies on the Library of Knowledge Integration (LOKI) for database integration and boundary definitions. LOKI contains resources such as: the National Center for Biotechnology (NCBI) dbSNP and gene Entrez database information [61], Kyoto Encyclopedia of Genes and Genomes (KEGG) [62], Reactome [63], Gene Ontology (GO) [64], Protein families database (Pfam) [65], NetPath - signal transduction pathways [66], Molecular INTeraction database (MINT) [67], Biological General Repository for Interaction Datasets (BioGrid) [68], Pharmacogenomics Knowledge Base (PharmGKB) [69], Open Regulatory Annotation Database (ORegAnno) [70], and evolutionary conserved regions from UCSC Genome Browser [71].

LOKI provides standardized interface and terminology to disparate sources, each containing individual means of representing data. The four main concepts used in LOKI are *positions*, *regions*, *groups*, and *sources*. The term *position* refers to single nucleotide polymorphisms (SNPs), single nucleotide variants (SNVs) or low frequency variants. The definition of *region* has a broader scope, any genomic segment with a start and stop position can be defined as a *region*, including genes, copy number variants (CNVs), insertions and deletions, and evolutionary conserved regions (ECRs). *Sources* are databases (such as those listed above) that contain *groups* of interconnected information, thus organizing the data in a standardized manner. For example, BioGrid ID:468346 defines a *group* from the BioGrid data *source*. This *group* contains the following *regions*: *HMGB1P1*, *CTCFL*, and *PRMT7*.

LOKI is implemented in SQLite, a relational database management system, which does not require a dedicated database server. The user must download and run installer scripts (python) and allow for 10-12 GB of data to be downloaded directly from the various sources. The updater script will automatically process and combine this information into a single

database file (∼6.7 GB range). A system running LOKI should have at least 50 GB of disk storage available. The script to build LOKI is open source and publicly available on the Ritchie lab website (`http://ritchielab.psu.edu/ritchielab/software/`). Users with knowledge of relational databases can customize their LOKI database by including or excluding sources, including additional sources, and updating source information as frequently as they like [Pendergrass et al., in preparation].

BioBin software overview

BioBin options can be configured via configuration file or command line input, which is helpful when developing low frequency variant analysis pipelines. Even in the same data, one might consider testing multiple hypotheses. For example, one could run BioBin with binning boundaries based on genes and then make a few small changes to the configuration file to run pathway binning analyses. BioBin also includes several novel features and options for evaluating low frequency variants, those are described in the Software Features section.

Input files

To run BioBin, the user **must** have a locally built LOKI database and two study files: 1) variant calling format file (VCF) and 2) phenotype file. The LOKI database is described in the previous section and detailed instructions can be found in the BioBin manual, which is available on the Ritchie lab website (`http://ritchielab.psu.edu/ritchielab/software/`). At this time, BioBin only accepts zipped and unzipped VCF files as study input. A single input VCF file should include all of the relevant individuals and study variants. The most recent genome build is preferred for genome coordinates to match LOKI information, but BioBin contains an internal algorithm derived from LiftOver to transform variants from older builds to the newest build if necessary [71]. Lastly, the user must include a phenotype file, a simple file with two columns indicating the sample identifiers

Table 4. Example phenotype input file

ID	PHE
ID1	0
ID2	0
ID3	1
ID4	1
ID5	0

(string value) found in the VCF file and corresponding phenotype (floating point value). Most of the current tests have focused on binary outcomes, but categorical or continuous outcomes are acceptable as phenotypes. The user needs to specifically indicate the desired "control" group to determine allele frequencies and to determine which variants in the data are binnable. For binary traits, one group should be designated as the control group. For quantitative traits, a single group can be designated as controls or all individuals can be considered together to determine binnable variants. An example phenotype file with binary outcomes is shown in Table 4. Other input files are optional and described in further detail in the Software Features section.

Output files

There are two main output files produced by BioBin: bins report and locus report. The bins report provides information on bins generated by BioBin. An example bins report output file is shown in Table 5. Lines 1-6 include the file header and summary information for each bin. Each line after 6th row corresponds to an individual in the study. After ID and Status columns, columns $i = 3..N$ represent all of the bins generated by BioBin. The values in a cell correspond to the contribution of variants of each individual (row) to the bin (column).

The summary rows summarize the variants and loci in each bin. Row 1 contains the total number of variants found within a bin. Row 2 represents the total number of loci binned together. With regard to the values in rows 1 and 2, a locus corresponds to the physical location of the variant. A single locus can represent multiple variants because there can be

Table 5. Example bins report output file

ID	Status	TTLL10	WRAP73
Total Variants	-1	32	63
Total Loci	-1	5	5
Control Loci Totals	-1	5	5
Case Loci Totals	-1	5	5
Control Bin Capacity	-1	134	172
Case Bin Capacity	-1	130	130
NA06984	0	0	1
NA06985	0	0	0
NA20504	1	0	1
NA20506	1	0	0

multiple alleles at a particular location in the population. Rows 3 and 4 exclude loci for which data are entirely missing from either the case or control populations. Rows 5 and 6 show the total bin capacity for either the cases or controls. The capacity is defined as the absolute maximum number of variants that could be contributed to a given bin.

The locus report contains information about the variants and bin statistics, but does not contain any information about individuals in the study. Each line corresponds to a locus in the study, similar to the VCF file. A sample of the locus report output file is shown in Table 6. Columns 1, 2, and 3 identify each locus. Column 4 represents the alleles and their frequencies, as calculated from the designated control population. A pipe (|) character separates individual alleles, and the allele and frequency are separated by a colon (:). The alleles are ordered from most frequent to least frequent, and the minor allele frequency (MAF) is defined to be the frequency of the second most common allele. Column 5 refers to the non-major allele frequency in the case population. The non-major allele frequency is defined to be the frequency of all alleles other than the most common allele in the control population. Column 6 represents the status of the locus. If the minor allele frequency is below the threshold for binning, this column will be 1, if the minor allele frequency exceeds the threshold, it will be 0. Column 7 lists all genes that contain the locus separated by a pipe (|). Column 8 lists the bin names that contain the locus, again separated by a pipe (|).

Generation of any of these reports or a few additional reports that include allelic or bin statistics can be turned on or off at the user's request using options available on the command line or in the configuration file. For information on additional reports, the BioBin

Table 6. Example locus report output file, including the chromosome (Chr), base pair location (BP), variant ID (ID), alleles, allele frequency in case group (AF_{Case}), whether or not the variant meets user designated minor allele frequency threshold (Rare), gene name, bin name.

Chr	**BP**	**ID**	**Alleles**	**AF_{Case}**	**Rare**	**Gene(s)**	**Bin(s)**
1	1115503	rs111751804	T:0.97|C:0.03	0.02	1	TTLL10	TTLL10
1	1115548	rs114390380	G:0.99|A:0.01	0.02	1	TTLL10	TTLL10
1	1118275	rs61733845	C:0.96|T:0.04	0.05	1	TTLL10	TTLL10
1	1120377	rs116321663	T:0.99|A:0.01	0.01	1	TTLL10	TTLL10
1	1120431	rs1320571	G:0.96|A:0.04	0.04	1	TTLL10	TTLL10
1	3548136	rs2760321	C:0.85|A:0.15	0.19	0	WRAP73	WRAP73
1	3548832	rs2760320	G:0.94|A:0.06	0.04	1	WRAP73	WRAP73
1	3548855	rs114376964	T:0.99|C:0.01	0.01	1	WRAP73	WRAP73
1	3551737	rs116230480	C:0.99|T:0.01	0	1	WRAP73	WRAP73

manual is available on the Ritchie lab website (`http://ritchielab.psu.edu/ritchielab/software/`).

BioBin software features

In addition to flexible and biologically informed binning strategies, several important features have been implemented to improve upon existing collapsing approaches. These parameters include: user-defined or customized knowledge, adjustable multi-level feature types, various filtering strategies, flexible loci selection, and individual variant weighting.

Customized knowledge

The LOKI database contains diverse and comprehensive knowledge from many databases, which together provide variant details, region annotations, and multiple region or group relationships (e.g. pathways or protein interactions). To accommodate a wide-variety of analyses, the user can choose to include or exclude any source available in LOKI from the command-line or configuration file. If provided by the user, BioBin accessible knowledge can also be expanded to include sources of knowledge outside of LOKI. For example, if a user wishes to bin specific regions based on his/her research knowledge that is not described in any public database loaded into LOKI, there are several options to include this novel information for a BioBin analysis. The first option is to add this knowledge to the LOKI database. This requires a relatively advanced understanding of the LOKI framework and SQL relational databases, but LOKI is open-source and can be modified on a local machine. A second and likely easier option is to input custom region files to transiently define bin boundaries. Custom feature files can be used in place of or in addition to LOKI knowledge. An example of a custom feature file is shown in Table 7.

Table 7. Custom region feature file.

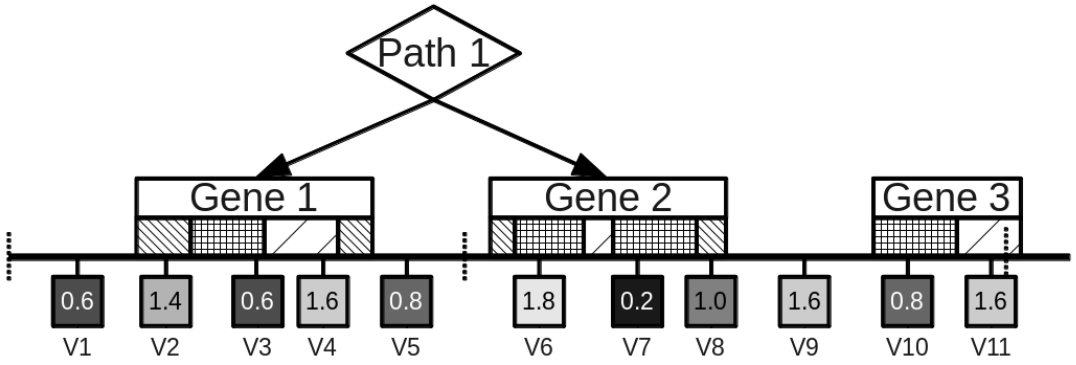
Chrom	ID	Start(bp)	Stop(bp)
9	Region1	1384729	1673929
13	Region2	14940582	16392837
18	Region3	27361833	29877254
22	Region4	188726	1208327

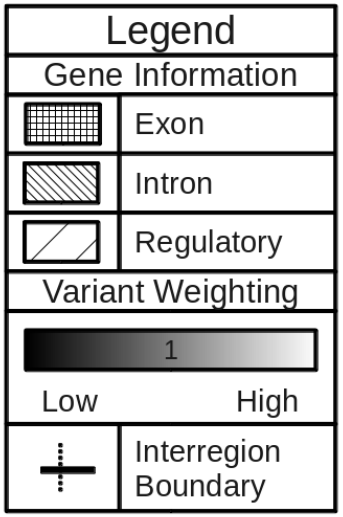
Multi-level feature binning

The most important component of BioBin is the ability to bin at multiple levels of biological knowledge. Example binning strategies can be seen in Figure 4. One can create gene-based bins for an exome study, but very quickly change the configuration file to collapse genes together to investigate evidence of protein-protein interactions. Using hierarchical biological relationships and optional functional or role information, BioBin can create bins based on many unique binning guidelines.

As a standard in the current iteration of LOKI, NCBI dbSNP and NCBI Entrez Gene have been selected as the primary sources of position and regional information due to the data quality, reliability, and clearly defined database schema. These sources also most closely correspond to the region and group IDs provided by other database sources integrated into LOKI.

In addition to binning variants based on knowledge, BioBin also provides an option to bin variants that do not associate with any available knowledge. These are known as interregion bins, or if generated between gene features, intergenic bins. After feature selection using LOKI and/or external custom files, interregion bins can be created using a configurable width parameter (in kb). These bins catch variants that do not fit into the user-defined or biologically defined feature types (see intergenic bin labels on Figure 4). For example, if one were testing low frequency burden differences between two groups across genes, all variants in genes would be collapsed into respective gene bins, and variants outside of gene boundaries would be binned based on genomic location in intergenic regions.





Sample Binning Strategies

Gene Burden Analysis

Gene 1	Gene 2	Gene 3	Intergenic 1	Intergenic 2
V2, V3, V4	V6, V7, V8	V10, V11	V1, V5	V9
3.6 / 3	3.0 / 3	2.4 / 2	1.4 / 2	1.6 / 1

Pathway Burden Analysis

Path 1	Gene 3	Intergenic 1	Intergenic 2
V2, V3, V4, V6, V7, V8	V10, V11	V1, V5	V9
6.6 / 6	2.4 / 2	1.4 / 2	1.6 / 1

Role Pathway Burden Analysis

Path 1 (E)	Path 1 (I)	Path 1 (R)	Gene 3 (E)	Gene 3 (R)	Intergenic 1	Intergenic 2
V3, V6, V7	V2, V8	V4	V10	V11	V1, V5	V9
2.6 / 3	2.4 / 2	1.6 / 1	0.8 / 1	1.6 / 1	1.4 / 2	1.6 / 1

Figure 4. Alternate binning strategies using biological knowledge (Gene Information) and functional or role annotations (Variant Information). Three example binning strategies are shown: gene burden analysis, pathway burden analysis, and functional pathway burden analysis. Note the intergenic bins that collects variants fall outside of the binning strategy.

Table 8. Custom region feature file.

Chrom	ID	Start(bp)	Stop(bp)
10	DEL	100010909	100010909
11	DEL	99715682	99715682
16	DEL	97443	97443
18	DEL	13029986	13029986

Filtering strategies

In published whole-exome studies, a series of filters are often applied to remove neutral or presumably low impact variants. Frequently this is accomplished by excluding variants found in datasets such as 1000 Genomes Project or excluding variants with certain properties, i.e. synonymous or predicted neutral variants. Similar to the custom region knowledge files described previously, BioBin accepts custom role files, which contain single variant or region annotations. These custom role files can be used to *exclude* or specifically *include* variants for a binning analysis. For example, one could use a role file to exclude variants based on the 1000 Genomes Project. Alternatively, with the same role file and slight parameter change, one can study variants exclusively present in the 1000 Genomes Project. This functionality is particularly useful if the user wants to filter based on protein coding variants or predicted damaging variants using an annotation tool such as Polyphen-2 or SIFT [72, 73, 74]. Table 8 shows an excerpt of annotated variants from the 1000 Genomes Project using Variant Effect Predictor Tools (VEP), an annotation tool that provides SIFT and PolyPhen-2 predictions [72, 73, 74, 75]. For example, if using a role file similar to one found in Table 8, BioBin could create gene feature bins containing only variants predicted to be deleterious or damaging (DEL).

Locus selection

In the binning method literature, it is common for studies to calculate allele frequencies in unaffected individuals to determine if a locus is binnable, i.e. less than the MAF binning

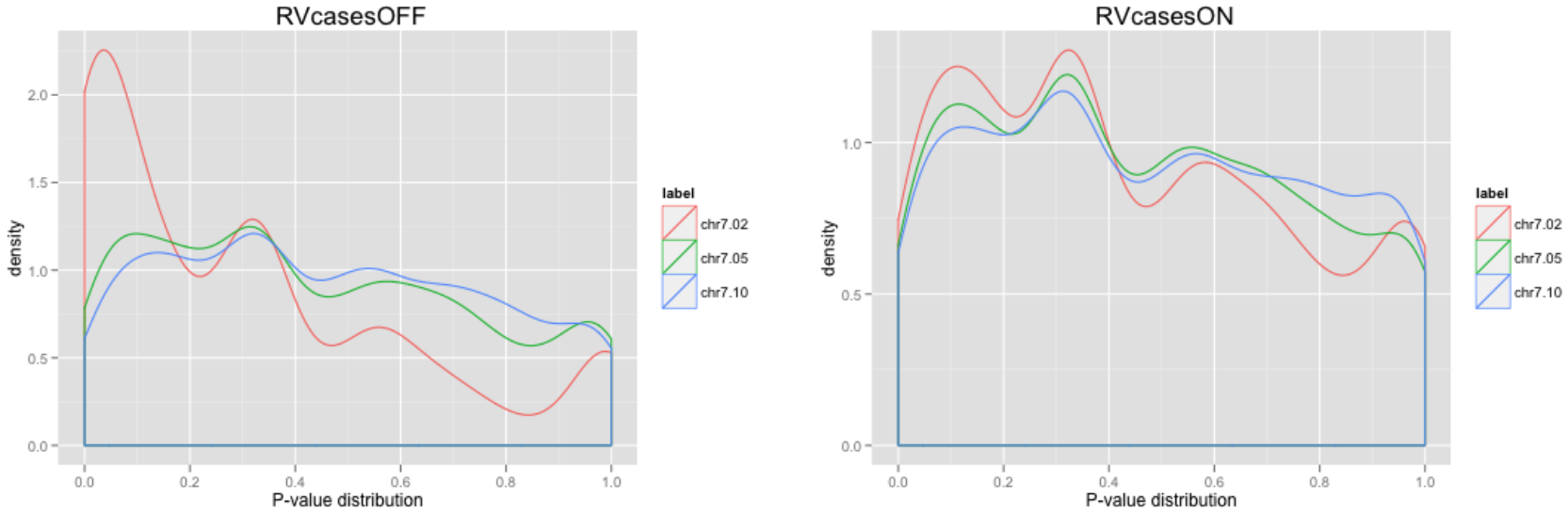
threshold. However, constraining binnable loci based only on controls leads to selection bias and an increase in type I error, which worsens in large bins. In this thesis, the term *locus* refers to a strict chromosome coordinate or position and the term *variant* describes alleles at that locus. The minor allele at a given locus is determined from the second most frequent allele in the control group.

Three parameters have been implemented in BioBin to manage type I error, reduce selection bias, and increase flexibility in selecting binnable loci. The first option is a configurable MAF binning threshold. Binning strategies are applied to low frequency variants, where the user defines "low frequency." Price et. al proposed a variable threshold approach; Price suggested that a single minor allele frequency threshold does not apply to all studies [76]. Although this is not available as an automatic optimization in BioBin, the MAF binning threshold can be tested and optimized by the user for his/her study data.

The second option, labeled Rare-Case-Control or "RCC," addresses how BioBin handles the minor allele frequency-binning threshold in two groups. When RCC is enabled, if the variant is low frequency in either group, the locus is binnable. This does not change which allele is considered the minor allele, the minor allele is annotated using unaffected group allele frequencies, but it does increase the number of binnable loci because it includes specific sites (or loci), which have a low frequency variant in cases and not in controls.

To illustrate the effect of "RCC," the CEU population from 1000 Genomes Project data was randomly divided into cases and controls. In Figure 5, each line corresponds to a MAF binning threshold. At a very low binning threshold (MAF ≤ 0.02), there is a distinct increase in bins with small p-values (red line). The low MAF threshold constricts the contribution of variants by each control to only one variant, whereas, each case can contribute one or more variants. In larger bins (bins containing more loci), this slow accumulation of variants in cases quickly leads to significance. However, when the RCC parameter is turned on, the p-value distribution is more uniform and selection bias is reduced.

A Kruskal-wallis test was run to determine if the mean p-value was the same in groups with varying loci per bin (accomplished by increasing the MAF binning threshold). Without using the RCC option, the Kruskal-wallis p-value was $2.2e^{-16}$ (see Figure 5a). Thus, at least one of the three MAF threshold groups had p-values that differed from the others more than



(a) RCC:OFF

(b) RCC:ON

Figure 5. P-value distribution under three different MAF binning thresholds to test RCC option: 0.02, 0.05, and 0.10. The left figure shows the p-value distribution when RCC is OFF, the right figure shows the p-value distribution with RCC is ON.

chance alone. In Figure 5b, the RCC option was utilized and the p-value was no longer significant (p-value = 0.288). When RCC was on, the difference between the three groups was no longer detectable. Therefore, adding the RCC option decreases error and specifically decreases the correlation between bin size and significance.

Although it is not advisable to perform genomic association analyses in low frequency variants in study data with considerable heterogeneity in ancestral background, there are still loci with considerable allelic heterogeneity within continental groups [4, 5, 6]. The third parameter option concerns choosing the minor allele. In a recent population comparison, switching the group status (case/control) changed the results by 1-3%. The last option for loci selection,"overall-major-allele," addresses this problem (denoted as "OMA" in Table 9). This option was added to allow BioBin to look at affected and unaffected groups before determining which alleles are the major and minor alleles. When the overall-major-allele option is turned on, the major allele is designated by the overall highest frequency allele.

In Table 9, each line corresponds to alternative ways a single variant would be handled under permutations of these options for 100 affected individuals (MAF=0.3) and 100 unaffected individuals (MAF=0.05). Generally, controls are assigned as individuals unaffected by the phenotype; however, to illustrate RCC and OMA options, the control group alternates between affected and unaffected individuals. Examples A-D show the function of the rare-case-control option. In example A, the variant is not rare enough in the designated control group to be binned. Under the same parameters, if the unaffected individuals are designated as controls, the variant is binned because it meets the minor allele frequency threshold (example B). Most often, it is not beneficial for results to change based on control designation. Examples C-D show how the variant is binned regardless of control group designation when the "RCC" option is used. The minor allele frequency within each group does not change. Therefore, the allele chosen as the minor allele is still dependent on the control group, but a variant can be binned if it is "rare" in either group. Since multiple variants can be present at a single locus, these rules are applied at each locus using the most frequent and second most frequent alleles. Once the locus is determined binnable, any variants at that locus will be considered in the binning analysis. When RCC is OFF at low binning thresholds, the number of loci in a bin is highly correlated with significance. Using

the RCC option reduces selection bias and bin size correlation with significance.

Examples E-H in Table 9 show the benefits of using the OMA option. Specifically, when comparing the low frequency variant count between affected individuals and unaffected individuals, examples C-D show more similar counts than A-B, but G-H have the exact variant counts which means that the designation of "control" status is unimportant. Using rare-case-control and overall-major-allele options require BioBin to review allele counts collectively between the two groups to choose the minor allele. The individual group minor allele frequency does not change, but which allele is considered the major (and thus minor) allele can change. The overall minor allele is not necessarily the control group minor allele. As shown in examples G-H, this is the ideal condition for a population comparison where results should be independent of which group is chosen as the control group.

Optional inheritance patterns

BioBin can alter the method of "counting" variants in a bin if the user wishes to employ an alternative inheritance pattern. The default option utilizes additive encoding, where each allelic variant adds to an individual bin score. It is also possible to use dominant or recessive encoding if the user wishes to test a specific hypothesis with those inheritance patterns.

Bin dependency

Bins are often not independent of each other and this should be considered in the statistical analysis. BioBin provides a measure of dependency in the screen output. Figure 6 shows a whole-exome analysis example, the total number of binnable variants was 238,145. Of those, 222,564 variants were binned only once, while 11,255 variants were found in more than one bin (between 2 and 22 bins). In a gene analysis, only a small proportion of variants are included in more than one bin ($< 5\%$). One should consider this information when determining the best method to account for multiple test correction.

Table 9. Iterations of major/minor allele selection and variant binning using parameters rare-case-control (RCC) and overall-major-allele (OMA). Using both options is necessary to make the results independent of control group selection and maintain a reasonable type I error rate.

Ex	**POP**	**Minor Allele**	**Major Allele**	**Group A**	**RCC ON**	**OMA ON**	**Minor Allele**	**Group A MAF**	**Low Freq. MAF≤0.05**	**Bin Count Group A**	**Bin Count Group B**
A	unaffected	T:60	A:140	unaffected	NO	NO	T	0.3	NO		
	affected	A:10	T:190								
B	unaffected	T:60	A:140	affected	NO	NO	A	0.05	YES	10	140
	affected	A:10	T:190								
C	unaffected	T:60	A:140	unaffected	YES	NO	T	0.3	YES	60	190
	affected	A:10	T:190								
D	unaffected	T:60	A:140	affected	YES	NO	A	0.05	YES	10	140
	affected	A:10	T:190								
E	unaffected	T:60	A:140	unaffected	NO	YES	A	0.7	NO		
	affected	A:10	T:190								
F	unaffected	T:60	A:140	affected	NO	YES	A	0.05	YES	10	140
	affected	A:10	T:190								
G	unaffected	T:60	A:140	unaffected	YES	YES	A	0.7	YES	140	10
	affected	A:10	T:190								
H	unaffected	T:60	A:140	affected	YES	YES	A	0.05	YES	10	140
	affected	A:10	T:190								

Figure 6. Screen output indicating bin dependency.

```
[ccb12@hammer4 pa_only_CBM]$ biobin-nightly pa_only_w_weight.cfg

   Total SNPS:       238145
   Variants:         238145
 * Rare Variants:    238145
   Total Bins:        18330

   * Rare variants are those whose minor alleles sum is below: 0.05


Number of Bins per locus
1       222564
2-22    11255

Executing Dataset File Generation
              pa_only_cases_w_weight_0.05-bins.csv : Bin Counts
             pa_only_cases_w_weight_0.05-locus.csv : Locus Data
```

Variant weighting

Madsen and Browning were the first to propose using individual variant weights to influence composite genetic scores. In their original paper, mutations were grouped into a bin and each individual was scored by a weighted sum of mutation counts. According to Madsen and Browning, the mutation frequency, q_i, for each locus is dependent on the number of variant alleles observed in unaffected individuals (m_i^U) and the number of affected and unaffected individuals (n_i, see Equation 7) The calculated weight (w_i) is the estimated standard deviation of the total number of mutations in the sample under the null hypothesis (see Equation 8). Finally, the genetic score for each individual is the sum of all variable loci in the bin divided by their respective weights [55](see Equation 9).

$$q_i = \frac{m_i^U + 1}{2n_i^U + 2} \tag{7}$$

$$w_i = \sqrt{n_i \cdot q_i \cdot (1 - q_i)} \tag{8}$$

$$\gamma_j = \sum_{i=1}^{L} \frac{I_{ij}}{w_i} \tag{9}$$

Although the weight sum test is implemented in a nonparametric framework using permutations on sum rank test statistics to estimate significance, it has been noted by us and others that the calculation of q_i based only on the observed control group (unaffected samples) results in inflated type I error [77]. The original Madsen and Browning implementation allows bias weighting such that weights in binned alleles with higher frequencies in cases are unbounded, while weights of alleles with higher frequencies in controls are bounded. Using this weighting scheme, there is a higher false positive rate even in the presence of no true genetic effect (see Chapter IV) [77]. To provide an unbiased weight, BioBin permits a weighting parameter with four options: control weight, overall weight, maximum weight, and minimum weight. The control weight is reflective of the original Madsen and Browning calculation, which uses only the number of alleles in unaffected individuals to calculate q_i. The overall weight uses the overall allele count in the calculation of q_i and is commonly used in popular software association packages [78, 79]. The maximum and minimum weights are calculated by using the maximum or minimum $1/w_i$ value when q_i is calculated using affected and unaffected individuals. The maximum and minimum weights allow for weights to reflect large differences in allele frequencies in cases and controls and are more powerful than using the overall weight option but do not lead to bias in results. Weights are incorporated similarly to Madsen and Browning to calculate the genetic score for each individual. Equation 10 shows this relationship and an additional custom weight (w_c) that can be implemented using custom weight input files. Custom weights allow the user to manage additional weights; for example, one might increase the weight of nonsynonymous variants to 1.1 to reflect the potential burden of damaging nonsynonymous variation.

$$\gamma_j = \sum_{i=1}^{L} I_{ij} \cdot \frac{1}{w_i} \cdot w_c \tag{10}$$

Statistical tests

The focus of the BioBin software is to build flexible and biologically relevant bins; therefore, BioBin does not include any particular statistical test in the software package. The lack of an implemented statistical test is preferable, since it allows the user the freedom to choose the most appropriate statistical test given their study data and hypothesis. The BioBin analyses presented in this thesis use burden tests, a composite genetic score has been used in multiple statistical frameworks to detect associations between independent genetic variables and a trait of interest. The following statistical tests were used: logistic regression, Wilcoxon two-sample rank sum test, and standard permutations. The most basic genetic score is just the individual's sum of variants within a single bin. Using weights, each variant can be influenced by weights based on allele frequency or custom weights provided by the user. The formula for combining those weights (if present) and calculating a genetic score is shown in Equation 10.

In the presented analyses, when a logistic regression was used, the null hypothesis of no effect was tested as $\beta_1 = 0$. When results were calculated using a Wilcoxon two-sample rank sum test, a nonparametric method to test if the mean ranks differ between two groups, the null hypothesis of no difference between the mean ranks was tested. In some analyses, permutations were used to affirm simulation results. Permutations were performed using either a Wilcoxon two-sample rank sum test statistic or the rank sum test as described by Madsen and Browning [55]. For the permutation test, the phenotype was randomly assigned and the resulting bin was tested 1000 times. The p-value was calculated as the proportion of permutations in the null distribution that were more extreme than the observed value.

Summary

There are challenges for association detection using binning analyses, variants in the same bin can have various functional effects (protective, detrimental or neutral), allele frequencies

at variant positions are often population specific, and there has not been a clear standard for statistical testing. However, collapsing algorithms improve power in low frequency variant analyses when large sample sizes ($>$ 10,000 individuals) are not available. Also, collapsing methods provide an avenue to embrace allelic heterogeneity, locus heterogeneity, and epistasis. Knowledge-based binning increases the likelihood that variants with similar functional properties will be binned together and that an association signal can be detected. Furthermore, collapsing method results, in particular BioBin results, are interpretable biologically.

CHAPTER IV

SIMULATION STUDIES

Since low frequency binning is a relatively new approach, BioBin had to be extensively tested utilizing simulations with multiple statistical approaches and weighting options. Simultations described in this chapter were generated using SimRare, a GUI interface for simuPOP, a forward time simulator [80, 81]. Together, these two software programs simulate introduction and evolution of rare variants and can allow complex fitness and selection modeling (http://simupop.sourceforge.net, https://code.google.com/p/simrare/) with a user-friendly approach. SimRare takes less time and is more computationally efficient because replicates are generated and data are stored as population averages rather than storing individual haplotypes. In this study, the term *replicate* is a realization of the forward-time simulation using the given input parameters. In each of the studies described in this chapter, 250 replicates were simulated. Each replicate uses the same evolutionary parameters but differs because of the random seed variables and random genomic size. After the replicate initializations, populations of any size with any genetic effect can be modeled from this population. In this chapter, the number of times populations are generated from this pool for testing will be referred to as *duplications*.

In all of the simulations described below, an additive multilocus model with a selection coefficient distribution was used, previously described by Kryukov [82]. The mutation rate was set at of $1.8e^{-8}$ per nucleotide per generation. The population sizes were N_e = 8100, 8100, 7900, and 900,000 with 5000 generations, 10 generations, and 370 generations respectively.

Type I error assessment

BioBin is a flexible binning algorithm; the variety of available weighting options and variable size of output bins should be adequately tested in simulated data to provide future users assurance in their study results. In this chapter, type I error results are presented from a gene region simulation study, large region simulation study, and two dependent group simulation studies. The gene region simulation study represents a biological gene region with a single start and stop position. The large region simulation study represents a much larger region with a single start and stop position that could be a very large gene, but is the average size of pathway bins. The first dependent group simulation study represents pathway bins by randomly selecting gene region simulations into a single bin. The second dependent group simulation study is similar to the first group simulation, except a single significant region is forced into each group.

An odds ratio of 1.0 was used for protective and detrimental mutations with an additive mode of inheritance for 500 cases and 500 controls in each study. Each simulation incorporated 1% missingness and 1% unphenotyped individuals. The type I error was calculated as the proportion of simulated duplicates with a p-value ≤ 0.05. An error rate above 5% would indicate a higher false-positive test and an error rate lower than 5% would indicate a conservative test.

Continuous region simulation

Type I error was tested using the evolutionary parameters described above to generate random length simulated regions for two continuous region simulations (each region has single start and stop position). For continuous region simulations, each duplicate was a single region of random size and all of the variants from the simulated region were binned together in a single bin. Each result was tested with all currently available allele frequency weight options and at least two statistical tests. First, 2000 duplicate regions were simulated

with random region lengths between 2.5kb and 100kb. Over the 2000 duplicates, the number of variants in a bin varied between 42 and 4254 ($\mu = 1813, SD = 1083.17$). The size of bins (number of variants per bin) in this simulation mimicked gene bins created from an exome study in a natural data set. However, the 2.5kb-100kb duplicates did not create bins large enough to resemble bins seen in pathway analyses. To address this, the simulation was extended to create 1012 duplicate larger continuous regions with random region lengths between 100kb-500kb. The larger regions are very memory intensive to simulate; therefore, the number of duplicates was less than the number generated for smaller regions. The number of variants per bin in the larger region study varied from 13,070 to 98,450 ($\mu = 48690, SD = 19207.4$).

The type I error results are shown in Table 10. For each simulation study in the 2.5kb-100kb continuous region study, five weight options (described further in Chapter III) and four statistical tests were used. The included statistical tests were: Wilcoxon two-sample rank sum test, logistic regression, Wilcoxon rank sum test with permutations, and rank sum test with permutations. For each 100kb-500kb continuous region, five weight options and two statistical tests were applied: Wilcoxon two-sample rank sum test and logistic regression.

The weight calculated using the original Madsen and Browning implementation (CTRL) had a very high type I error in every test except permutation tests. The weights calculated from overall allele frequency (OVERALL), which is the most common implementation of the Madsen and Browning test in current online methods, are mostly well controlled, but increase slightly in analyses with larger bins. Of the novel unbiased weights, the minimum and maximum weight (MIN and MAX, respectively), the minimum weight had the lowest type I error and in most analyses was overly conservative.

Using the p-values from each of the type I error simulation results, quantile-quantile plots were generated to visualize the log p-value distribution (see Figure 7). On each plot, the size, weight, and statistical test are indicated. The null uniform distribution is shown in red. Each column represents results from a specific allele frequency weight. For example, the first column shows all of the results from the CTRL weight analyses. The first two rows are 2.5kb-100kb analysis results with 2000 replicates. The third and fourth rows are

Table 10. Type I error simulation results from continuous region simulation studies.

Study	Statistical Test	Weight	Type I error	Permutation*
2.5kb-100kb	Wilcoxon	CTRL	0.5267	
		OVERALL	0.0525	
		MAX	0.0715	
		MIN	0.048	
		NO WEIGHT	0.0505	
	Regression	CTRL	0.7856	
		OVERALL	0.0455	
		MAX	0.0685	
		MIN	0.0385	
		NO WEIGHT	0.043	
	Wilcoxon	CTRL	0.0535	Y
		OVERALL	0.0545	Y
		MAX	0.0535	Y
		MIN	0.056	Y
		NO WEIGHT	0.0505	Y
	Sum rank	CTRL	0.0495	Y
		OVERALL	0.0485	Y
		MAX	0.0475	Y
		MIN	0.051	Y
		NO WEIGHT	0.0435	Y
100kb-500kb	Wilcoxon	CTRL	0.999	
		OVERALL	0.0613	
		MAX	0.0771	
		MIN	0.0464	
		NO WEIGHT	0.0563	
	Regression	CTRL	0.999	
		OVERALL	0.0514	
		MAX	0.085	
		MIN	0.0425	
		NO WEIGHT	0.0573	

* N permutations = 1000

permutation results using 1000 permutations on each of the 2000 replicates from the same data. The fifth and sixth rows are the type I error results from the 1016 replicates of the 100kb-500kb large region analyses.

The type I error is vastly inflated in every analysis with the exception of the permutation tests. For comparison, the last column contains the results from BioBin when no weights are used and the type I error is well controlled.

Pathway simulation studies

From the simulated region tests described above, new knowledge was generated for type I error under different weight and test conditions; however, the simulations were not quite comparable to pathway analyses. Variants binned using pathway knowledge are often binned in multiple pathway bins because genes recur in multiple pathways with high frequency. This dependency could affect the type I error rate.

To address this, two additional simulation approaches were developed. Between 2-50 bins from the 2000 gene region duplicates were grouped two ways to test type I error. For both group simulations, five weight options and two statistical tests were applied: Wilcoxon two-sample rank sum test and logistic regression.

First, between 2 and 50 bins were randomly grouped together (simulating 2-50 genes in a pathway). Each 2.5kb-100kb bin could only appear once in a group simulation bin, but could appear multiple times across the 1000 group simulations. Second, this was repeated and the most significant bin (p-value $\sim 1e^{-04}$) from the null 2.5kb-100kb bin analysis was forced into each group. One thousand new groups were created by randomly combining between 2 and 50 bins including the forced the false positive bin into each group. The type I error was measured as the proportion of total replicates with a p-value of less than or equal to 0.05. In the second group analysis, since a known signal was forced into each bin, the reported error was not truly type I error. However, it was important to consider how a single bin or region of signal can be propagated to larger bins and how the signal is balanced by noise.

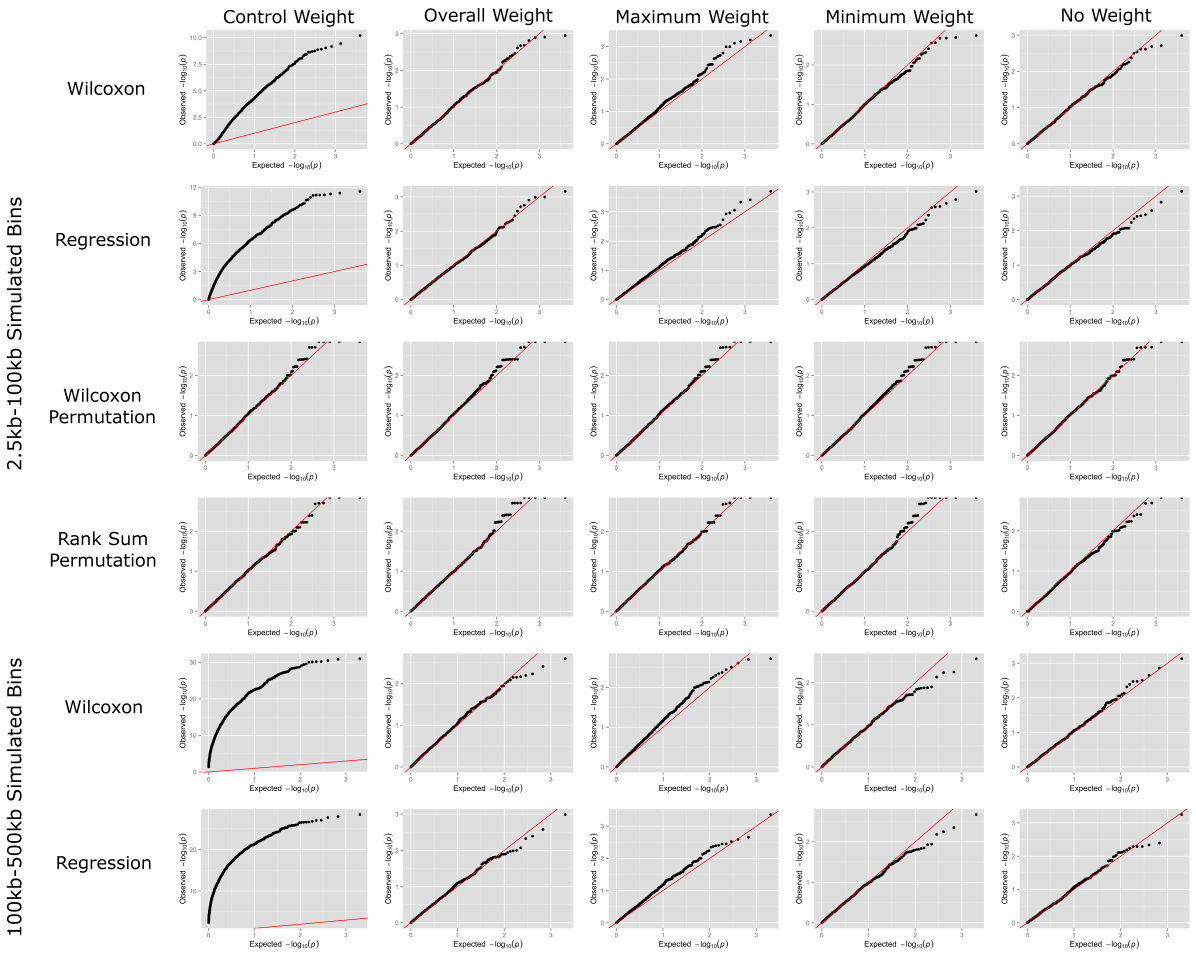


Figure 7. Quantile-quantile plots for type I error simulation studies in continuous regions. The top four rows are the QQ plots under various weighting conditions (labeled columns) and various statistical tests (labeled rows) for the smaller gene region simulation (2.5kb-100kb regions). The bottom two rows are the QQ plots under various weighting conditions (labeled columns) and various statistical tests (labeled rows) for the larger continuous region simulation (100kb-500kb regions)

Table 11. Type I error simulation results from group simulation studies.

Study	Statistical Test	Weight	Type I error
2.5-100kb : Random Groups (2-50)	Wilcoxon	CTRL	0.996
		OVERALL	0.06
		MAX	0.076
		MIN	0.051
		NO WEIGHT	0.048
	Regression	CTRL	0.997
		OVERALL	0.063
		MAX	0.079
		MIN	0.056
		NO WEIGHT	0.052
2.5-100kb : Forced Signif. Groups (2-50)	Wilcoxon	CTRL	0.999
		OVERALL	0.126
		MAX	0.148
		MIN	0.111
		NO WEIGHT	0.086
	Regression	CTRL	0.999
		OVERALL	0.128
		MAX	0.141
		MIN	0.113
		NO WEIGHT	0.081

Table 11 shows the type I error results from the group simulations. Quantile-quantile plots are shown in Figure 8. The top two rows show the p-value distribution from 1000 duplicates of the first group simulation which contained between 2-50 random bins from the 2.5kb-100kb region analyses. The last two rows show the p-value distribution from the 1000 duplicates of the group simulations which contained between 2-50 bins from the 2.5kb-100kb region analyses with the forced false positive bin in each group. As shown in Table 11 and Figure 8, the control weight does not manage type I error under any simulated conditions without permutation. In the completely random group simulation study, the other four weights manage type I error reasonably well. Maximum weight is slightly anti-conservative while the other three are conservative. In the random group with a forced signal, the type I error is always anti-conservative because the bins are not designed for a true type I error assessment. Of the four weights, minimum weight and no weight are the most conservative.

Correlation between significance and bin size

Shown in Table 10 and Table 11, the type I error regardless of weight increases in the three large bin simulation studies (100kb-500kb and two group simulations). The analyses using the MIN weight or no weight were the least affected, but type I error did still increase. It is important to understand if the increase in type I error was completely explained by the size of the bin or if the increase in type I error was compounded by bin dependency (bins are present in more than one group, which is common in pathway analyses). In each of the simulation studies, the correlation between each bin p-value and the number of variants in that bin was evaluated.

In Figure 9, each plot shows the fitted linear correlation line between bin p-values and the number of variants in that bin. The colors represent the five weight conditions used in the simulation testing. There is also a black line at $y = 0.5$ to represent the null correlation between p-value and number of variants in a bin. In each simulation study, the CTRL weight is highly affected by the number of variants in a bin because of the selection bias described in a later section. This effect is most evident in the three large simulation

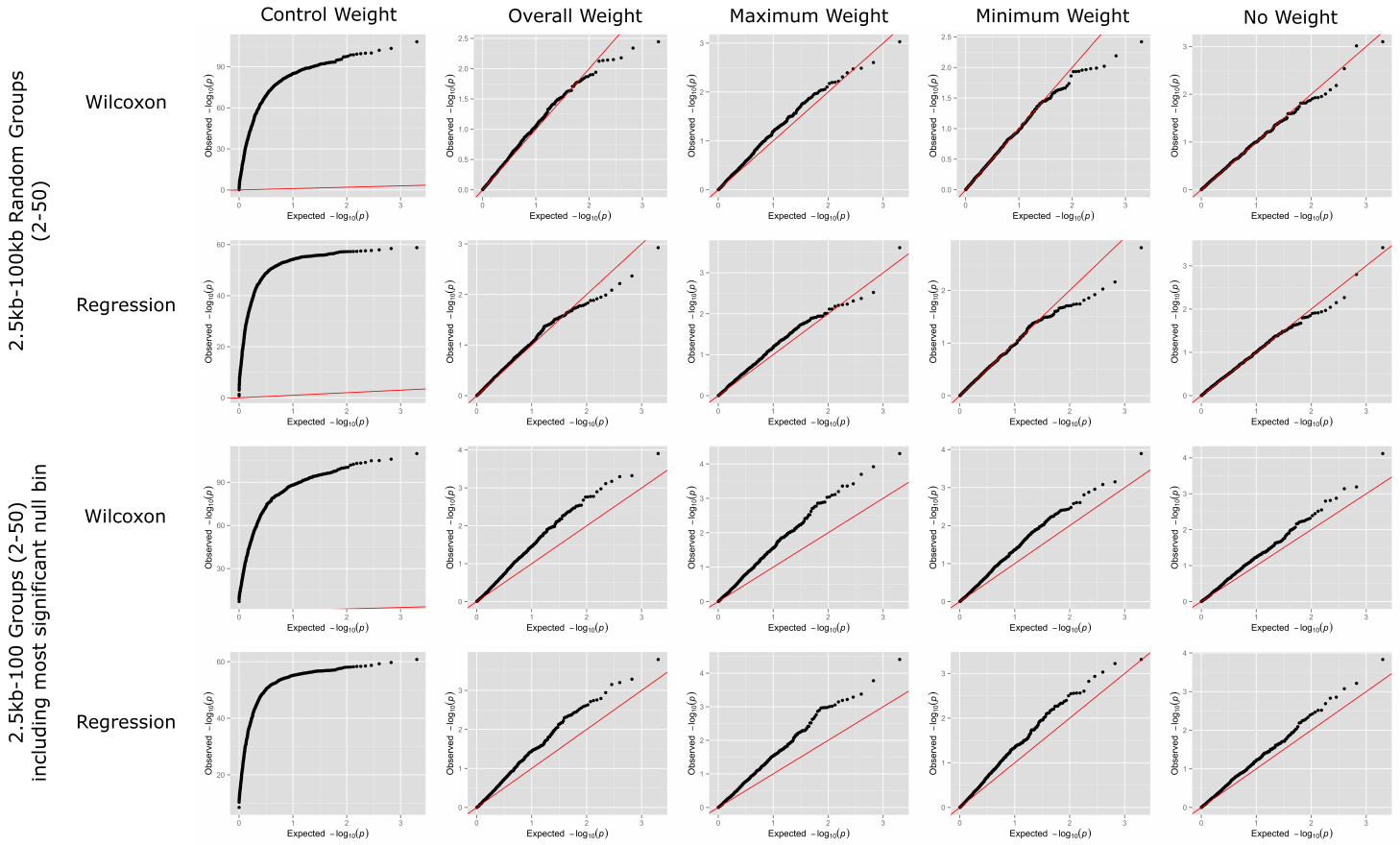


Figure 8. Quantile-quantile plots for type I error simulation studies using grouped regions. The top two rows are the QQ plots under various weighting conditions (labeled columns) and various statistical tests (labeled rows) for random groups of 2.5kb-100kb regions, where groups varied in size from 2 to 50 randomly selected regions. The bottom two rows are the QQ plots under various weighting conditions (labeled columns) and various statistical tests (labeled rows) for random groups of 2.5kb-100kb regions, where groups varied in size from 2 to 50 randomly selected regions EXCEPT one of the regions selected is necessarily the most significant bin in the null distribution.

studies, note the type I error for the large region simulation study and the two group simulation studies was over 99% (see Table 10 and Table 11). The Spearman correlation was tested in each of the four simulations with minimum variant weights applied using the regression p-value and size of bin. The top left plot shows the results from the 2.5kb-100kb continuous region analysis. The number of variants in a bin varied between 42 and 4254 ($\mu = 1813, SD = 1083.17, Spearman\ correlation\ \rho_{MIN} = -0.0258, p - value_{MIN} = 0.102$). Weights other than the CTRL weight have only a marginal decrease in p-values at the high end of gene region simulation study. The MIN weight does not appear to decrease at all, but stays with the null line at $y = 0.5$. The top right plot shows the results from the 100kb-500kb large continuous region simulation study, the number of variants per bin varied from between 13,070 to 98,450 ($\mu = 48690, SD = 19207.4, Spearman\ correlation\ \rho_{MIN} = 0.00729, p - value_{MIN} = 0.743$). The non-CTRL weights do not show a strong trend between p-value and the number of variants in a bin. The bottom left plot shows the results from the group simulation study, the number of variants per bin varied between 922 variants to 106,300 variants ($\mu = 47670, SD = 26325.97, Spearman\ correlation\ \rho_{MIN} = -0.0203, p - value_{MIN} = 0.365$). Without regard to the CTRL weight, each of the other weight conditions show a minor decreasing trend, indicating that larger group bins contain dependencies with slightly lower p-values. The bottom right plot shows the results from the 2.5kb-100kb group simulation study, where the most significant null bin was forced into each group. The number of variants per bin varied from 3330 to 115,000 variants ($\mu = 48400, SD = 27206.31, Spearman\ correlation\ \rho_{MIN} = 0.287, p - value_{MIN} = 3.9e^{-39}$). There is a noticeable correlation between p-value and bin size. As the bin size increases, the signal from the single false positive bin is mitigated.

Non-bias variant weighting comparison

The simulation results indicate that relying on the variant weights based entirely on controls drastically increases the type I error in all statistical tests except permutation testing. Even worse, this bias is magnified with the size of the bin, introducing a spurious correlation that

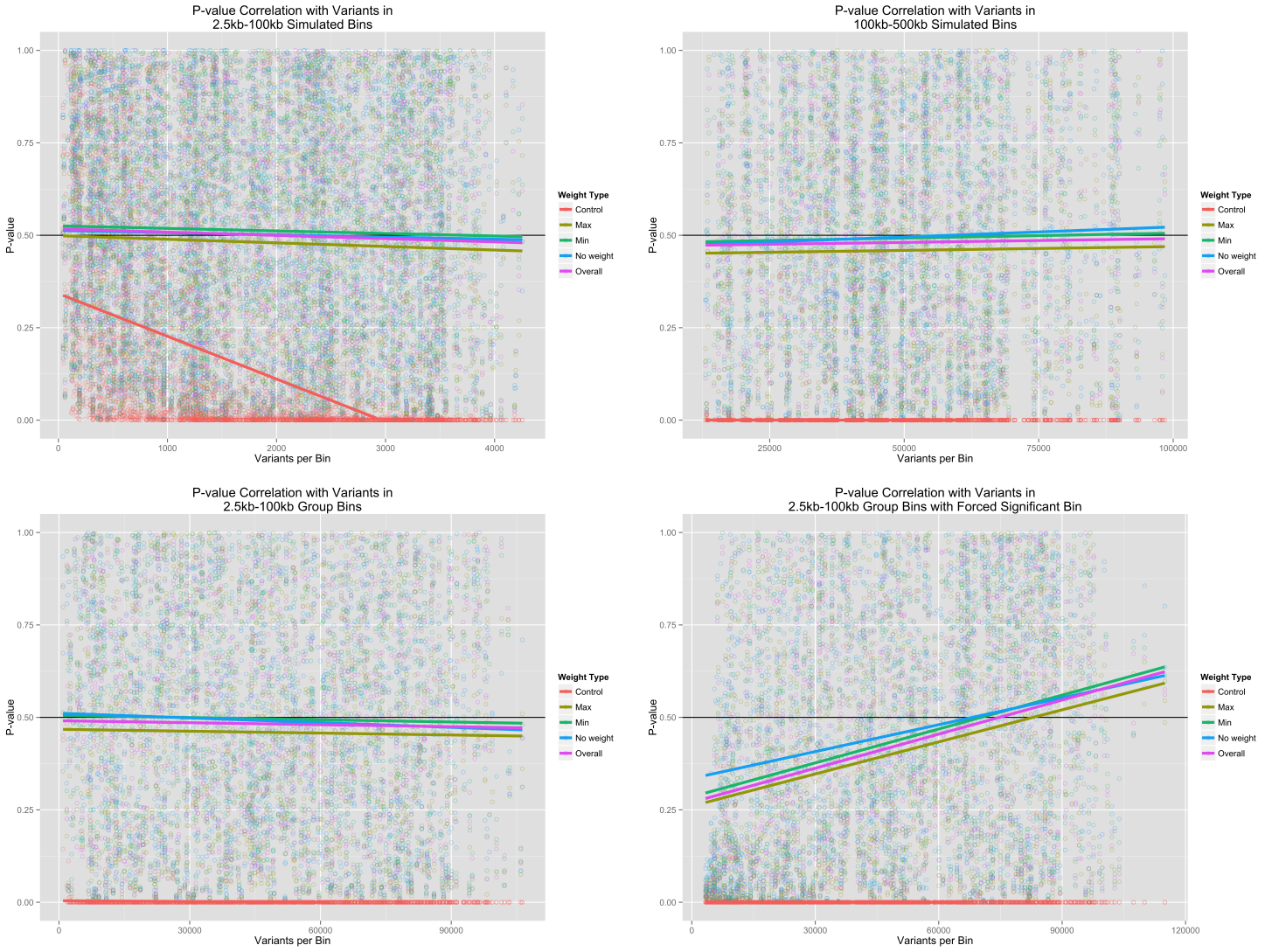
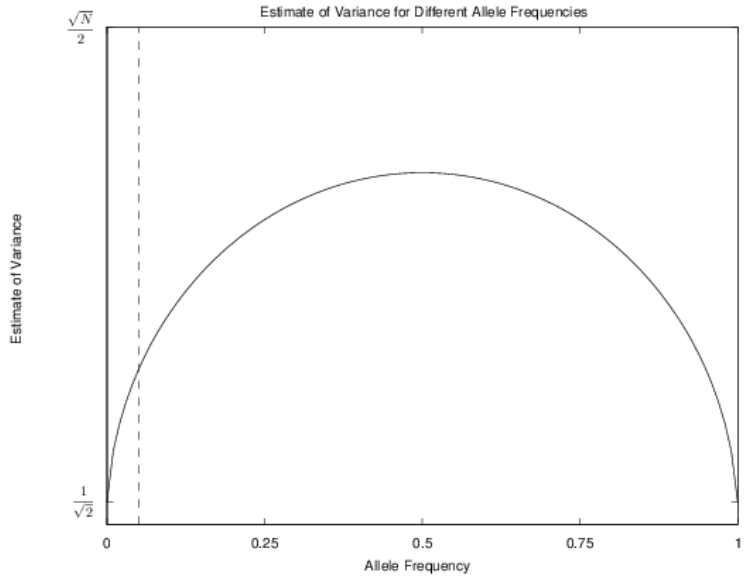


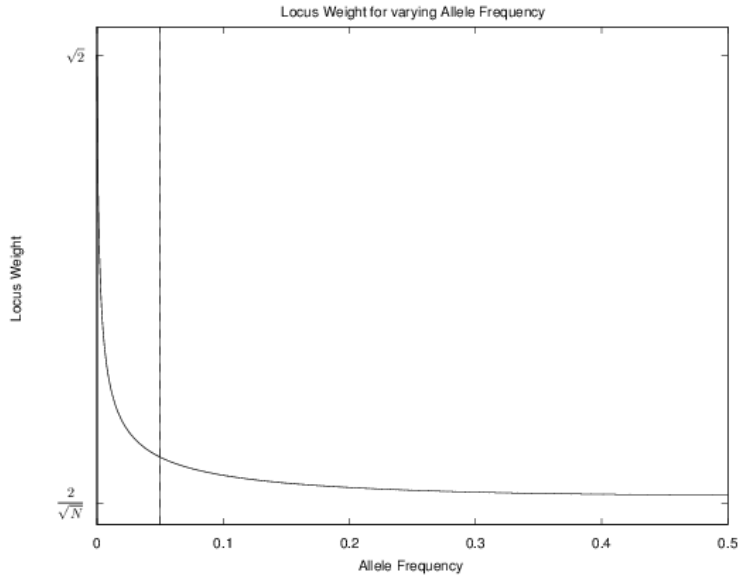
Figure 9. Correlation between bin p-value and number of variants in the bin. The results are presented using the Wilcoxon p-values for each of the weight conditions.

can confound results. By weighting solely on the control population, the user introduces a bias similar to the bias described by Lemire [77]. Three options for weighting were implemented that eliminate selection bias. The most common method is to weight loci by their overall frequency in the case and control population combined. While this reduces type I error, it also reduces statistical power. In a recent population comparison of low frequency variants, many loci that were nearly fixed in opposite directions were found. If one used the overall weight, these loci would be extremely down-weighted, even though they are incredibly relevant [83].

The other two unbiased methods use either the minimum or the maximum of the weights calculated for each population individually (e.g. $\min(\frac{1}{w_i^U}, \frac{1}{w_i^A})$ or $\max(\frac{1}{w_i^U}, \frac{1}{w_i^A})$. In simulation tests, the minimum weight better controlled the type I error and in most cases was overly conservative. Figure 10 shows the Madsen and Browning estimate of variance (w_i) and the respective locus weight for large populations and varying allele frequencies. As can be seen in Figure 10, when using the locus maximum weight with a minor allele frequency cutoff (vertical line), the locus weights are constrained within a limited range; however, there is no such constraint when using the minimum weighting option.



(a) Variance



(b) Weight

Figure 10. Estimate graphs of variance (w_i) and locus weight ($\frac{1}{w_i}$) for varying allele frequencies. The vertical lines correspond to a 5% minor allele frequency.

Power simulation assessment

Varying sample size

In addition to the parameters described in the Simrare section, to evaluate power, a sample data set was generated with the following parameters: fixed 5kb simulated region, 0.9 odds ratio for protective mutations, 2.5 odds ratio for detrimental mutations, and an additive mode of inheritance. A protective odds ratio was 0.9 to add noise to the data. The detrimental odds ratio was designated as 2.5 to provide a conservative estimate of power. Most literature reviews expect causative low frequency alleles to have odds ratios ≥ 2 [2]. Four sample sizes were created: 2000, 1000, 500, 250. Case/control status was evenly and randomly assigned in each of the 4000 duplicates. Missingness or unphenotyped individuals were not incorporated. The power was computed as the percentage of the 4,000 duplicates with a p-value ≤ 0.05. As shown in Table 12, the power is greater than 90% at sample sizes of 1000 and 2000 individuals. The power drops to 75% at a sample size of 500 and 50% at a sample size of 250. While the power drops dramatically with decreasing sample size, it is important to note that power to detect associations relies heavily on the effect size of the variants.

Table 12. Power analysis using fixed 5kb simulations with 4000 replicates for each of four sample sizes (N=2000, 1000, 500, and 250). Note: for each sample size, the number reflects the total number of individuals (i.e. N=2000 translates to 1000 cases and 1000 controls).

Sample Size	**Power**
2000	0.9910
1000	0.9340
500	0.7575
250	0.5030

Unequal sample size and comparison with other methods

The last test was designed to generate a sample data set evaluating power similar to the cystic fibrosis study sample (see Chapter VI), the following parameters were used: randomly generated bins ranging from 1kb to 10kb; 0.9 odds ratio for protective mutations; 2.5 odds ratio for detrimental mutations; and an additive mode of inheritance for 100 cases and 300 controls under three scenarios: 100% functional variants, 50% functional variants, and 25% functional variants. In this case, missingness or unphenotyped individuals were not incorporated. The power was calculated as the proportion of the 1,000 duplicates with a p-value ≤ 0.05.

The power of BioBin using multiple weight conditions in a logistic regression and Wilcoxon two-sample rank sum test framework was compared with other published methods: combined multivariate and collapsing method (CMC), Kernel-based adaptive cluster test (KBAC), rare variant threshold test (MZ), weighted sum statistic (WSS) and variable threshold test (VT) [81]. The power analysis results are shown in Figure 11, Figure 12, and Figure 13. The corresponding type I error for each method is shown in text on each graph. The most powerful BioBin result uses CTRL weight, which has a tremendous type I error rate (red line). The most powerful BioBin weight with controlled type I error is the minimum weight using a logistic regression test; it is second only to the variable threshold method first described by Price et al. [56].

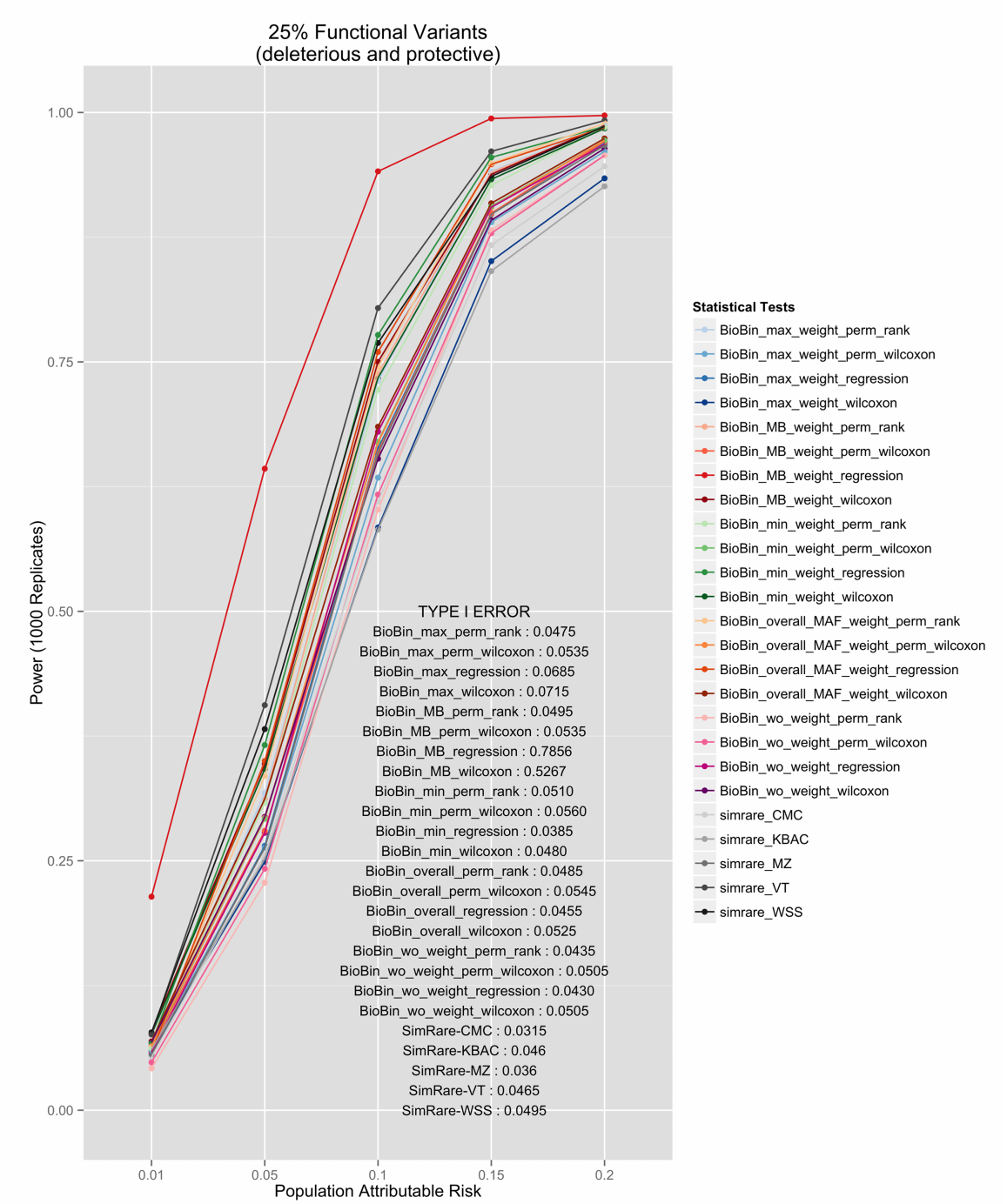


Figure 11. Power estimates for multiple binning strategies on simulated data with 100 cases and 300 controls where only 25% of variants are functional. For reference, the corresponding Type I error values are provided for each test.

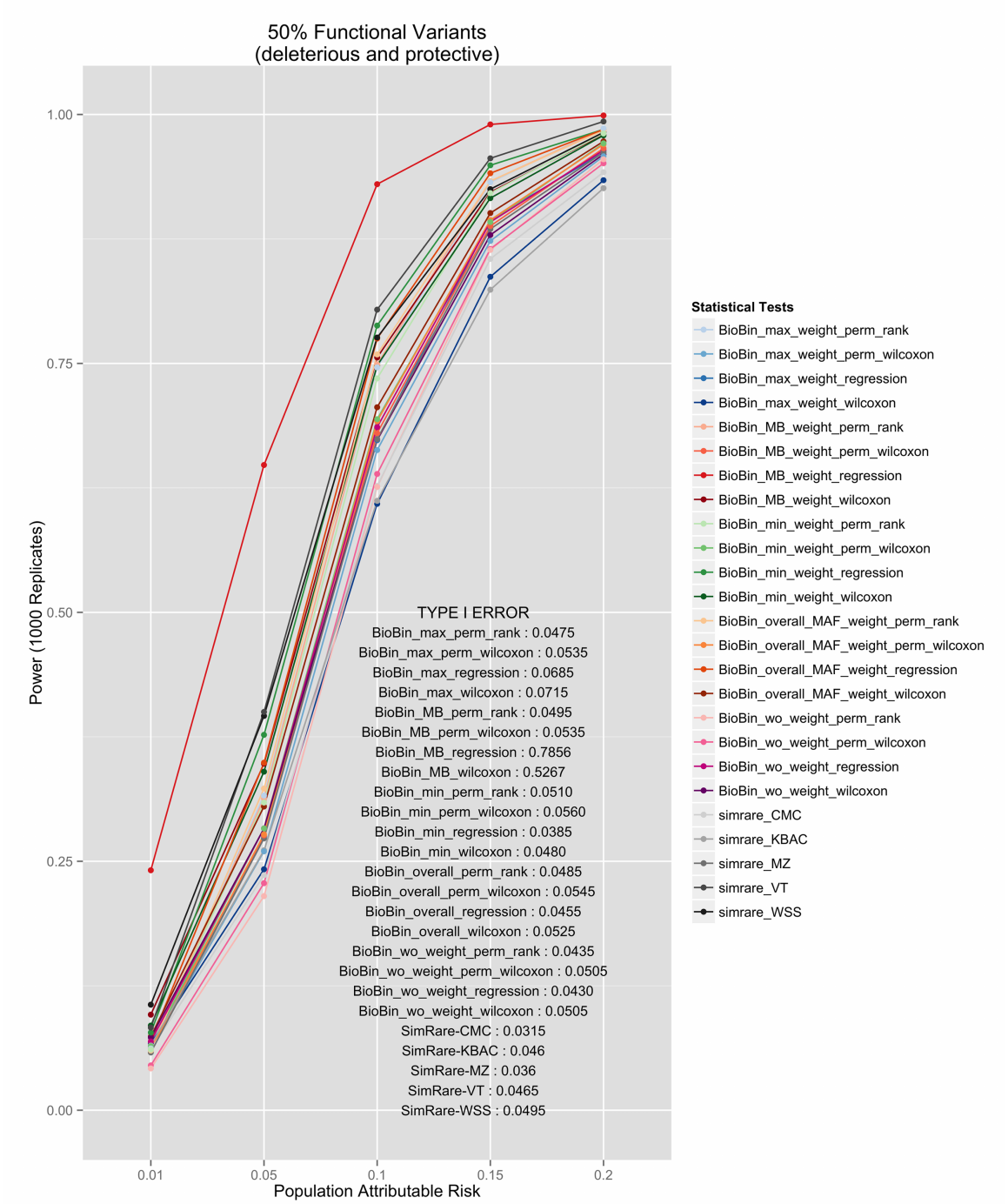


Figure 12. Power estimates for multiple binning strategies on simulated data with 100 cases and 300 controls where only 50% of variants are functional. For reference, the corresponding Type I error values are provided for each test.

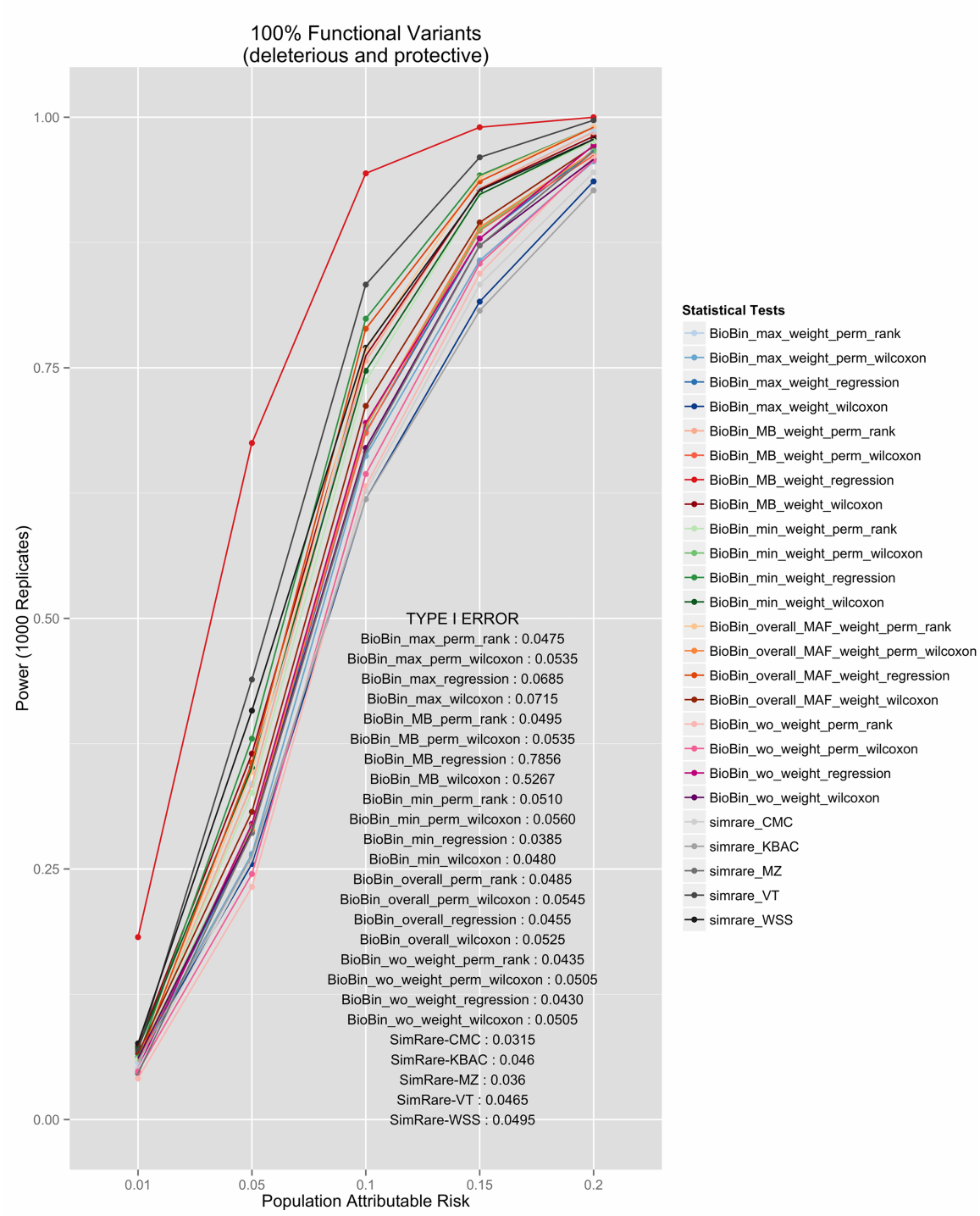


Figure 13. Power estimates for multiple binning strategies on simulated data with 100 cases and 300 controls where 100% of variants are functional. For reference, the corresponding Type I error values are provided for each test.

Summary

The power and type I error estimates are dependent on sample size, effect size, and BioBin parameters. Minimum variant weighting was the most powerful and most conservative weighting framework. Permutations were used for some of the calculations provided but were not necessary to maintain reasonable type I error rates as long as the variant weights were unbiased. Lastly, bin size does not appear to increase type I error. Dependency within bins (such as found in pathway analyses) can minimally increase the type I error. Additionally, a child bin or variant with a strong signal can propagate signal through larger parent bins in a size dependent manner. Overall, the bins provided by BioBin are most powerful if the total sample size is greater than 500 individuals and minimum weighting is used.

CHAPTER V

1000 GENOMES PROJECT DATA: POPULATION COMPARISON OF LOW FREQUENCY BURDEN

Low frequency variants are likely to play an important role in uncovering complex trait heritability; however, they are often population specific or unique to populations within a continent. This specificity complicates genetic analyses investigating low frequency variants for two reasons: low frequency variant signals in an association test are often difficult to generalize beyond a single population or continental group and there is an increase in false positive results in association analyses due to underlying population stratification. In order to reveal the magnitude of low frequency population stratification, pairwise population comparisons were performed using the 1000 Genomes Project Phase I data to investigate differences in low frequency variant burden across multiple biological features.

Methods and results

Binning approach

NCBI dbSNP and NCBI Entrez Gene were chosen as the primary sources of position and regional information [61]. Pathway/group bins, regulatory regions, and evolutionary conserved regions were created using sources available in LOKI (sources detailed in Chapter III). Some sources explicitly provide lists of genes in pathways, others provide groups of genes which share a biological connection (e.g. protein-protein interactions). For the purposes of this study, any bin created by multiple regions/genes was analyzed in the Pathway-Groups feature analysis. External custom input files were generated using boundaries of annotated exon regions from UCSC to bin exon and intron specific variants. For example, if Gene A

Table 13. Excerpt of custom region file containing regions with signatures of natural selection

Chr	ID	Start(bp)	Stop(bp)
1	CMS_EUR:reg1	1490074	1509034
1	CMS_ASN:reg2	16414784	16417992
1	CMS_AFR:reg3	26917014	26936774
1	CMS_EUR:reg4	30707291	30724056

has three exons and two introns, only two bins would be created: GeneA-exons and GeneA-introns. GeneA-exons would contain all variants that fell within any of the three Gene A exon boundaries. External custom feature files were also generated for regions under natural selection by combining regions provided by previously published work [84, 85]. An excerpt of the custom region natural selection file is shown in Table 13.

Statistical analysis

BioBin is a bioinformatics tool used to create new feature sets that can then be analyzed in subsequent statistical analyses. Statistical tests used with BioBin can be chosen according to the hypothesis being tested, the question of interest, or the type of data being tested. Unless otherwise noted, the results presented here were calculated using a Wilcoxon 2-sample rank sum test implemented and graphed in the R statistical package [86, 87]. P-values presented have been corrected using a standard Bonferroni correction, adjusting for the number of bins created and tested in a given analysis.

1000 Genomes Project data

To investigate low frequency variant population stratification using BioBin, 1000 Genomes Project Phase I data were analyzed. The 1000 Genomes Project was started in 2008 with the mission to provide deep characterization of variation in the human genome. As of October 2011, the sequencing project included whole-genome sequence data for 1094 individuals, and aimed to sequence 2,500 individuals by its completion [88]. Table 14 provides the total

number of variants (common and low frequency) and individuals included in Phase I VCF files of 1000 Genomes Project data for 1094 individuals in all 14 populations. Cryptically related individuals (N=75) were removed and a pairwise comparison of low frequency variant burden differences between 14 populations was conducted.

In addition to the differences in overall magnitude of variation between these population groups, there were also differences in the distribution of this variation. In Figure 14, the allele frequency density distribution plot of chromosome 1 for all 14 populations is presented. On chromosome 1, African descent populations have the highest density of low frequency variation. Others have found a similar trend genome-wide [60]. In general, the African ancestral populations not only have more variants overall than other ancestral groups (see Table 14), these populations also have a higher distribution of low frequency variants than other ancestral groups (see Figure 14). Although the number of individuals in a given population affects the identification of low frequency variants, the trends seen in Figure 14 reflect ancestry, not population size. The cyclic blue line corresponds to the Iberian population, which only contains 14 individuals. The smaller sample size is responsible for discrete allele frequency values and irregular allele frequency distribution.

Although low coverage next generation sequence data are prone to errors, no evidence exists to support the theory that sequence technology led to differential bias in a way that could explain the trends found in this chapter. In a recent publication by the 1000 Genomes Project, the authors declared sequence errors to be relevant to the technology used but not to have any correlation with population identity. In order to determine if there was differential bias across populations or continental groups based on sequence technology, principal component analyses in each continental group were performed and global variation differences in the context of sequence technology were reviewed (similar approach to recent 1000 Genomes Consortium paper). Sequence technologies used for each population in the Phase I release (see Table 15) [60] were examined. Before removing the 75 cryptically related individuals, only the TSI population was sequenced on a single technology. However, after dropping cryptically related individuals, CHB, CHS, and JPT were also sequenced exclusively with Illumina technology.

In Figure 15, the first two principal components calculated from each of the four conti-

Table 14. 1000 Genomes Project Phase I data characteristics. Fourteen populations released in the Phase I 1000 Genomes Project data release, including the continental group, population abbreviation (POP), short description of each population (POPULATION), number of individuals (N), number of cryptically related individuals dropped in final analyses (REL), total number of loci, variants, low frequency variants (MAF<=0.03), and private variants (unique to population). Only autosomal variants were considered. The total loci column refers to the number of variant lines in the VCF file, but not all of these lines contained variants, likely due to filtering and missing data.

Continental Group	**POP**	**Relevant Population**	**N**	**REL**	**Total Loci**	**Total Variants**	**Low Freq Variants**	**Private Variants**
African descent (AFR)	ASW	HapMap African ancestry individuals from SW US	61	5	18819173	18762530	7948290	1059215
	LWK	Luhya individuals	97	10	19936728	19857956	8781777	2600039
	YRI	Yoruba individuals	88	0	18022152	17926400	7328288	1032847
Asian descent (ASN)	CHB	Han Chinese Beijing	97	16	10566371	10292757	3673350	860493
	CHS	Han Chinese South	100	16	10547019	10251069	3872508	1102270
	JPT	Japanese individuals	89	17	10368186	10063756	3535488	1233969
European descent (EUR)	CEU	CEPH individuals	87	0	11198921	10994490	4028071	520730
	FIN	Finnish individuals	93	0	11005104	10799742	3549441	524199
	GBR	British individuals from England and Scotland	89	3	11411688	11212275	4064515	576664
	TSI	Toscan individuals	98	0	11858607	11668150	4502592	818043
Spanish/Mexican descent (SPN)	CLM	Colombian in Medellin, Columbia	60	1	13869201	13753047	6063724	729009
	IBS	Iberian population in Spain	14	0	8424366	8155987	0	129800
	MXL	Mexican individuals from LA, California	66	7	12929352	12788406	5322835	840056
	PUR	individuals from Puerto Rico	55	0	14066653	13958200	6266201	561551

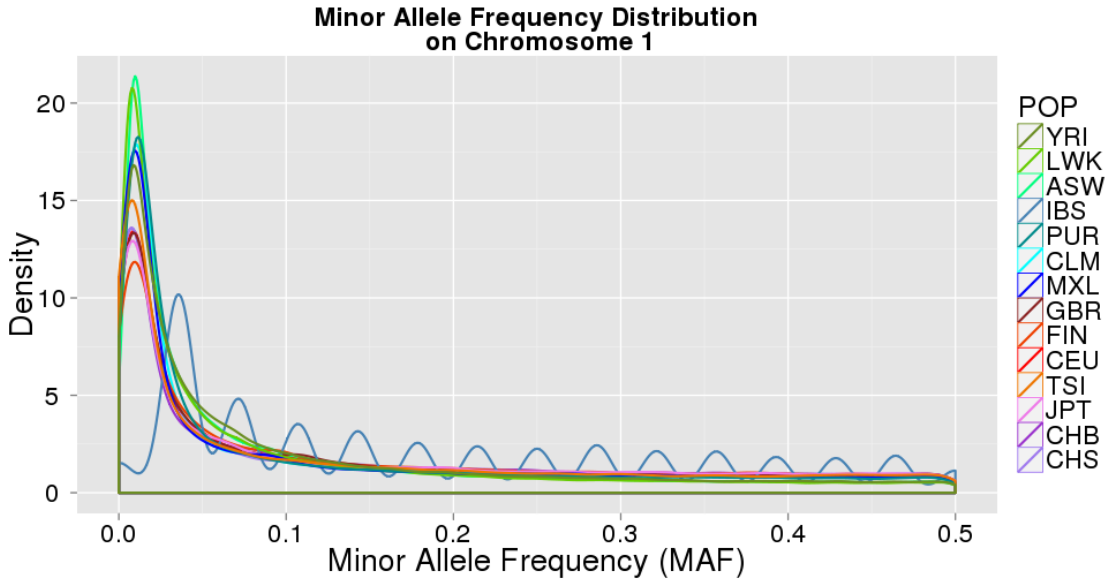


Figure 14. Minor allele frequency distribution on chromosome 1 for 14 1000 Genomes Project Phase I populations. Groups are color coordinated by continental ancestry: greens=African descent (YRI, LWK, ASW); blues=Mexican/Spanish descent (IBS, PUR, CLM, MXL); orange/reds=European descent (GBR, FIN, CEU, TSI); and pink/purple colors=Asian descent (JPT, CHB, CHS). The populations of African descent have the highest proportion of low frequency variation. The cyclic blue line is the IBS (Iberian) population, which only contains 14 individuals. This reduces the overall available spectrum of variant frequency; IBS are thus an outlier in many of the presented analyses.

Table 15. Phase I 1000 Genomes Project sequence technology data characteristics. 1000 Genomes Project Phase I populations (1094 individuals, 14 populations) and number of individuals from each population sequenced on ABI solid, Illumina, LS454 or Illumina and LS454.

Continental Group	**Population**	**ABI_SOLID**	**ILLUMINA**	**ILLUMINA-LS454**	**LS454**
African descent (AFR)	ASW	11	50	0	0
	LWK	14	83	0	0
	YRI	12	76	0	0
Asian descent (ASN)	CHB	16	81	0	0
	CHS	8	92	0	0
	JPT	11	78	0	0
European descent (EUR)	CEU	0	72	9	6
	FIN	18	75	0	0
	GBR	19	70	0	0
	TSI	0	98	0	0
Spanish/Mexican descent (SPN)	CLM	10	50	0	0
	IBS	8	6	0	0
	MXL	12	54	0	0
	PUR	3	52	0	0

nental groups shown in Table 15 are plotted. The scatter plots are colored using population identity and then sequence technology. From Figure 15, it is clear that within continental groups, the largest source of variation is sequence technology. In all four groups, the first principal component perfectly separates based on technology. However, the variation does not also coincide with population identity and there is overlap between populations since few populations were sequenced on a single technology. This reduces the likelihood that sequence technology causes differential bias in the resulting trends of the presented analyses (see Table 15 and Figure 15).

Investigation of allele sharing

In any genetic study, and especially in consideration of low frequency variants, it is important to evaluate sample relatedness and allele sharing. To accomplish this, identity-by-descent (IBD) was investigated in very common variants (MAF $> 5\%$), to assess relatedness. Second, a more traditional method of assessing cryptic relatedness was used, LD pruned variants with MAF $> 5\%$ in continental groups were used to parsimoniously eliminate cryptically related individuals. Third, identity-by-state (IBS) within populations was assessed with and without cryptically related individuals. Lastly, IBS was calculated within continental groups in low frequency variants (MAF $< 5\%$) and very common variants (MAF $> 25\%$).

Population groups were combined into continental populations (i.e. AFR continental group included ASW, LWK, YRI) and sample relatedness was evaluated between and within the general ancestry groups using identity-by-state (IBS) and identity-by-descent (IBD). Pairwise IBS represents the number of shared alleles at a specific locus between two individuals. IBS can be observed as 0, 1, or 2 depending on how many alleles are in common between the pair. If the shared alleles are inherited from a recent common ancestor, they are also considered IBD. Pairwise IBS calculations for low-frequency variants approximate IBD since the variants are likely to be recent and the chance of being identical because of recurrence is rare [89]. PLINK and PLINK-SEQ were used to estimate pairwise IBS and IBD

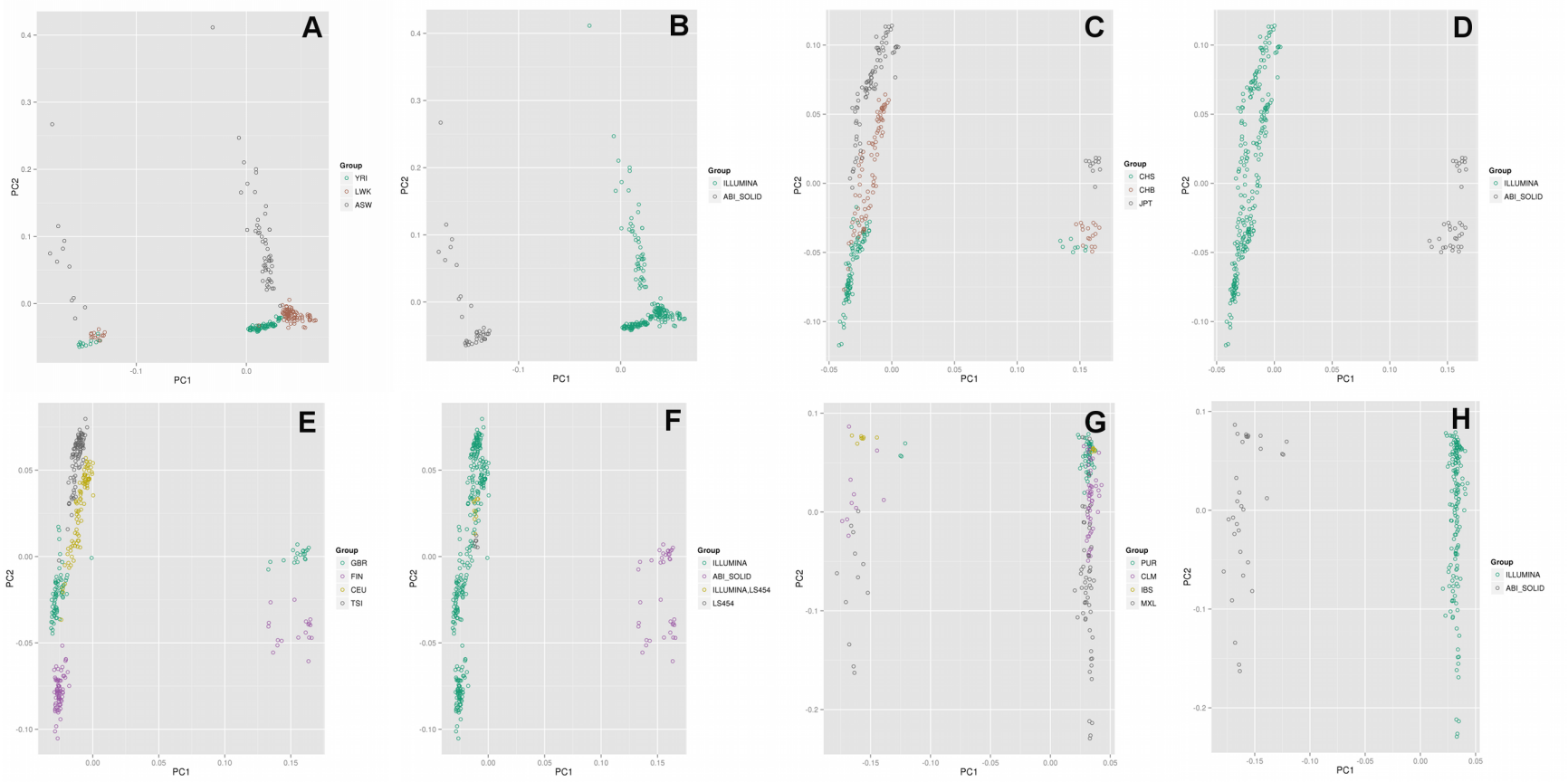


Figure 15. Investigating differential bias in 1000 Genomes Project data using principal components analysis. Each plots shows the first two principal components calculated from each continental group colored by population identity (A, C, E, G) and sequence technology (B, D, F, H). The labels correspond to populations from four continental groups: (A/B) AFR continental group, (C/D) ASN continental group, (E/F) EUR continental group, and (G/H) SPN continental group. Since the global variation is caused primarily by sequence technology, and very few populations are actually sequenced on a single technology, sequence technology likely contributes little bias to the trends seen in the presented results.

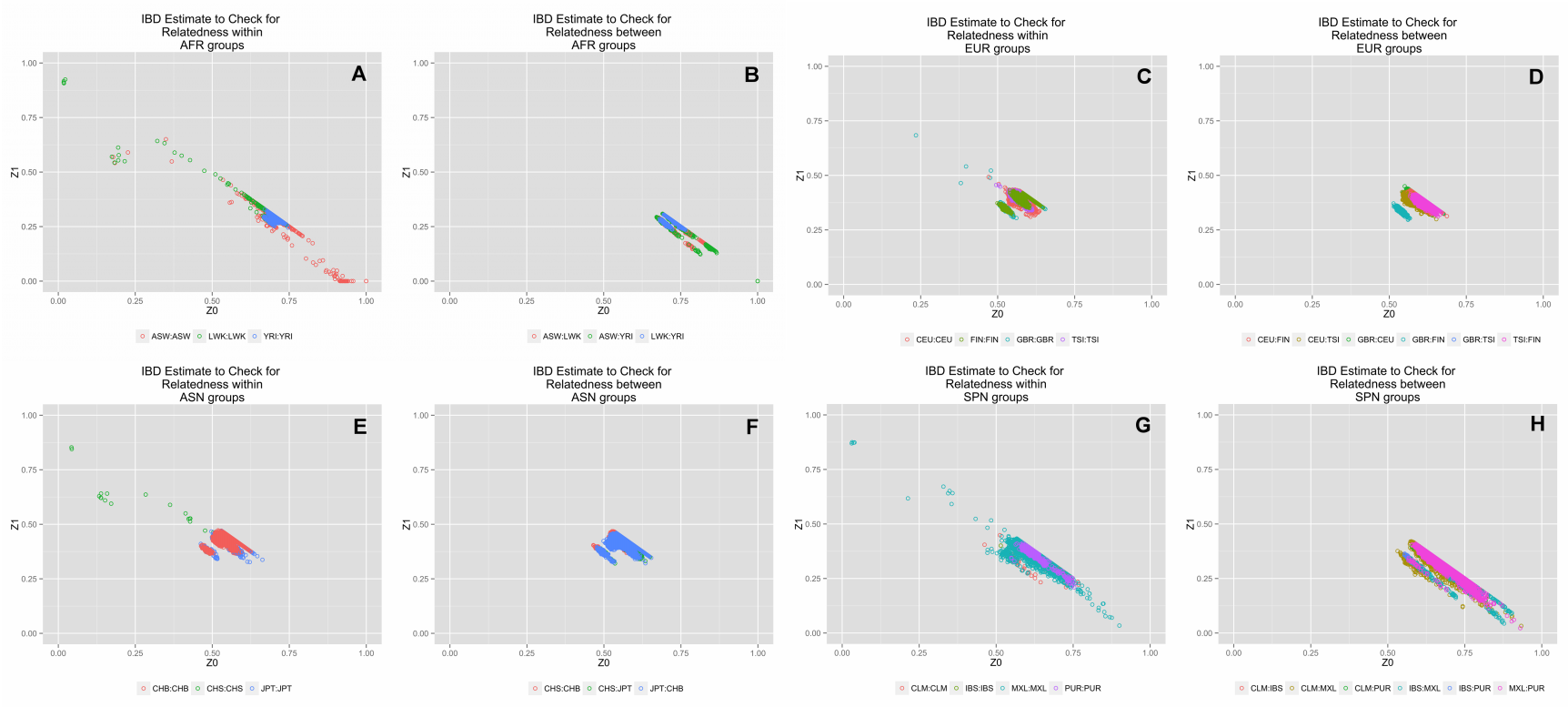
for individuals of the same general ancestry group (http://atgu.mgh.harvard.edu/plinkseq/, http://pngu.mgh.harvard.edu/ purcell/plink/) [78]. In each analysis, evidence of increased relatedness was found in ASW (African ancestry, USA), CHB (Han Chinese Beijing, China), CHS (Han Chinese Shanghai, China), CLM (Medellin, Columbia), GBR (England and Scotland), JPT (Japan), LWK (Luhya, Kenya), and MXL (Mexican Ancestry, California) populations.

Identity by descent (IBD) in all 1094 individuals

In most genomic studies, subject relatedness is calculated using common variants. Therefore, in the first allele sharing analysis, IBD was estimated using only common variants (MAF $>$ 10%) in all 1094 individuals available in the Phase I release. In Figure 16, the y-axis and x-axis correspond to the proportion of markers identical by descent between a pair of individuals sharing one allele versus none. Small clusters of individuals in the top left quadrant correspond to more allele sharing than expected from unrelated individuals. For example, the green points in the left plot of Figure 16A represent a subset of approximately 10 related individuals in the LWK population. Each point represents the IBD estimate between two LWK individuals. The pairs that share one allele at almost 100% of possible loci (Z1$\sim$1) and share at least one allele at all loci (Z0$\sim$0) represent a parent-child relationship. Siblings cluster near the center of the plot while pairwise IBD estimates for completely unrelated individuals cluster in the lower right quadrant. In Figure 16, there are several within-population plots that show increased allele sharing within population groups (A,C,E,G). However, there does not appear to be any increased sharing between population groups (B,D,F,H). For example, even though there are individuals in the LWK population that appear to be related (see Figure 16A), none of the three African descent populations appear to have closely related individuals across populations (i.e. LWK-YRI are not related, Figure 16B).

In Figure 16, the pairwise calculations with a proportion of IBD greater than 0.5 were composed of 16 LWK individuals, each with evidence of first and second-degree relationships within the LWK population. Two of the four top related IBD pairwise comparisons in LWK have been calculated in other studies as parent-child relationships [90].

Figure 16. IBD estimates using variants with MAF $> 10\%$ within and between ancestral groups. There are two frames for each ancestral group. The labels correspond to populations from four continental groups: (A/B) AFR continental group, (C/D) ASN continental group, (E/F) EUR continental group, and (G/H) SPN continental group. The left frame from each continental group corresponds to the IBD estimate within each population (A, C, E, G). The right frame from each continental group corresponds to the IBD estimate between populations within the ancestral group (B, D, F, H).



Most of the apparent relationships in the IBD plots above have been identified previously and are available on the 1000 Genomes Project website [`http://www.1000genomes.org/phase1-analysis-results-directory`, cryptic relation analysis].

Evaluating cryptic relatedness

For common variants, an independent subset of SNPs with a minor allele frequency greater than 5% and r^2 linkage disequilibrium values less than 0.2 was created to calculate pairwise IBD between individuals. For example, the populations of African descent (LWK, ASW, and YRI) were grouped, and the IBD calculated using all of the individuals from these three populations. Maximally connected or related individuals were removed in a parsimonious and iterative manner and the IBD analysis was repeated until the maximum pairwise pi_hat score was less than or equal to 0.3. After repeating this analysis in each continental group, 75 individuals were dropped from BioBin analyses based on the threshold for cryptic relatedness. The remaining 1,019 individuals were used for the binning analyses presented in this paper.

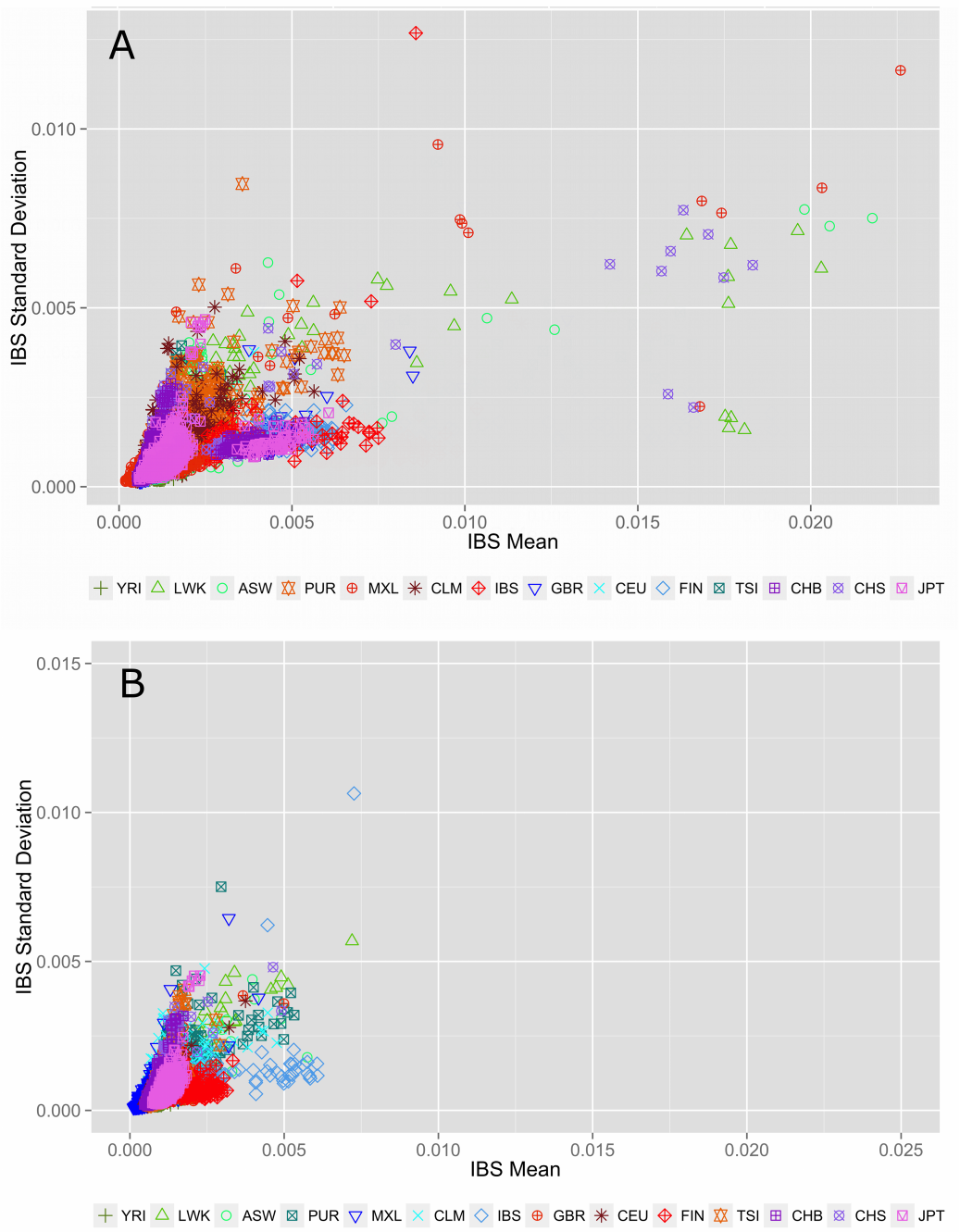
Within population identity by sharing (IBS)

An alternate allele sharing method described by Abecasis et al. uses IBS rather than IBD to review allele sharing [91, 92]. In the case of low frequency or rare variants, IBS approximates IBD. Figure 17 shows within population IBS for all 14 populations for variants with a MAF $< 3\%$, where each point represents a pairwise IBS calculation *within* the same population (i.e. IBS calculation between YRI-YRI individuals but not YRI-CEU individuals). In Figure 17A, the pairs with average IBS calculations that fall outside of the cluster are cryptically related individuals with increased allele sharing. Figure 17B shows the IBS calculations after removing 75 individuals with cryptic relatedness.

Within continental group, IBS calculations in low frequency and common variants

Allele sharing was evaluated between major ancestral groups using PLINK-SEQ to calculate IBS for low frequency variants and common variants (threshold MAF $< 3\%$ and MAF $> 25\%$, respectively). Using the ratio of shared alleles divided by the total number of

Figure 17. Within population identity-by-state (IBS) estimations A) before and B) after removing individuals with cryptic relatedness. The x-axis represents the IBS mean for low frequency variants averaged over 22 autosomal chromosomes. The y-axis corresponds to the standard deviation of IBS scores across 22 autosomal chromosomes. The colors and point types correspond to each population; color schemes correspond to general ancestry groups as defined for Figure 14. Each point represents a population pairwise IBS calculation (i.e. YRI-YRI, not YRI-CEU). Identifying and excluding related individuals removes the outliers seen in the top plot.



genotyped alleles between two individuals, excess sharing of low frequency variants was compared to excess sharing of common variants. Again, there was increased sharing among ASW, CHB, CHS, CLM, GBR, JPT, LWK, and MXL populations before removing the 75 cryptically related individuals.

Figure 18 shows the mean IBS calculations (averaged across 22 autosomal chromosomes) in low frequency variants for all pairwise individuals within a continental ancestry group. The left plot (Figure 18A) corresponds to the IBS calculations for all 1094 individuals; the right plot (Figure 18B) shows the IBS calculations after removing cryptically related individuals. The x-axis corresponds to the index number comparison; each x index value represents one pairwise comparison. The comparisons are grouped and colored by type (i.e. CHS-CHS and CHS-JPT). The y-axis corresponds to the mean IBS calculation across all 22 autosomal chromosomes. In Figure 18, low IBS means correspond to very little allele sharing for variants with MAF < 0.03. Higher IBS means correspond to more allele sharing (and perhaps relatedness) among individuals in that pair. For example, there is increased sharing of alleles with $< 3\%$ MAF among LWK pairs (teal peaks, Figure 18A).

Common variant IBS calculations alone overestimate IBD; however, the analysis was repeated for common variants, results are shown in Figure 19. For common variants, the IBS calculations were measured using only variants with a continental group minor allele frequency of 25% or higher. In Figure 19, the left plot shows the IBS calculations in all 1094 individuals; the right plot shows the IBS calculations after removing 75 cryptically related individuals. The same peaks of increased sharing in LWK, ASW, GBR, CHS, and MXL are seen and the removal of those cryptically related individuals reduces the amount of sharing in those populations.

Genomic feature exploration

After determining which individuals to exclude from this study, the feature options of BioBin were used to investigate a variety of biologically relevant bins for differences in low frequency variant burden across 14 populations. Feature selection in BioBin is a clear innovation over

Figure 18. Pairwise IBS calculations for low frequency variants (MAF < 3%) within continental groups. Plots (A-D) show the IBS calculations within continental groups for all 1094 individuals. The plots to the right (E-H) show the IBS calculations within continental groups for 1019 individuals (cryptically related individuals removed). Each dot represents a pair of individuals; the colors correspond to population comparisons. Points with the higher mean IBS indicate increased sharing.

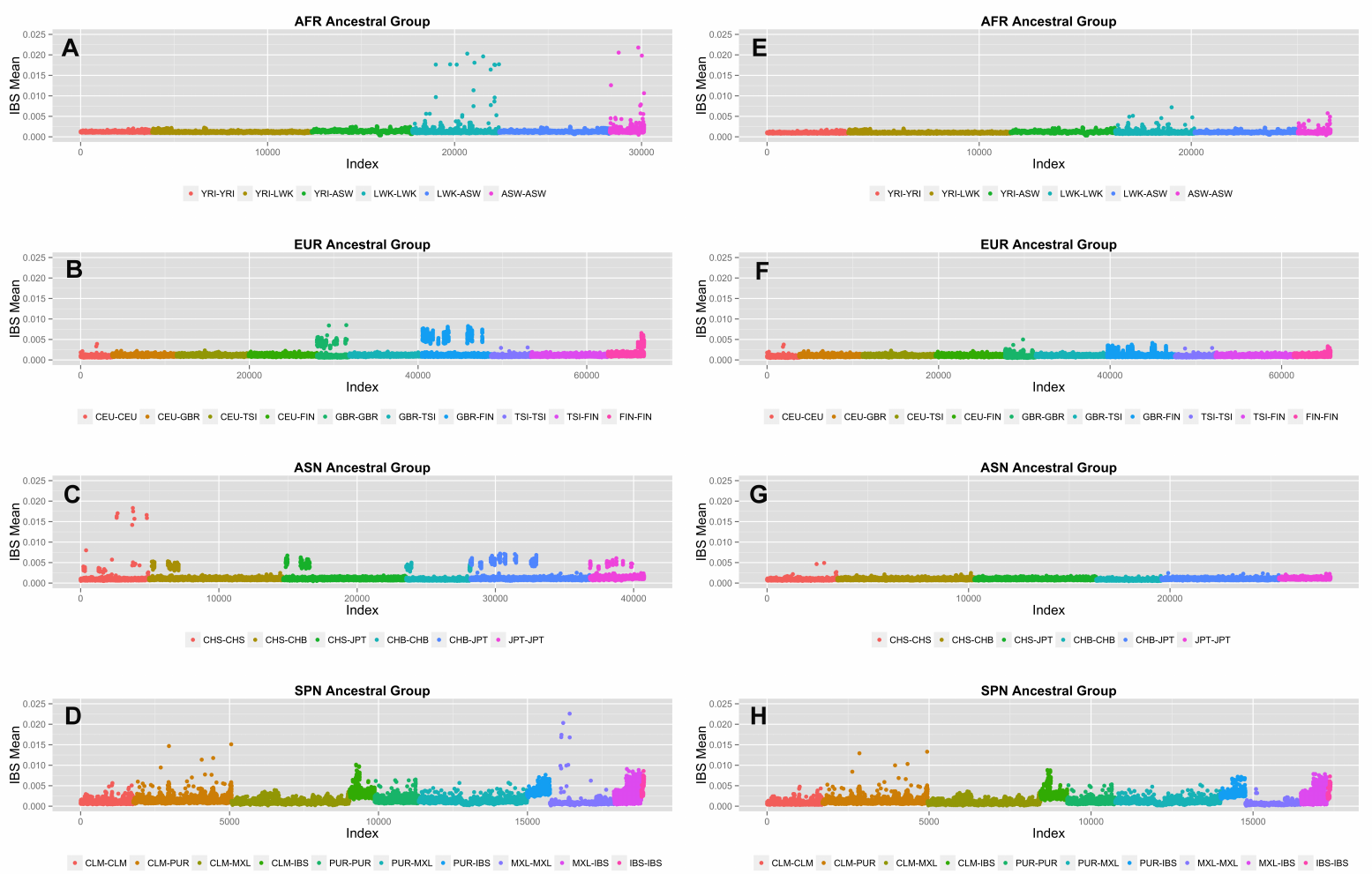
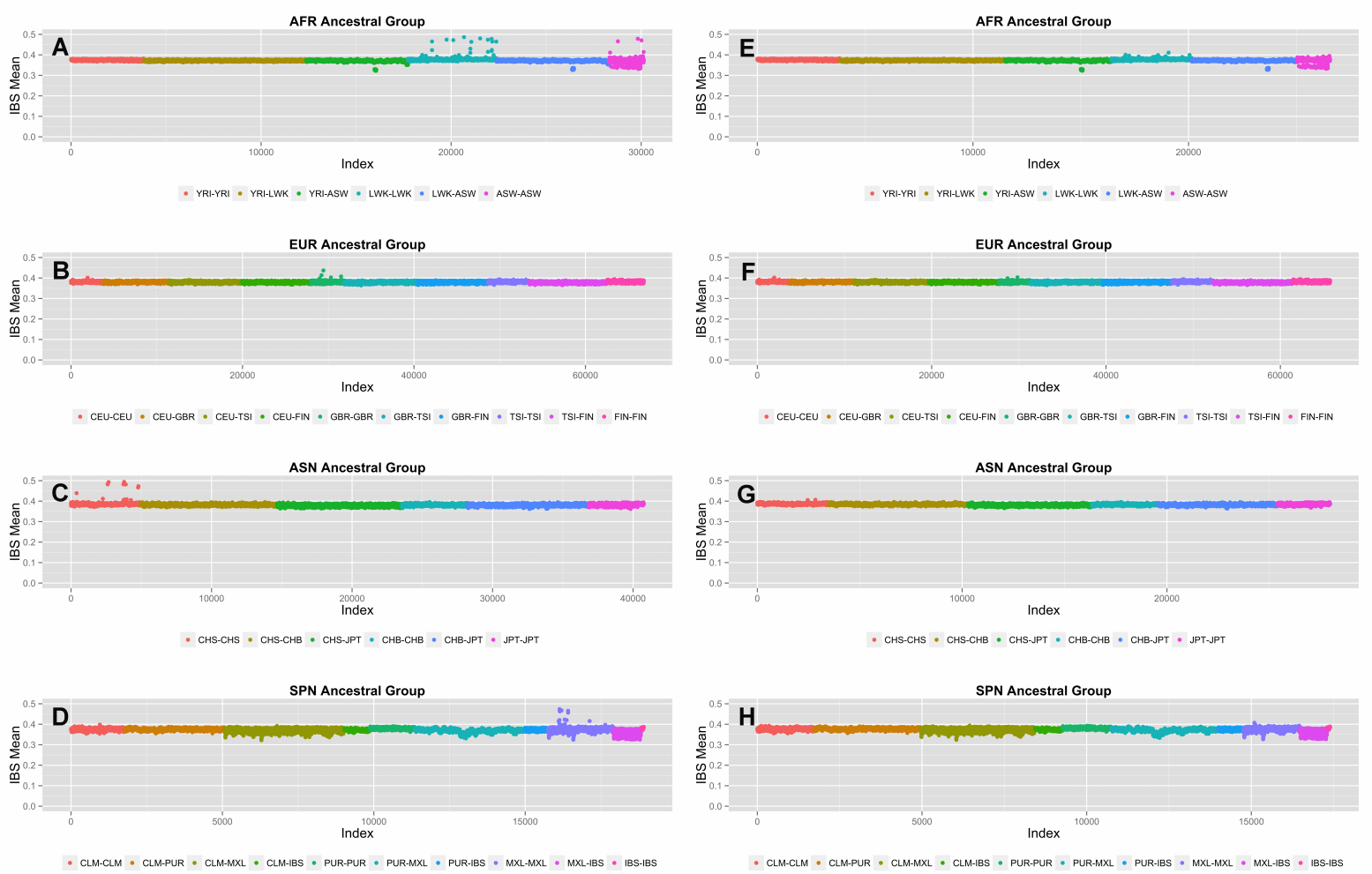


Figure 19. Pairwise IBS calculations for common variants (MAF $> 25\%$) within continental groups (A-D). Plots (A-D) show the IBS calculations within continental groups for all 1094 individuals. The plots to the right (E-H) show the IBS calculations within continental groups for 1019 individuals (cryptically related individuals removed). Each dot represents a pair of individuals; the colors correspond to population comparisons. Points with the higher mean IBS indicate increased sharing.



other available collapsing methods. Knowledge of biological features, such as genes and pathways, are available through LOKI for binning. The minimum bin size was set to two variants, the interregion bin size was chosen to be 50kb, and a MAF binning threshold of 0.03 was implemented. A 3% MAF binning threshold was chosen to focus the analysis on rare and near rare variation that differs between population groups. Genes (introns, exons, nonsynonymous variants, and predicted deleterious variants), intergenic regions, pathways, pathway-exons, regulatory regions, evolutionary conserved regions, and regions thought to be under natural selection were binned.

Results are shown in Figure 20, Figure 21, Figure 22, Figure 23, Figure 24, and Figure 25. Each matrix plot indicates the proportion of significant bins (after Bonferroni correction) out of the total number of bins generated between two populations. The color intensity represents the proportion of total bins that were significant [0, 1]. Overall, there are large differences across populations with regard to low frequency variant burden, and the distribution of low frequency variants is not random across the genome. The magnitude of stratification corresponds to the mutational landscape of the region. Note: although Iberian (IBS) populations are often clustered with European groups, and it makes sense to do so, a hierarchical clustering algorithm grouped the small population with Spanish/Mexican populations in the matrix plot results.

Coding and noncoding regions

NCBI Entrez database was chosen to provide the boundaries for gene regions and created a custom role file of intron and exon boundaries using data provided from UCSC Genome Browser [71]. In Figure 20, the top matrix corresponds to bins created using gene exon boundaries, the middle matrix corresponds to bins created using gene intron boundaries, and the bottom matrix corresponds to bins created using regions between genes (intergenic). The abbreviations for the each population are found on the x and y-axes. The numbers in each block and the color intensity [0, 1] indicate the proportion of significant bins (after Bonferroni correction) for the 1000 Genomes populations on each axis, where the darker the color, the higher the proportion of significant bins. In general, the x-axis is organized with African descent populations on the far right and increasing differentiation with regard

to low frequency burden towards the left (i.e. populations of Asian descent have the highest proportion of significant bins compared to African descent groups).

The coding regions show a trend of a lower proportion of significant bins with low frequency variant burden differences than either the intron or intergenic bins. For example, in the CEU (Northern/Western European Ancestry, USA), YRI (Yoruba African) comparison, approximately 44% of the gene exon bins had significant differences in low frequency variant burden. In contrast, the noncoding region bins, gene-introns and intergenic bins had 66% and 70% of bins with significant differences in low frequency variant burden. The coding regions appear to be under more constraint across populations than noncoding regions. Comparing only the noncoding regions, introns tend to have slightly fewer variation differences than intergenic bins, most likely because introns are by default nearest neighbors to the selective pressures on coding regions.

The gene exon bins were filtered using annotations from the Variant Effect Predictor Software (VEP) [75]. Gene bins were created with only nonsynonymous variants and a second analysis using only predicted damaging variants annotated by SIFT or PolyPhen-2 [72, 74, 75]. The results in Figure 21 indicate that these potentially functional and significant changes are even more conserved between populations than coding regions (Figure 20A).

ORegAnno annotated regions

The database ORegAnno (Open Regulatory Annotation database) was used to define regulatory region boundaries for the bin analysis. The top matrix of Figure 22 shows the 91 population comparisons for the ORegAnno regulatory feature analysis.

In comparison to Figure 20, the annotated regulatory regions have fewer significant bins. For example, in gene exon analysis shown in Figure 20, approximately 44% of the ASW-CHB gene-exon bins contained significant differences in low frequency burden. However, in Figure 22, only 28% of the ASW-CHB annotated regulatory bins contained significant differences in low frequency burden. This trend is consistent across the matrix of population comparisons; regulatory regions have fewer significant bins than the coding or noncoding features of the same population comparison.

Figure 20. Proportion of significantly different bins in A) gene exon, B) gene intron, and C) intergenic regions. The proportion of significant bins across all population comparisons increases from coding (A) to noncoding (B) and finally intergenic (C) regions.

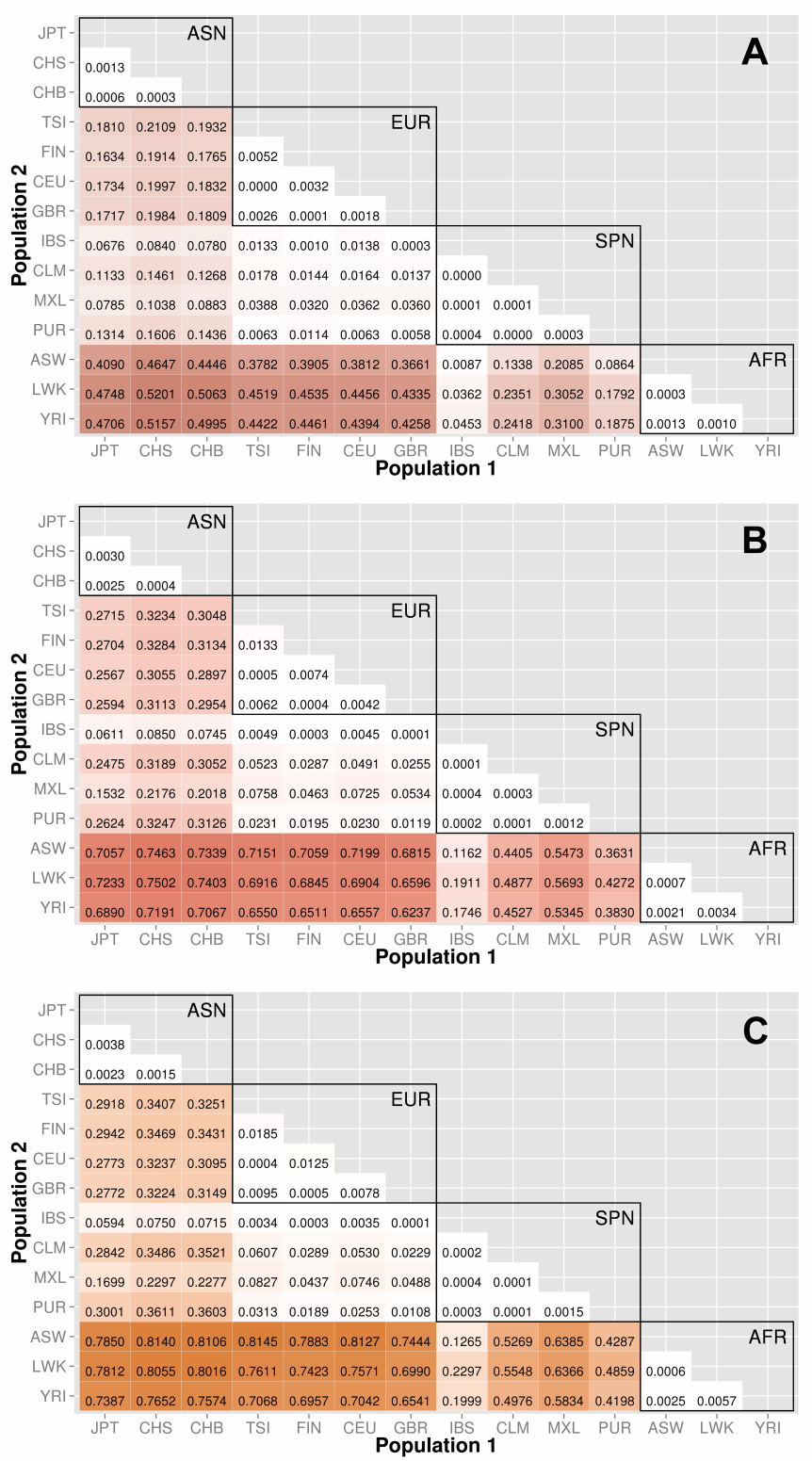
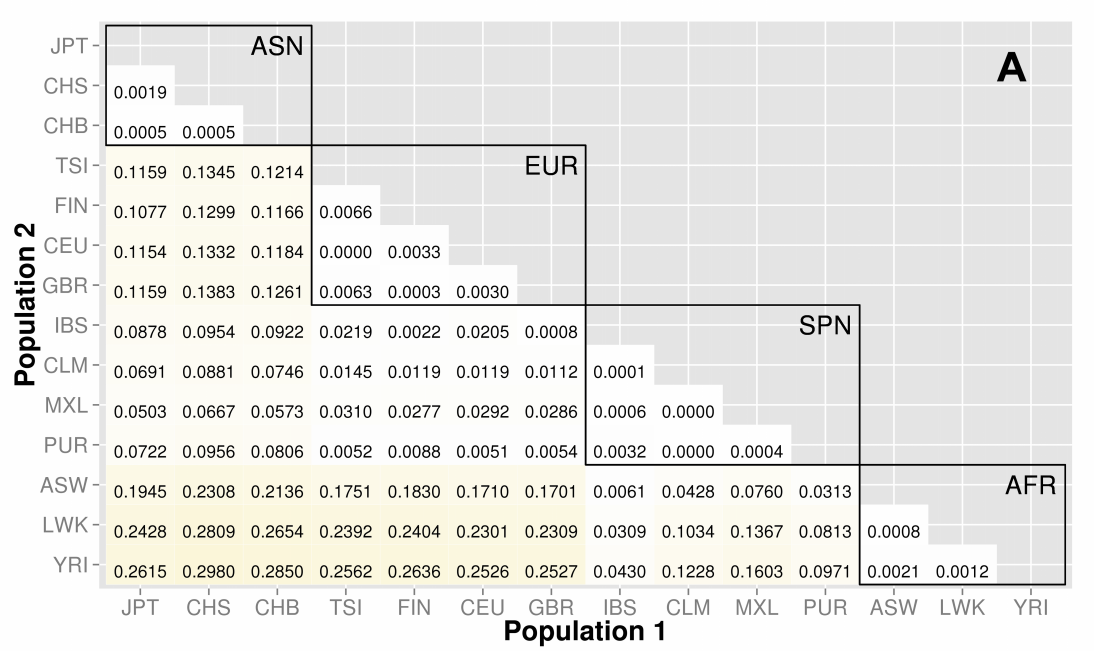


Figure 21. Proportion of significantly different bins for gene exon filters: A) nonsynonymous and B) predicted deleterious variants. The abbreviations for the each population on are on the x and y-axes. The numbers in each block and the color intensity [0, 1] indicate the proportion of significant bins (after Bonferroni correction) for the 1000 Genomes populations on each axis, where the darker the color, the higher the proportion of significant bins. In general, the x-axis is organized with African descent populations on the far right and increasing differentiation with regard to low frequency burden towards the left (i.e. populations of Asian descent have the highest proportion of significant bins compared to African descent groups). Filtering gene exon regions by mutation type and predicted functional significance lead to smaller bins and overall greatly reduced proportions of significance.



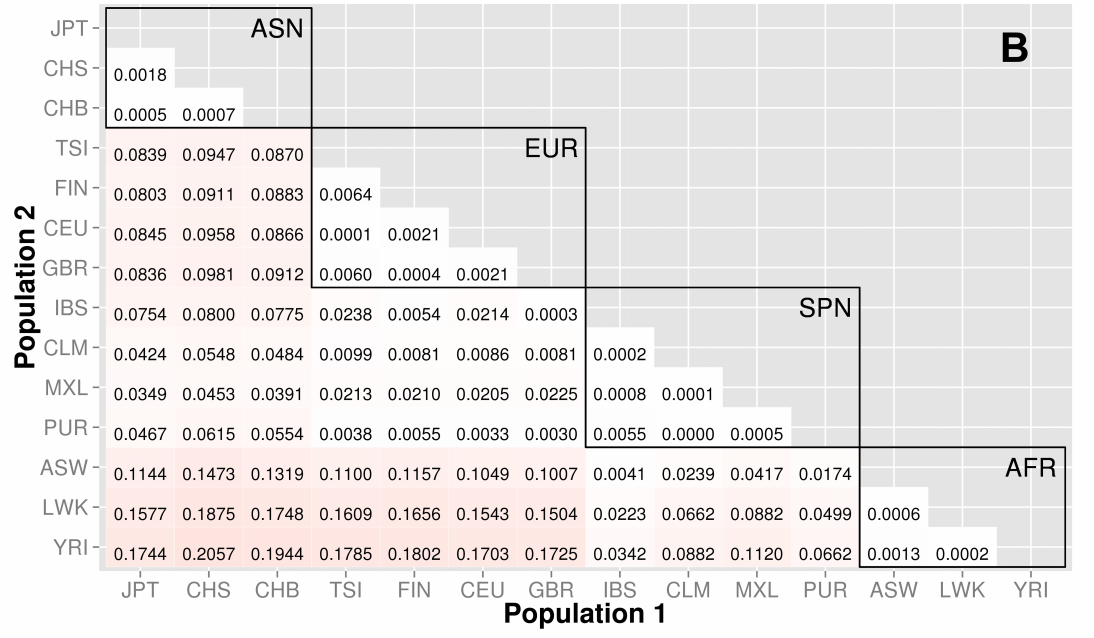
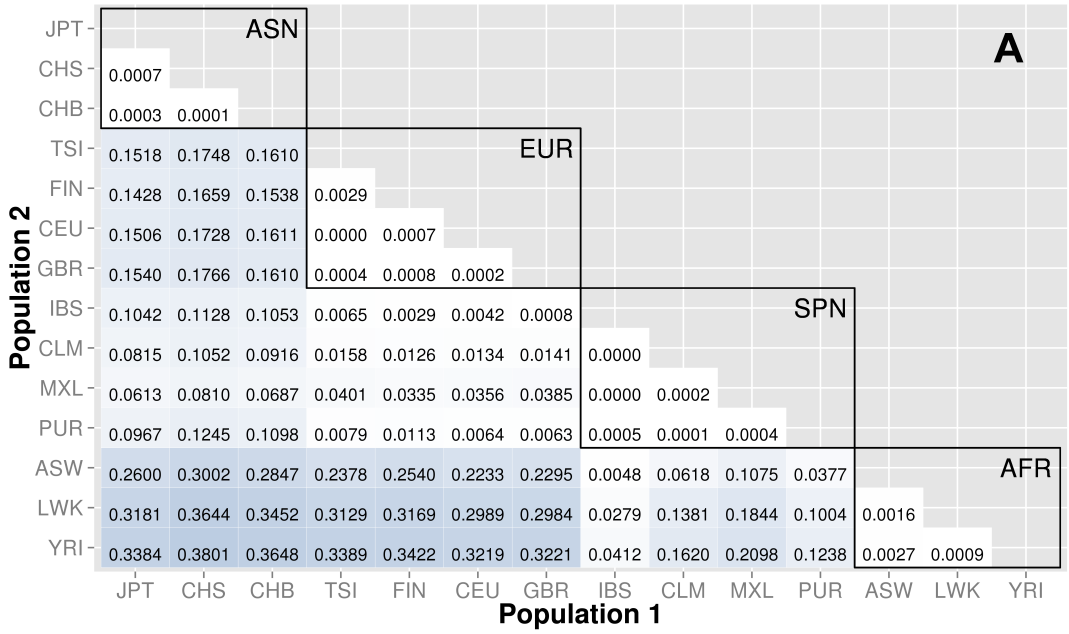
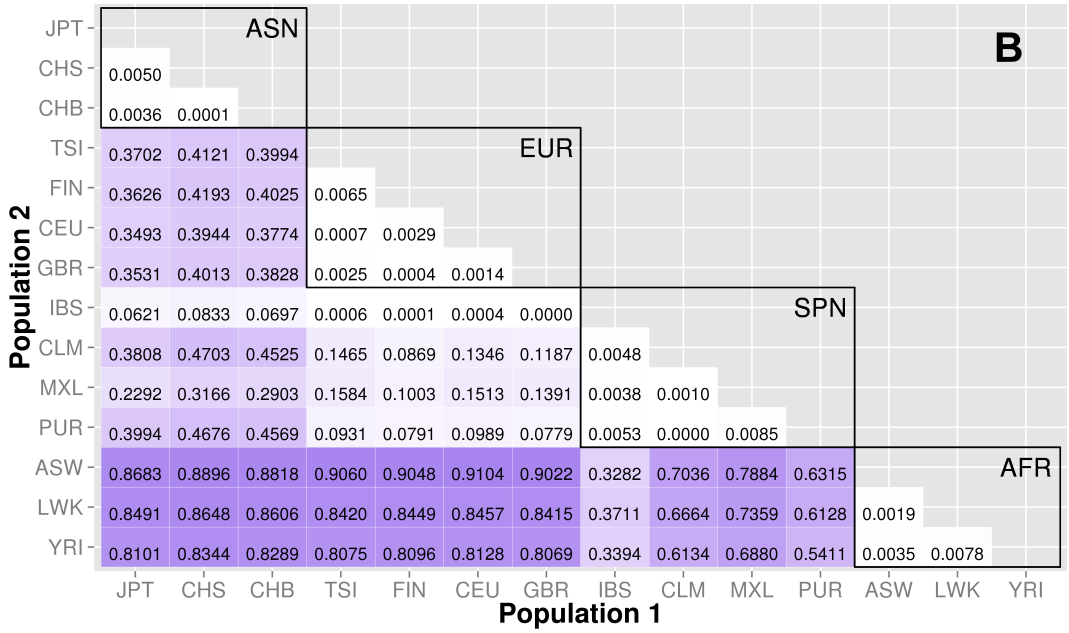


Figure 22. Proportion of significantly different bins in A) ORegAnno regulatory and B) pathway feature analysis. The abbreviations for the each population on are on the x and y-axes. The numbers in each block and the color intensity [0, 1] indicate the proportion of significant bins (after Bonferroni correction) for the 1000 Genomes populations on each axis, where the darker the color, the higher the proportion of significant bins. In general, the x-axis is organized with African descent populations on the far right and increasing differentiation with regard to low frequency burden towards the left (i.e. populations of Asian descent have the highest proportion of significant bins compared to African descent groups). From more conserved regulatory regions to relatively large binned pathways, Figure 22A shows conservation in comparison to genic regions (Figure 20) and Figure 22B shows occasionally very high proportions of significant bins in parent pathway bins in comparison to genic regions (Figure 20).





Pathway and group features

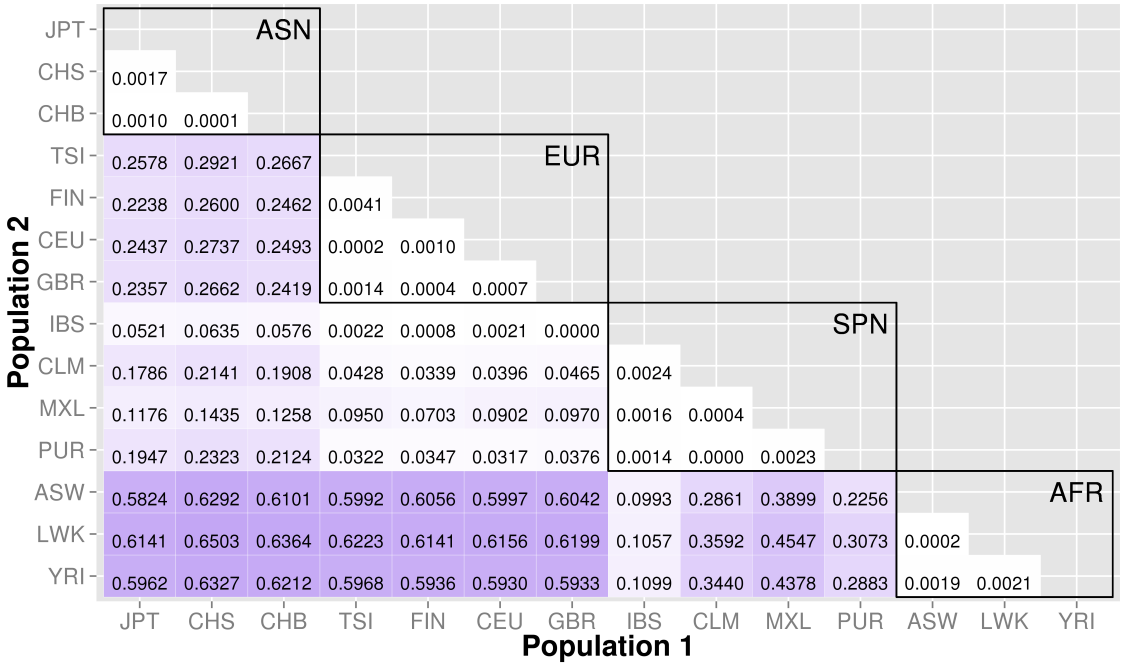
Several biological pathway and group sources from LOKI (the Library of Knowledge Integration, described in detail in Chapter III) were used to generate low frequency variant bins: PFAM, KEGG, NetPath, PharmGKB, MINT, GO, dbSNP, Entrez, and Reactome. The Figure 22B shows the 91 population comparisons for the pathway group feature analysis.

Of all of the feature analyses, pathway bins consistently show the highest proportion of significant differences in low frequency variant burden between populations. There are several potential explanations. First, since pathway bins are generally much larger than the other feature types, it is possible that large bins increase the false positive rate. Second, the same genes and regions can recur in multiple pathways. If the region has significant differences in low frequency variant burden, then each pathway or group containing that region will have a higher chance of having significant differences in low frequency variant burden. Following this logic, a pathway containing many genes has a higher chance of having at least one gene with extreme low frequency variant stratification. To compare only coding regions within a pathway, the pathway analysis was filtered to include only variants within exons. The proportions are reduced (shown in Figure 23) but still higher than the gene exon proportions shown in Figure 20A.

Evolutionary conserved regions (ECRs)

PhastCons output downloaded from UCSC Genome Browser was used to derive evolutionary conserved feature boundaries for primates, mammals, and more than 40 species of vertebrates. Figure 24 shows the 91 population comparisons for the ECR feature analysis. The numbers in each block and the color density indicate the proportion of significant bins for the 1000 Genomes populations on each axis. For example, in the ECR: vertebrate matrix, 16.38% of the ECR bins have significant differences in low frequency burden between YRI and CHS populations. In general, the x-axis is organized with African descent populations on the far right and increasing differentiation with regard to low frequency burden towards the left (i.e. populations of Asian descent have the highest proportion of significant bins compared to African descent groups). Of all of the feature analyses, ECR bins had the

Figure 23. Proportion of significantly different bins for the pathway-exon feature analysis. The numbers in each block and the color intensity [0, 1] indicate the proportion of significant bins for the 1000 Genomes populations on each axis. In general, the x-axis is organized with African descent populations on the far right and increasing differentiation with regard to low frequency burden towards the left (i.e. populations of Asian descent have the highest proportion of significant bins compared to African descent groups). The overall proportion of significant bins is much less in this pathway-exon analysis than the pathway analysis shown in Figure 22B.



smallest proportion of significant bins. More ancestrally similar populations tended to have negligibly low frequency burden differences in these conserved segments. For example, approximately 7% of the ECR region bins (vertebrate alignment) were significantly different between FIN (Finnish) and JPT (Japanese) individuals. However, the significant number of bins between the two ancestrally similar GBR (British) and CEU individuals was less than 1%.

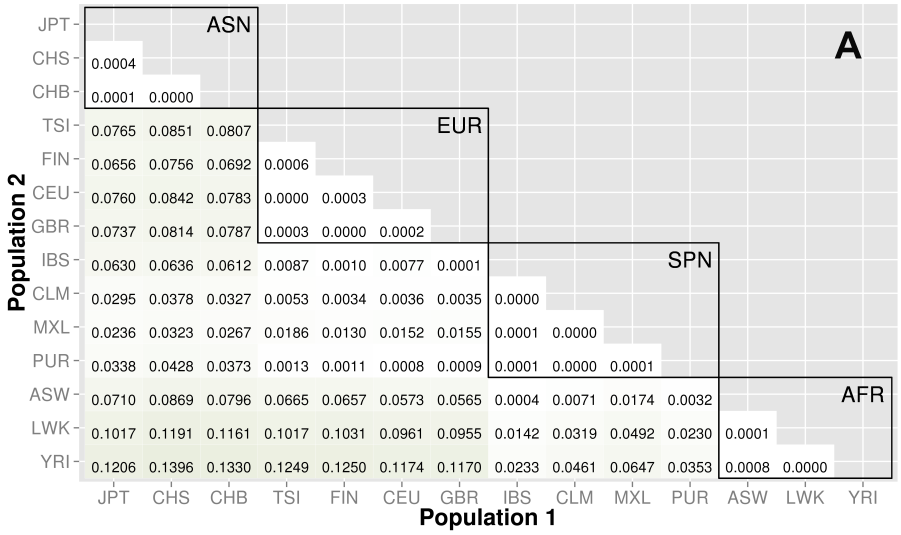
Regions of natural selection

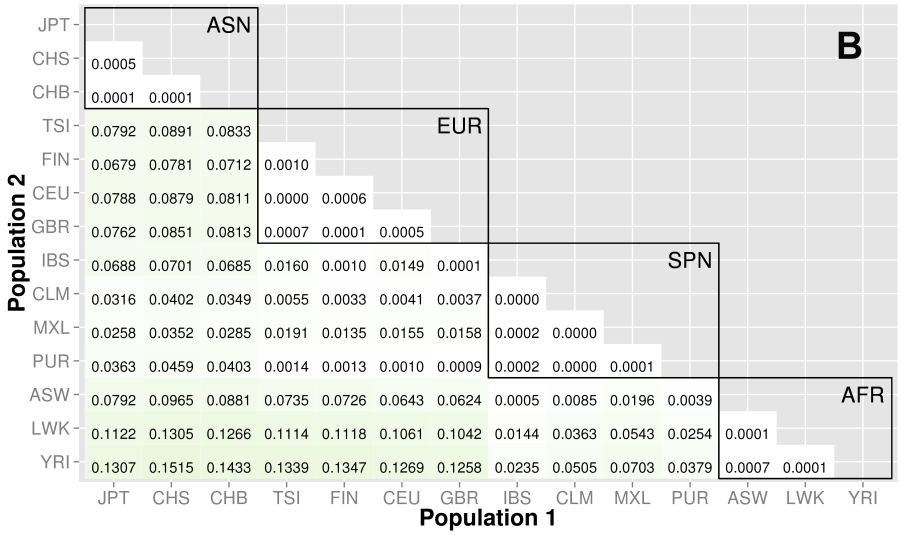
Natural selection can alter genomic variation in features, particularly in regions with some impact on protein function (regulatory regions, coding regions). Positive selection on a specific variant allows the advantageous variant to sweep through a population, which can lead to an excess of common variants. Alternatively, weak negative selection or purifying selection can result in selective removal of deleterious alleles. This can decrease variation around the locus under selection and lead to an excess of rare or low frequency variation [93]. Commonly, evidence of natural selection is found only in one ancestral group, which is consistent with the idea that these selection events postdate population separation [94]. Because of this differentiation among populations, regions identified as being under selective pressures were used as features in a BioBin analysis. Table 16 shows the analysis plan, features tested, sources used, and the mean number of bins generated across all pairwise comparisons.

To investigate regions of natural selection, a feature list was created using regions recently identified/confirmed by Grossman et al. with the Composite Multiple Signals algorithm on the 1000 Genomes Project data [84]. In addition, a publication by Barreiro et al. provided a list of specific genes with the strongest signatures of positive selection; i.e. genes that contained at least one nonsynonymous or 5' UTR mutation with an FST value greater than 0.65 [85].

After lifting positions to human genome build 37, there were only 368 remaining regions from the original regions identified by Grossman et al. The results are shown in Figure 25. The top plot corresponds to regions identified in African ancestry, the middle plot corresponds to regions identified in populations of Asian ancestry, and the bottom

Figure 24. Proportion of significantly different bins in evolutionary conserved region feature analysis A) conserved with primates, B) conserved with mammals, and C) conserved with vertebrates.





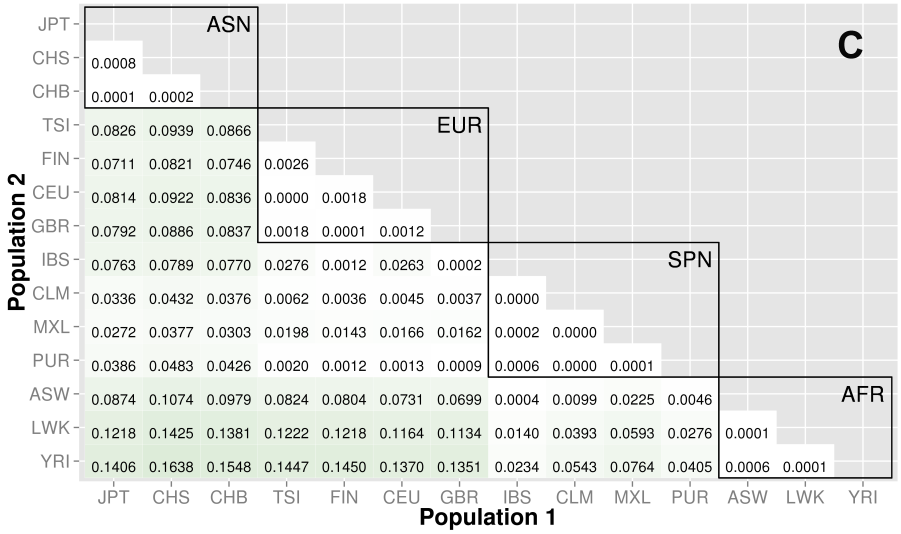


Table 16. Analyses performed for each population comparison: features tested, contributing sources, and total of bins generated for each analysis.

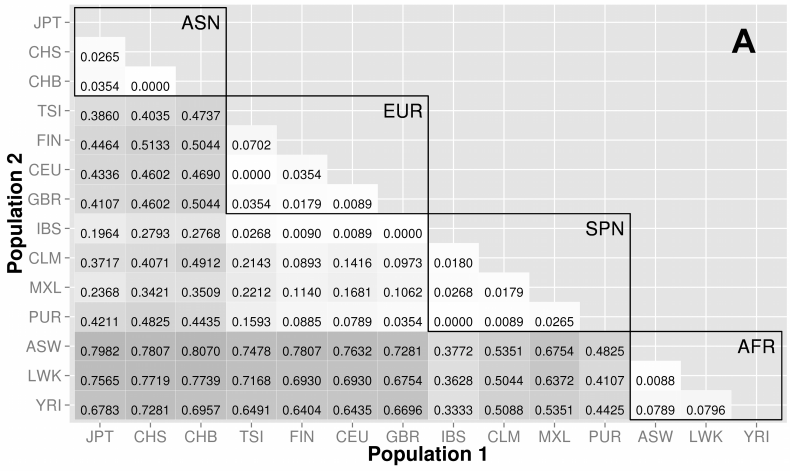
Analysis	Feature	Source	AVG Bin Total
A	Genes-Exons(NS/DEL)	NCBI Entrez, UCSC roles	80786
	Genes-Introns	NCBI Entrez, UCSC roles	
	Genes-Unknown	NCBI Entrez, UCSC roles	
	Intergenic (50kb)	-	
B	Pathway/Groups	PFAM, KEGG, NetPath, PharmGKB, MINT, GO, dbSNP, Entrez, Reactome	178497
C	Natural Selection	Pritchard/Stoneking	?
		Grossman	368
D	ORegAnno	UCSC-ORegAnno	11293
E	ECR-vertebrates	UCSC-PhaseCons	319269
	ECR-placental mammals		
	ECR-primates		

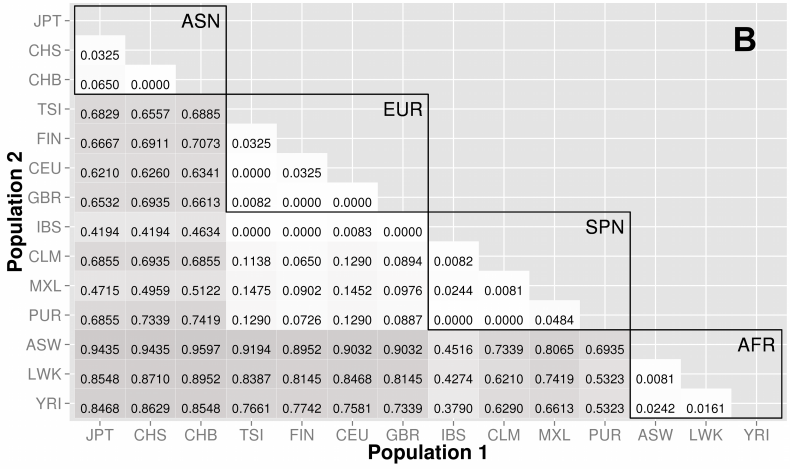
plot corresponds to regions identified specifically in populations of European ancestry. The trends in these three matrix plots are distinctly different from the trends shown in Figures 20, Figure 21, and Figure 22. The blocks of comparisons within a continental group (shown in black boxes on each matrix plot) still have very little color, indicating that the low frequency variant burden between populations within a continental group is very similar. The main difference is the gain of intensity outside of the continental groups. For example, in Figure 25B (regions identified in Asian populations), the European continental group and Spanish continental group are most likely to have proportions over 60% when compared to populations of Asian descent. In the same plot, the populations in the African group have proportions over 85% when compared to populations in the Asian group.

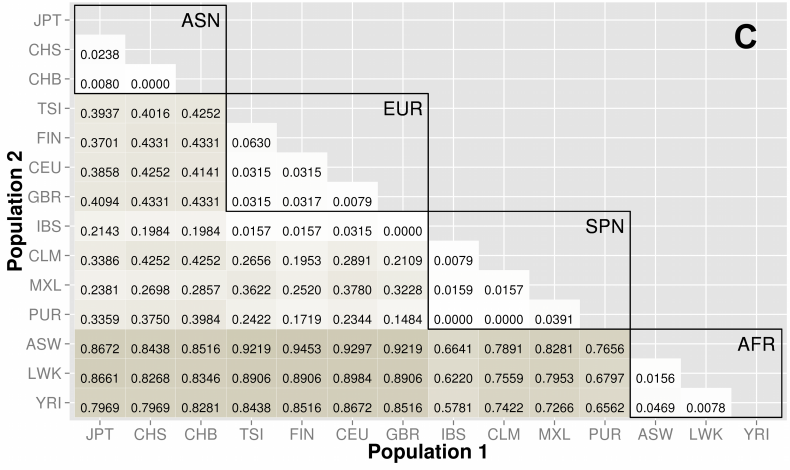
In general, regions considered to be under natural selection were unlikely to have significant differences in low frequency burden between ancestrally similar populations, and very likely to have significant differences in regions considered to be under natural selection between ancestrally distant populations (see Figure 25).

Regions of natural selection have been identified using various methods often in a population specific manner. Therefore, large differences were expected in low frequency variant burden between populations that have not shared similar evolutionary history. Specific genes provided by Barreiro et al. that have been found to show the strongest signatures of

Figure 25. Proportion of significantly different bins in natural selection analysis by region of identification: A) AFR continental group, B) ASN continental group, and C) EUR continental group. The regions of natural selection, particularly negative selection, are often accompanied by excess low frequency variants. As world populations evolved, selective forces were often unique and location specific. Therefore, the evolution of low frequency variants compared across world populations can be markers of past selective events. Populations within a continental group are very similar and there are high proportions of statistically significant bins between populations of different continental groups.







positive selection were investigated. These genes contain at least one nonsynonymous or 5' UTR mutation with an F_{ST} value greater than 0.65 [85].

Using CEU/CHB/YRI populations as representative populations from the European, Asian, and African ancestral groups, the regions of natural selection associated with the gene list provided by Barreiro with overlap in one of two other publications are shown in Table 17 and Table 18 [84, 95, 96]. This table includes the number of loci in the bin, total binned variants from both populations, and the bin p-value. The source author corresponds to the paper for that particular region. The "relevant population" describes the population where the signature of selection was found.

Next, particular genes known to have allele frequency differences between populations were of interest, including Lactase, Phenylalanine Hydroxylase, CTCF, and CFTR. Table 19 shows the BioBin p-value results for three representative populations (YRI, CEU, CHB) in each gene bin. For example, the Lactase gene had a significant low frequency variant burden difference between CEU/CHB and CEU/YRI but not YRI/CHB. On the other hand, CFTR was significantly different between CEU/CHB, CEU/YRI, and YRI/CHB.

Linkage disequilibrium in binned low frequency variants

Although low frequency variants are commonly assumed as independent (in low linkage disequilibrium (LD) with other variants), there are rare haplotypes within related individuals and populations [97]. Linkage disequilibrium (LD) was investigated in 10 top-ranked bins for three population comparisons, CEU-CHB, CHB-YRI, CEU-YRI. LD was calculated between binned variants and the number of variants inside of a bin in LD with an $r^2 > 0.3$ was calculated.

In Figure 26, three pairwise population comparisons are shown. The top 10 ranked bins were investigated from the CEU-CHB (A), CHB-YRI (B), and CEU-YRI (C) coding and noncoding analyses for presence of LD ($r^2 > 0.3$) between two variants in the same bin. The abundance of white space illustrates the lack of rare haplotypes in the top most significant bins.

Table 17. Reviewing regions of interest found in Barreiro study. Genes identified in Barreiro study between CEU/CHB and CEU/YRI population comparisons with an F_{ST} value > 0.65 that were also found in the regions identified by Pritchard, Stoneking, and Grossman.

Natural Selection					**POP 2**							
					CHB				**YRI**			
POP 1	**Source Author**	**Relevant POP**	**Genes**	**Gene Info**	**# Loci**	**POP 1 Variants**	**POP 2 Variants**	**P-value**	**# Loci**	**POP 1 Variants**	**POP 2 Variants**	**P-value**
CEU	M.S.	EUR	LCT	Lactase	2258	8164	17181	$7.255e^{-07}$	4133	10618	34844	$1.314e^{-26}$
	J.P.	EUR	LCT		75	158	1088	1	185	602	1129	1
	S.G.	EUR	LCT		697	1413	7863	$1.505e^{-03}$	1207	3064	11430	$4.447e^{-22}$
	J.P.	AFR	ABCC11	Morpho. Traits	277	3478	150	$1.096e^{-17}$	303	432	1420	$6.173e^{-10}$
	J.P.	ASN	EDAR		360	1524	1184	$2.710e^{-02}$	574	1226	4264	$1.828e^{-19}$
	S.G.	ASN	EDAR		134	4462	1216	$2.432e^{-26}$	168	140	766	$1.388e^{-14}$
	M.S.	EUR	SLC24A5		987	1826	6736	$7.382e^{-18}$	1604	2670	10047	$3.686e^{-24}$
	J.P.	ASN	SLC24A2		1042	5522	3838	1	1647	2777	10374	$1.753e^{-24}$
	J.P.	EUR	DUOX2	Immunological Response	44	115	75	1	130	86	1075	$2.319e^{-26}$
	J.P.	EUR	ALMS1	Insulin Regulation	780	8624	688	$5.052e^{-08}$	1116	7977	2043	$2.001e^{-24}$
	S.G.	EUR	ALMS1		544	5734	568	$5.346e^{-04}$	803	1502	5947	$5.060e^{-23}$
	J.P.	ASN	ADH1B	Metabolic Regulation	41	223	190	1	77	167	455	$6.642e^{-05}$
	J.P.	EUR	CPSF3L	Misc. Unknown	46	135	241	$8.589e^{-01}$	92	145	688	$1.274e^{-16}$
	J.P.	EUR	FAIM		64	1250	462	$9.653e^{-05}$	127	249	934	$2.902e^{-24}$
	M.S.	ASN	LIMCH1		1584	7267	3567	$3.297e^{-06}$	1083	2067	6123	$1.059e^{-20}$
	J.P.	EUR	PCGF1		7	8	55	$2.905e^{-06}$	14	8	55	$3.756e^{-04}$
	J.P.	AFR	RNF135		83	183	75	$7.000e^{-05}$	157	284	4311	$2.142e^{-26}$
	J.P.	EUR/ASN	SLC30A9		323	2483	525	$4.786e^{-11}$	541	1464	3847	$1.991e^{-20}$
	M.S.	ASN	SLC30A9		735	9261	1204	$1.547e^{-10}$	1083	4509	6435	$5.799e^{-02}$
	S.G.	AFR	SLC30A9		112	2191	253	$4.786e^{-11}$	131	883	458	$9.199e^{-09}$
	J.P.	EUR	TTC31		29	63	25	1	5	28	303	$2.072e^{-18}$

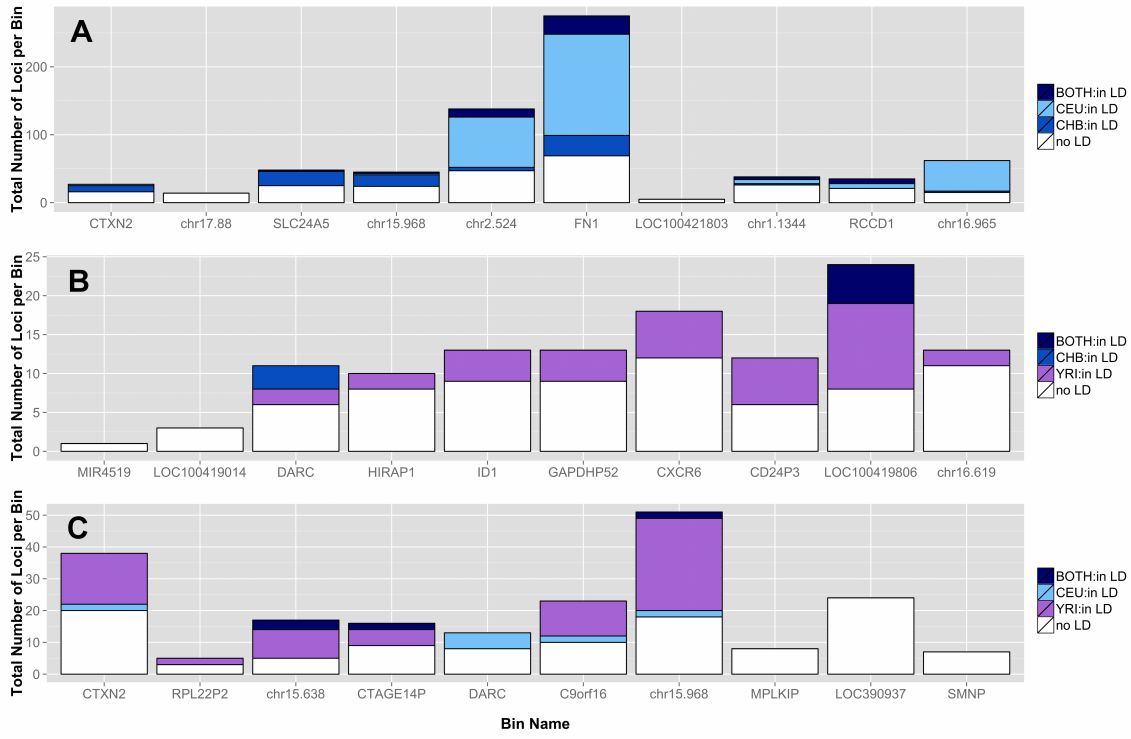
Table 18. Reviewing regions of interest found in Barreiro study. Genes identified in Barreiro study in YRI-CHB comparison with an F_{ST} value > 0.65 that overlapped with the regions identified by Pritchard, Stoneking, and Grossman.

Natural Selection					**POP 2**			
					CHB			
POP 1	**Source Author**	**Relevant POP**	**Genes**	**Gene Info**	**# Loci**	**POP 1 Variants**	**POP 2 Variants**	**P-value**
YRI	M.S.	EUR	LCT	Lactase	4179	26603	14402	$7.397e^{-12}$
	J.P.	EUR	LCT		215	1123	1293	1
	S.G.	EUR	LCT		1247	7633	6310	1
	J.P.	AFR	ABCC11	Morpho. Traits	420	8166	306	$3.249e^{-18}$
	J.P.	ASN	EDAR		601	3828	1016	$6.360e^{-18}$
	S.G.	ASN	EDAR		174	2916	124	$7.385e^{-23}$
	M.S.	EUR	SLC24A5		1483	8190	4216	$2.789e^{-11}$
	J.P.	ASN	SLC24A2		1729	18627	4085	$4.140e^{-23}$
	J.P.	EUR	DUOX2	Immunological Response	125	1226	52	$3.869e^{-26}$
	J.P.	EUR	ALMS1	Insulin Regulation	1239	28301	1565	$1.252e^{-25}$
	S.G.	EUR	ALMS1		922	19783	1311	$1.248e^{-25}$
	J.P.	ASN	ADH1B	Metabolic Regulation	62	213	22	$1.469e^{-14}$
	J.P.	EUR	CPSF3L	Misc. Unknown	77	622	223	$7.590e^{-08}$
	J.P.	EUR	FAIM		130	1993	396	$1.784e^{-24}$
	M.S.	ASN	LIMCH1		1095	11207	1086	$4.804e^{-25}$
	J.P.	EUR	PCGF1		15	55	55	1
	J.P.	AFR	RNF135		146	4112	172	$3.003e^{-25}$
	J.P.	EUR/ASN	SLC30A9		568	6538	1023	$2.897e^{-24}$
	M.S.	ASN	SLC30A9		1301	27895	7627	$1.388e^{-25}$
	S.G.	AFR	SLC30A9		194	6665	1947	$8.165e^{-23}$
	J.P.	EUR	TTC31		55	483	25	$1.853e^{-17}$

Table 19. Gene results for specific genes of interest with known allele frequency differences between ancestral populations.

POP 2		**CHB**				**YRI**			
POP 1		**Number of Loci**	**POP 1 Variants**	**POP 2 Variants**	**P-value**	**Number of Loci**	**POP 1 Variants**	**POP 2 Variants**	**P-value**
CEU	LCT	140	114	1030	4.24E-10	213	143	1799	1.03E-22
	PAH	246	1520	168	2.01E-07	492	1670	2737	1
	CTCF	225	836	207	8.83E-05	323	801	2412	4.13E-04
	CFTR	487	3340	428	4.54E-11	731	1737	4355	1.66E-16
YRI	LCT	214	1424	788	5.88E-01				
	PAH	530	4764	2892	2.10E-05				
	CTCF	311	3700	429	2.08E-21				
	CFTR	792	7396	1544	5.00E-23				

Figure 26. Proportion of loci in top bins in high LD with other variants in the same bin. Each bar represents a gene or intergenic bin. For a particular population comparison, the total height of the bar corresponds to the number of loci in that bin. The shades of blue and purple correspond to loci with r^2 LD values greater than 0.3 for a specific population shown in the legend. The variant can be in LD in one population, the other population, or both (described in each legend). Almost all of the low frequency loci in LD had r^2 values of approximately 0.5 or 1, corresponding to almost perfect LD. The white space corresponds to loci in the bin with LD values less than 0.3. The top bins are therefore, mostly composed of independent loci.



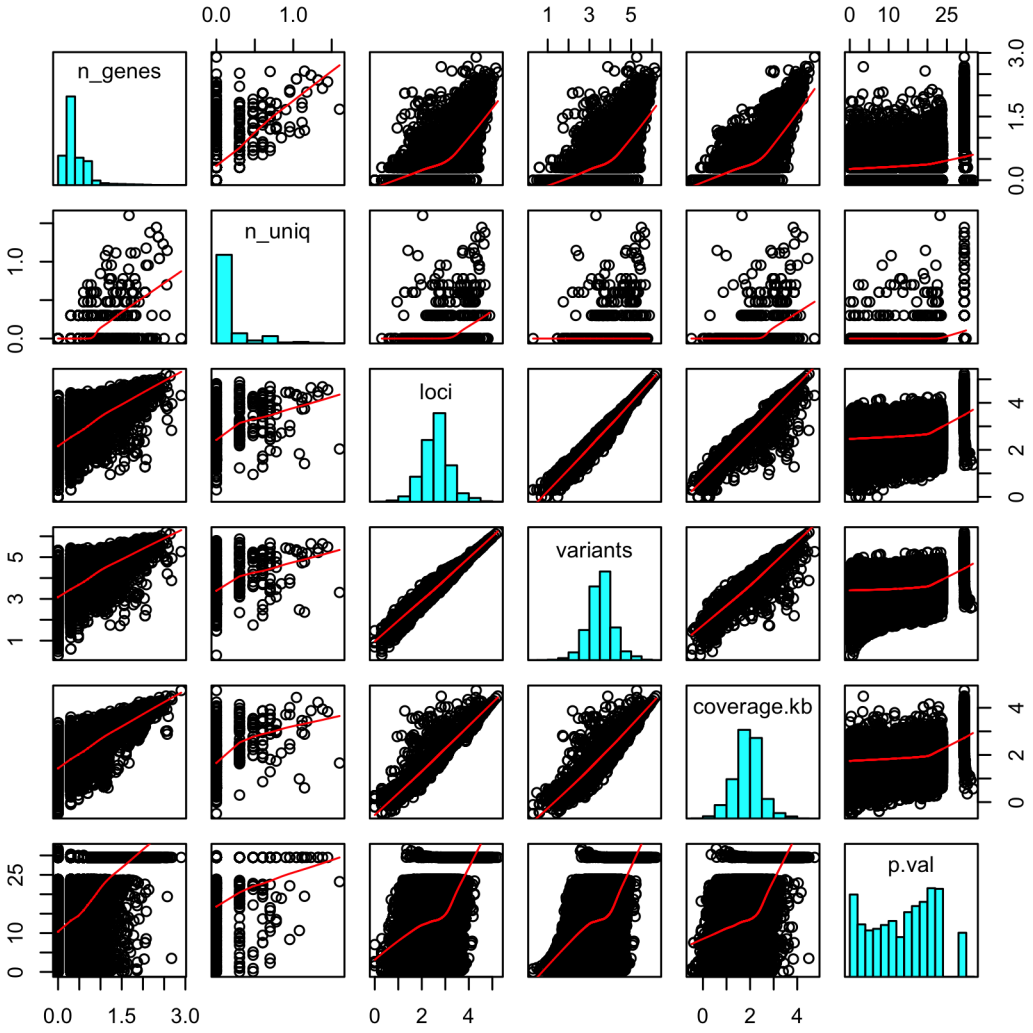
Pathway correlation with bin size

Since the proportion of significant bins in the feature analyses is considerably higher for pathway bins than any other feature, the correlation between pathway p-value and bin size was of particular interest. The pathway feature for the CEU-YRI population comparison was chosen to assess the correlation between significance and several characteristics of the pathways. All of the pathways in the YRI/CEU analysis were compiled, and the following information for each pathway bin was obtained: total genomic coverage, number of genes, number of independent genes, number of loci, number of variants, and BioBin p-value. Because the majority of pathways or groups were not very large, the data was heavily skewed (see Figure 27). A $\log_{10}$ transformation was performed on all six variables: number of genes in the pathway or group, number of unique genes (not present in any other pathway or group), number of loci in the pathway bin, number of variants in the pathway bin, genomic coverage of the pathway bin, and the BioBin reported Bonferroni adjusted p-value. Because of the skewedness, any pathway bins that had transformed loci values outside of 2.5 standard deviations of the log-transformed loci mean were removed.

Figure 27 and Figure 28 show the correlations between six untransformed and transformed variables (with outliers removed), where each pairwise correlation is significant (p-value < 0.05). A bin was considered an outlier if the number of loci in the bin was more than 2.5 standard deviations from the mean transformed loci value.

The number of loci, number of variants, and size of genomic region were significantly and linearly correlated with each other (correlation coefficients > 0.95). The most interesting correlations were the nonlinear correlations between the loci/variants/genomic coverage and p-values. Figure 28B is a higher magnification of the highlighted correlation in Figure 28A; specifically, the correlation between $\log_{10}$ p-values and $\log_{10}$ variants were plotted. The lowess smoothing function is shown in red, and the function appears to change slope twice. On the x-axis, the slope from x=1 to x=3 is relatively linear and the $\log_{10}$ p-value increases with an increasing number of variants (p-value becomes more significant). From x=3 to x=4, the slope is near 0. From x=4 to x=5, the slope appears nonlinear and with a larger slope than the left slope, indicating again most significant p-values with higher numbers

Figure 27. Investigation of pathway significant correlation with bin size using untransformed pathway variables. Correlation scatterplot matrix for six untransformed variables: the number of genes in a pathway (n_genes), the number of unique genes in the pathway (n_uniq), the number of loci in the pathway bin (loci), the number of variants in the pathway bin (variants), the genomic coverage of pathway (coverage_kb), and the bin p-value (p-val). Bins considered outliers were removed before generating the correlations (http://stat.ethz.ch/R-manual/R-patched/library/graphics/html/pairs.html). The variables are right skewed and require transformation.



of variants in a bin. Although these are transformed values, the p-values are not perfectly uniform. Therefore, the tails are possibly unreliable.

Lastly, boxplots describing certain characteristics from each data source were created. Figure 29 shows that specific sources (i.e. KEGG) consistently have larger bin characteristics (number of loci, number of genes, coverage (kb), etc.) and also have much more significant bin p-values (Figure 29B). It appears that certain sources might inherently have more significant groups by nature of the information that these sources provide, or because of the size of groups found in the source.

Figure 28. Investigation of pathway significant correlation with bin size using $\log_{10}$ transformed pathway variables. Correlation scatterplot matrix for six $\log_{10}$ transformed variables: the number of genes in a pathway (n_genes), the number of unique genes in the pathway (n_uniq), the number of loci in the pathway bin (loci), the number of variants in the pathway bin (variants), the genomic coverage of pathway (coverage_kb), and the bin p-value (p-val). Bins considered outliers were removed before generating the correlations. Figure 28B is a higher magnification of the correlation highlighted in Figure 28A, but instead of the $+\log_{10}$ transform of p-values, it is showing the the $\log_{10}$ transformed p-values and $\log_{10}$ transformed variants with a loess smoothing function (red line) and 95% confidence intervals (gray shading). (http://stat.ethz.ch/R-manual/R-patched/library/graphics/html/pairs.html).

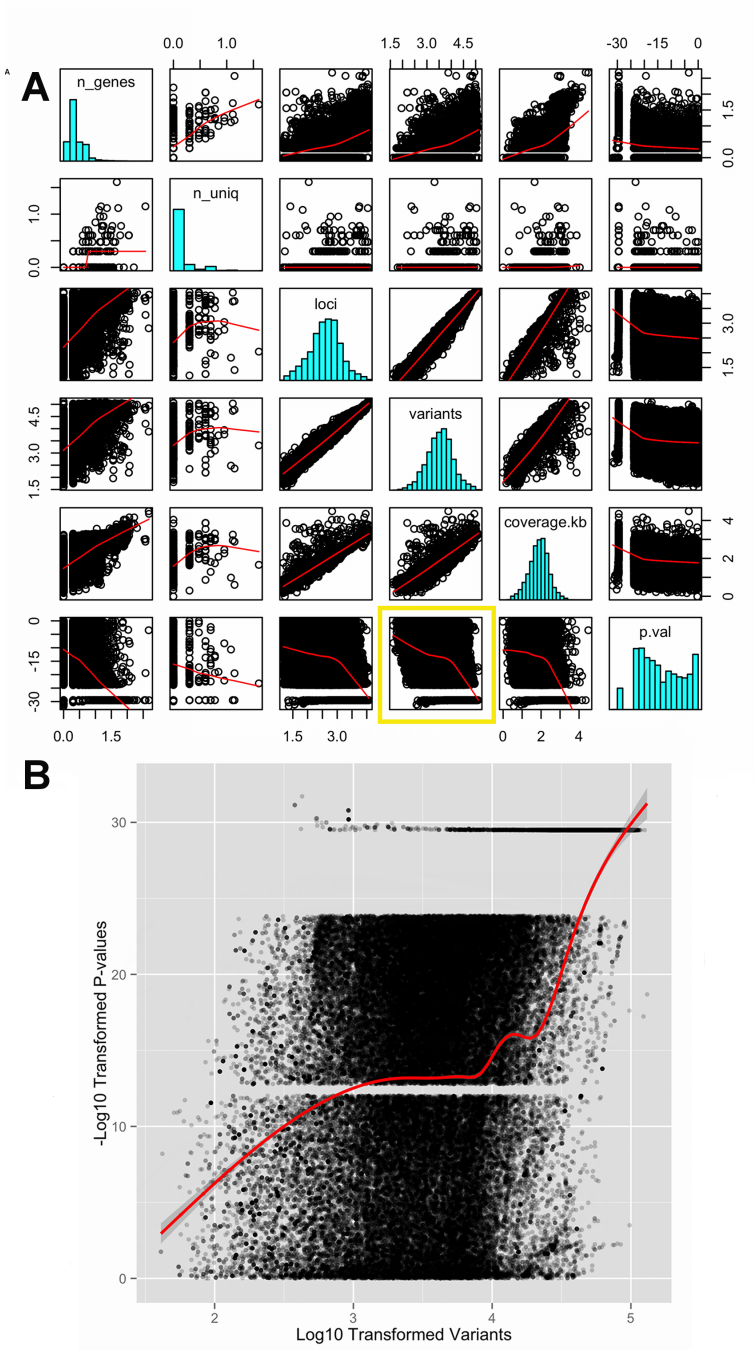
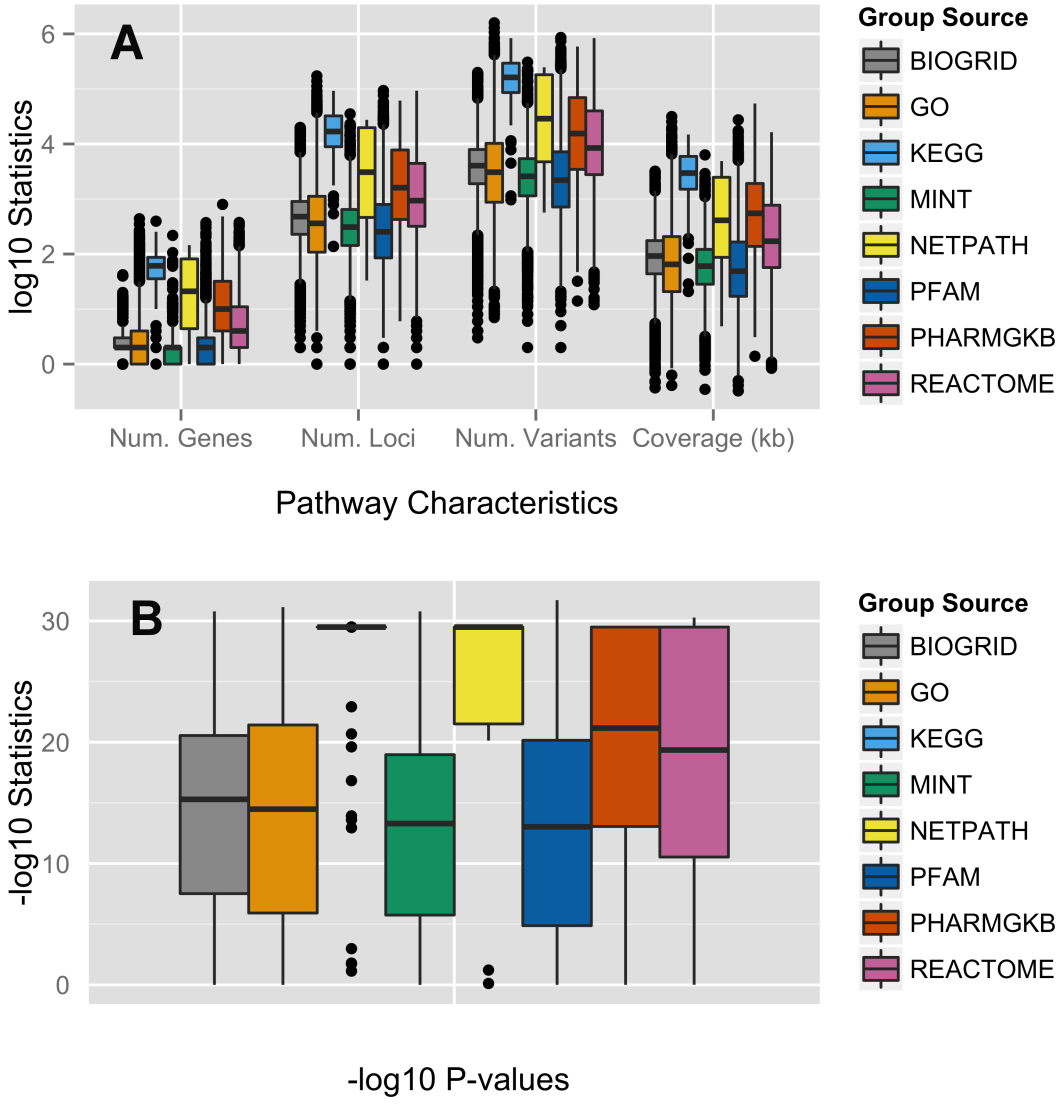


Figure 29. Pathway characteristics presented by LOKI source. Different pathway characteristics presented in box plots: A) The y-axis shows the $\log_{10}$ frequency of each source statistic for the number of genes (Num. Genes), the number of loci (Num. Loci), the number of variants (Num. Variants), and the coverage in kb, B) The distribution of p-values for the various knowledge sources. On average, the same four sources listed above also tend to have bins with smaller p-values. Each boxplot and color corresponds to the biological knowledge sources listed in the legend. KEGG, NetPATH, PharmGKB, and Reactome show consistently larger bins (higher number of loci, variants, and coverage).



Discussion

1000 Genomes Project data

Since the reference genome is predominantly of European ancestry [98, 99, 100], populations with non-European ancestry generally have more variation with respect to the reference genome than populations of European ancestry (see Table 14). Therefore, to interpret the results of this study, one might conclude that non-European populations have higher rates of sequencing error than European descent populations. However, in the most recent 1000 Genomes Project publication, the authors report an accuracy of individual genotype calls at heterozygous sites more than 99% for common SNPs and 95% for SNPs at a frequency $\geq 0.5\%$ [60]. Furthermore, the authors found that variation in genotype accuracy was more related to sequencing depth and technical issues than population-level characteristics [60]. Therefore, neither the sequencing error nor the predominantly European reference genome adequately explain the trends seen in the genomic feature exploration.

Both sequence generation (technology and/or site) and population identity strongly contribute to underlying stratification in next-generation sequence data. After removing individuals with cryptic relatedness, 4 out of 14 Phase I populations were sequenced entirely using a single sequence technology (CHB, CHS, JPT, and TSI). The other 10 populations had between 3-18 individuals or 5%-57% of the population sequenced on technologies other than Illumina (ABI SOLID or LS454). Note: all three of the Asian populations (after removing individuals with cryptic relatedness) were sequenced only with Illumina technologies.

Investigation of allele sharing

To identify cryptically related individuals, LD-pruned common variants (minor allele frequency $\geq 5\%$, linkage disequilibrium $r^2 \leq 0.2$) were used to calculate identity-by-descent.

Seventy-five individuals of various population backgrounds were identified and eliminated. In addition to the previously documented relatedness in 1000 Genomes Project [`http://www.1000genomes.org/phase1-analysis-results-directory`], additional cryptic relatedness was found [90, 101]. The differences are likely because continental groups were used (not a single population or the entire 1094 individuals) to identify cryptically related individuals; in this analysis, continental groups could include variants with fixed opposite frequencies that are overall common. This is infrequent in populations of the same continental group, but could be stratification introduced by different sequencing technologies.

Genomic feature exploration

The major goal of this study was to investigate population stratification across multiple biological features. Matrix plots were created to illustrate the proportion of significant bins in each comparison (shown in Figure 20, Figure 21, Figure 22, Figure 23, Figure 24, and Figure 25). These results show an interesting trend between functional regions of the genome and variant tolerance, where variant tolerance refers to the balance of mutation load and resulting functional impact. Mutations appear to be less tolerated in functional regions. Similarly, ECRs, which are known to be conserved among species, are also the features least likely to have variation burden differences between two populations. There is some debate about selection and functional significance in these conserved regions; it is unknown what factors have the largest effect on mutation rates [102], but it is possible that consistently low mutation rates in these features have generated conserved regions throughout evolution [103]. There are two potential explanations: 1) additional level of repair of DNA damage in transcriptional active regions by transcription coupled repair (TCR), 2) approximately 3% of the genome is subject to negative selection; however, it is estimated that functionally dense regions contain up to 20% of the sites under selection [102, 103].

A number of the top results in each comparison have an interesting context, particularly in light of natural selection. Perhaps one of the most notable is *SLC24A5* (Ensembl ID:ENSG00000188467), which is one of the top ten results in 19 out of 91 populations

comparisons in the gene feature analysis. European specific selective sweeps estimated in the last 20,000 years suggest that *SLC24A5* is key in skin pigmentation; Zebrafish with "golden" mutations in this gene exhibit melanosomal changes [104, 105, 106]. The presence of selection in this gene only in particular populations is most likely due to environmental factors such as distance to the equator, which has led to the evolution and expansion of low frequency variants in some populations but not others.

A second notable top result is *DARC* (Ensembl ID:ENSG00000213088), which encodes the Duffy antigen. The *DARC* gene bin was in the top ten results in 14 out of 91 population comparisons in the gene feature analysis. It has long been known that populations of African descent have increased diversity due to natural selection at this location, which prevents *Plasmodium vivax* infection.

The top result from the regulatory region analysis was a region on chromosome 20 (chr20:45395536- 45396346) which was in the top ten bins in 24 out of 91 population comparisons in the ORegAnno feature analysis. This region also overlaps several ENCODE transcription factor binding sites in multiple cell lines: CTCF, POLR2A, NFYA, E2F1, FOS, and more. It was also annotated as an insulator in multiple cell lines in ENCODE Chromatin State Segmentation analyses using Hidden Markov Models [71, 107].

As one last example, the intergenic bin chr15.968 contains variants in the genome location chr15:48400199-48412256. This bin is one of the top ten bins in 17 out of 91 population comparisons in the intergenic analysis. The region covered by the chr15.968 bin is less than 1kb upstream of *SCL24A5* on chromosome 15 and overlaps with several transcription factor-binding sites (including CTCF), regions thought to be weak enhancers, and regions thought to be insulators. According to Grossman et al., there are defined regions under natural selection before and after this region (chr15:45145764-45258860 and chr15:48539026-48633153), and all are very likely to participate in the transcriptional regulation of *SLC24A5*.

The natural selection features require knowledge of three things for interpretation: 1) population A, 2) population B, and 3) the population where this signature was identified. When all three of these are within the same ancestral or continental group, very few differences are expected in low frequency burden. However, if population A is the same or similar to the population possessing the selection signature and population B is different, signifi-

cant differences are expected in low frequency burden between population A and population B. In these results, the vast majority of regions considered to be under natural selection had significant differences in low frequency burden between disparate ancestral populations, which supports the theory of selection in these regions.

Proportion of LD between variants in a bin

Low frequency variants can form rare haplotypes, which inflate the signal in a feature bin [43, 97]. The top 10 ranked bins from the CEU-CHB, CHB-YRI, and CEU-YRI coding and noncoding analyses for presence of LD between two variants in the same bin were investigated. Figure 26 shows bins predominately filled with white-space indicating low to no pairwise LD between variants in those bins. In the top ten bins from these three analyses, rare haplotypes do not appear to be driving the significant differences seen in low frequency variant burden.

Pathway size correlations

In general, size of bins can influence the number of stratified variants contained and thus the significance of that bin. It is important to prove that this is because larger bins have a greater opportunity to collect variants that are stratified and not because of inflated type I error. Type I error rates in bins between approximately 40 variants to over 100,000 variants were tested and no correlation between bin size and Type I error rate was found (see Chapter IV). However, it should also be noted that while larger bins have more chances to collect stratified variants, there is also a larger capacity to collect neutral variants that contribute noise and decrease the signal.

Using CEU-YRI pathway burden analysis, correlation between pathway size and significance was reviewed. The number of genes in pathways ranged from 1 to over 700 genes, with the average around 5 genes per group. Correlations for this data are shown in Figure 28. Not surprisingly, there were very linear and positive correlations between number of loci, number of variants, and genomic coverage. However, each of these had a nonlinear and somewhat complex relationship with the log-transformed p-value. This is highlighted in Figure 28B, which shows the relationship between the $\log_{10}$ transformed p-value and

the $\log_{10}$ transformed number of variants in the bin. The trend indicates that p-values are positively correlated (become more significant) with numbers of variants in a bin when the numbers of variants are relatively small or very large.

Two reasons could explain this correlation: 1) the false-positive rate is influenced by bin size (number of variants per bin), and 2) true signals from child bins (genes) with burden differences which perpetuate higher numbers of significant parent bins (pathways). After extensive simulation testing (see Chapter IV) and recent publications in the literature, the latter is likely true [108]. A single or small number of child bins (gene bins in this example) can drive parent bins (pathways in this example) to be significant even if no other child bin contains stratification. The comparison in Figure 29 between group sources available in LOKI suggests KEGG, NetPATH, PharmGKB, and Reactome have consistently larger bins (higher number of loci, variants, and coverage). On average, these same four sources also tend to have bins with smaller p-values. Therefore, larger pathways are more likely to contain a gene with extreme low frequency variant stratification.

Trends in the Asian continental group

The x-axis of each matrix plot (i.e. Figure 20) is oriented with African continental populations on the far right and the continental group with the highest proportion of significantly different low frequency variant bins on the far left. Asian populations are generally more different from African populations than European populations. There are at least three possible explanations; first, the Asian populations were the only continental group to be sequenced on the same technology, which could introduce a different bias when testing any of these populations with populations outside of Asian ancestry. While this is true of the 1,019 unrelated individuals, there were cryptically related individuals sequenced using SOLID technologies in all three of the Asian populations. The only population (including cryptically related individuals) to be sequenced exclusively on Illumina was TSI. When the Asian populations were examined, the cryptically related individuals were added back into the analysis to see if individuals of Asian descent sequenced with different technologies changed the results. The trend was the same, Asian populations are the most different from African populations with regard to low frequency variant burden. The second potential explanation

is that Asian populations had considerable proportions of cryptic relatedness that had to be removed for this analysis, 49 of the 75 individuals removed were from Asian populations. Perhaps there was something unique about how those samples were collected. The third and most interesting explanation involves the journey for early Asian populations leaving Africa. Travelling east was much different geographically than travelling west. For example, early Asian migrants would have traversed the Himalayan Mountains. The harsh travel could have induced bottlenecks and other evolutionary mechanisms that would uniquely change the genetic architecture, specifically the architecture of low frequency variation.

Conclusion

As researchers continue in pursuit of genetic etiologies explaining heritability in common, complex disease, it is important to consider multiple types of genomic data, specifically variation beyond common variants. Low frequency variants are more frequent in the genome than common variants and are likely to have significant functional impact on human health. Many successes are expected in next-generation data analysis; however, it has become increasingly clear that the same methods and corrections used in GWAS cannot be applied to low frequency variant analyses. Since low frequency variants are often recent mutations, they are specific to continental ancestry groups. This provides two important conclusions: first, low-frequency variants that influence disease are likely not the same across distantly related individuals (allelic and locus heterogeneity); second, low frequency population substructure leads to substantial differentiation and cannot be ignored in low frequency analyses [60].

Population stratification is a pertinent and very present confounder in genomic studies. In common variant association studies, stratification is often managed using ancestry correction components. Until relatively recently, the challenges presented by low frequency population stratification to genomic analyses has been overlooked. Current methods used for GWAS to correct for ancestry are not likely adequate for low frequency stratification [76, 109]. Therefore, it is imperative that researchers are aware of potential pitfalls

stratification can introduce to low frequency genomic analyses. Additionally, this study highlights potential limitation in using the 1000 Genomes samples as population based controls in case-control association studies. If the case population belongs to a different ancestry than the 1000 Genomes control population and/or they are exclusively sequenced using a different technology, this can introduce significant stratification. This level of stratification may or may not be adequately corrected using PCA or other analysis methods. Thus, it is clear that proper evaluation of stratification would be prudent if 1000 Genomes Project data are being proposed as a population control set.

In summary, the results presented in this chapter expose the magnitude of low frequency population stratification between all populations available in 1000 Genomes Project Phase I release across multiple interesting biological features. The magnitude of low frequency stratification appears to be dependent on the functional location of the variation. For example, there were fewer differences in low frequency burden in coding regions than intergenic regions. Features with less variant tolerance and possibly more evolutionary constraint had fewer differences in low frequency variant burden between different populations, i.e. significant low frequency bins seemed to be consistent with mutation theory. African descent populations overall varied most greatly from populations of Asian descent. However, low levels of stratification existed even between populations of the same continental group. Future studies should focus on methods to accurately control for low frequency population stratification.

CHAPTER VI

BIOBIN ANALYSES IN NATURAL DATA

Kabuki analysis

BioBin was applied to ten samples with Kabuki syndrome. Kabuki syndrome is a rare disorder that affects multiple systems; it is commonly associated with the following characteristics: cardiac anomalies, skeletal abnormalities, characteristic facial appearance, and intellectual disability [110]. In a published analysis using these data, Ng et al. applied a filtering method to identify MLL2 as a possible causative gene for Kabuki syndrome [111]. Although a sample of ten individuals across multiple ancestries do not provide reasonable power to achieve statistical significance for identified rare variant trends in a burden analysis, it was a useful exercise to show how BioBin can be used to prioritize bins based on rare variant burden differences.

Study sample

The NHLBI Kabuki dataset available on dbGaP (http://www.ncbi.nlm.nih.gov/gap/) was downloaded April 2012. According to the authors of the original study, ten unrelated individuals with Kabuki syndrome were sequenced: 7 of European ancestry, 2 of Hispanic ancestry, and one of mixed European and Afro-Caribbean ancestry [111]. Variants were identified using a custom Agilent array capture kit targeting all protein coding regions annotated by RefSeq 36.3 [111]. Shotgun fragment libraries were hybridized to these custom microarrays, and then enriched using massively parallel sequencing for an average coverage of 40x on the mappable, targeted exome [111]. The raw fastq files were downloaded from dbGaP and processed using standard exome algorithms: bwa, samtools, picard, GATK, and

bedtools.

Since the number of identified variants correlate with the depth of coverage (to a threshold of approximately 30x) and the Kabuki cases are from multiple ancestral backgrounds, it was imperative to use a multi-ethnic control set with a high depth of coverage. Publically available Complete Genomics whole-genome sequences for 54 unrelated individuals from 11 populations were used as the control group for this experiment [112]. The 54 Complete Genomics samples were sequenced with an average genome-wide coverage of 80x [112].

Methods and results

In this study of 10 cases and 54 controls, there was very little power to detect a reasonable association. If the probability of exposure (allele frequency) was 0.03, type I error was 0.05, and the true odds ratio for disease in exposed subjects relative to unexposed subjects was 3, the power to reject the null hypothesis using a chi-square test was only 0.226 [113].

BioBin was used to collapse the whole-exome data for 10 Kabuki individuals with 54 individuals from Complete Genomics whole-genome data. In the original Kabuki analysis, Ng et al. used a filtering method to identify MLL2 as a possible causative gene for Kabuki syndrome. In this analysis, in order to compare the cases and controls, we filtered both datasets by exome boundaries (available from UCSC) and filtered out variants present in the 1000 Genomes Project Phase I data (October 2011 release: 14 populations, 1094 individuals) [71, 88]. A MAF binning threshold of 0.05 was used to collapse rare variants based on known gene regions (start and stop positions form bin boundaries) and known pathways (gene bins in the same pathway are collapsed into one pathway bin). BioBin produced the MLL2 gene bin with 125 total variant loci (184 total variants) at a minor allele frequency threshold of 0.05, but was not significant (p-value = 0.4718).

While we did not replicate the *MLL2* finding, one of the top pathways included *EMG1*, a gene previously associated with Bowen-Conradi syndrome (pathway adjusted p-value $<$ 0.001, gene adjusted p-value $<$ 0.001). Bowen-Conradi syndrome has a much more severe phenotype, but shares two disease characteristics with Kabuki: impaired growth and mental

retardation [114].

Conclusions

Ng et al. filtered out 1000 Genome variants and other non-causative variants identified from previous Kabuki studies. They also considered only nonsynonymous variants with predicted changes in function. As shown in the results section, ten individuals with whole-exome data are not a large enough case sample size for sufficient power in a Wilcoxon 2-sample rank sum test. The uneven sample size and different sequencing approach for cases and controls were major limitations in this study. To compare the cases and controls, both datasets were filtered by exome boundaries and variants present in the 1000 Genomes Project Phase 1 release were removed. These steps helped reduce potentially noise contributing neutral variants.

The same filtering process from Ng et al. was not used and the sample size affords very little power for a case-control study, which together, likely explain why *MLL2* was not significant in this analysis. In addition, population stratification exists in the cases and controls and was not accounted for in this analysis. *MLL2* has 54 exons and quite a bit of neutral variation. As shown in Chapter IV, larger bins with increased background variation make causative signals harder to detect.

In this underpowered analysis, BioBin results should be utilized as a prioritization method; thus, it would require a much larger sample size to investigate the robustness of the *EMG1* association. To improve this analysis, one could potentially use a principle components analysis to adjust for the variant confounding between the two groups in a regression analysis or perform a permutation test to help adjust for unknown confounding. A better test data set would include at least 500 individuals that were sequenced with the same technology. Overall, BioBin can be used as a filtering mechanism to group data and evaluate rare variant burdens between two groups, but requires a more substantial sample size to gain power to detect significance.

Cystic fibrosis analysis

Cystic fibrosis (CF) is a genetic disease affecting multiple organ systems. According to the Cystic Fibrosis Foundation, cystic fibrosis affects approximately 1 out of every 2000-3000 live births in Caucasian individuals and the median life expectancy for patients with cystic fibrosis in 2011 was 36.8 years [115]. It is caused by homozygous or compound heterozygous mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene, a regulated chloride channel. More than 1500 catalogued mutations with disease causing potential in the *CFTR* gene have been catalogued [116]. Of the various complications and comorbidities that accompany CF, the most debilitating organ dysfunction involves abnormalities in airway secretion and resulting chronic lung infections.

Pseudomonas aerunginosa (PA) infection is the most common cause of respiratory failure in patients with CF, and is responsible for deteriorating lung function and mortality in patients with persistent infection [117]. Although there are clear environmental exposures that contribute to infection frequency, there are also genetic factors that modify recurrent PA infection risk. Recently, Green et al. calculated the heritability of chronic PA infection as 0.76 using monozygotic and dizygotic twins with Class I, II, and III mutations (severely affected CFTR function) [118]. The results of this study and the known severity of chronic infection to individuals with cystic fibrosis make a strong case for continued genomic research in this area.

Study sample

Data were accessed from the National Heart, Lung, and Blood Institute (NHLBI) from September 2012 through January 2013. Study Accession ID phs000254.v2.p1 contains exome sequence data from 431 subjects across two cohorts, 189 limited to cystic fibrosis research from the University of Washington (UW) and 242 for general research use from University of North Carolina (UNC). Selected subjects were divided based on two extremes

Table 20. Data characteristics for cystic fibrosis study sample for 416 European descent individuals.

Covariates	Overall (%) N=416	PA (%) N=181	PF (%) N=235
Data release version			
Phase I	90 (21.63)	90 (49.72)	0 (0)
Phase II	326 (78.37)	91 (50.28)	236 (100)
Sample Site			
UW	174 (41.83)	117 (64.64)	58 (24.58)
UNC	242 (58.17)	64 (35.36)	178 (75.42)

of lung disease phenotypes: youngest individuals with chronic *Pseudomonas aeruginosa* (Pa) and those exhibiting extremely poor pulmonary function (PF) as defined by survival corrected FEV percentile [119].

Before starting analyses, a principal component analysis (PCA) was run using the 1000 Genomes Project Phase I data and CF data to assess and designate ancestry [88]. Figure 30 shows the first two principal components of the merged 1000 Genomes (labeled by color in legend) and dbGaP cystic fibrosis study data (labeled gray in legend). In Figure 30, individuals of European descent from the 1000 Genomes Project data are clustered in orange. In the cystic fibrosis dataset, individuals with values for PC1 ≤ 0.005 or PC2 ≤ 0.01 were eliminated from the data set based on how closely individuals clustered with the European descent group from 1000 Genomes Project Phase I data. Using these criteria, 15 individuals from the UW site were dropped from further analyses.

Table 20 shows phase and site data characteristics of the remaining 416 individuals. Phase refers to the dbGaP release, 90 individuals used in this study were released in the first phase through dbGaP. An additional 326 individuals were released in the second phase. All of the individuals from the PF study were ascertained in the second phase. In addition more individuals overall were collected from the UNC site than the UW site.

All 416 samples were collected from individuals with cystic fibrosis, but technically from two studies: recurrent *Pseudomonas aeruginosa* (PA) infection and pulmonary function (PF) phenotype. Therefore, the analyses presented in this chapter are stratified by phenotype. Several of the covariates were measured over the clinical history for a given patient; in this case, median values were used for statistical analyses. Median values were used for the following variables: Forced Expiratory Volume (FEV) measures, age, and height. The

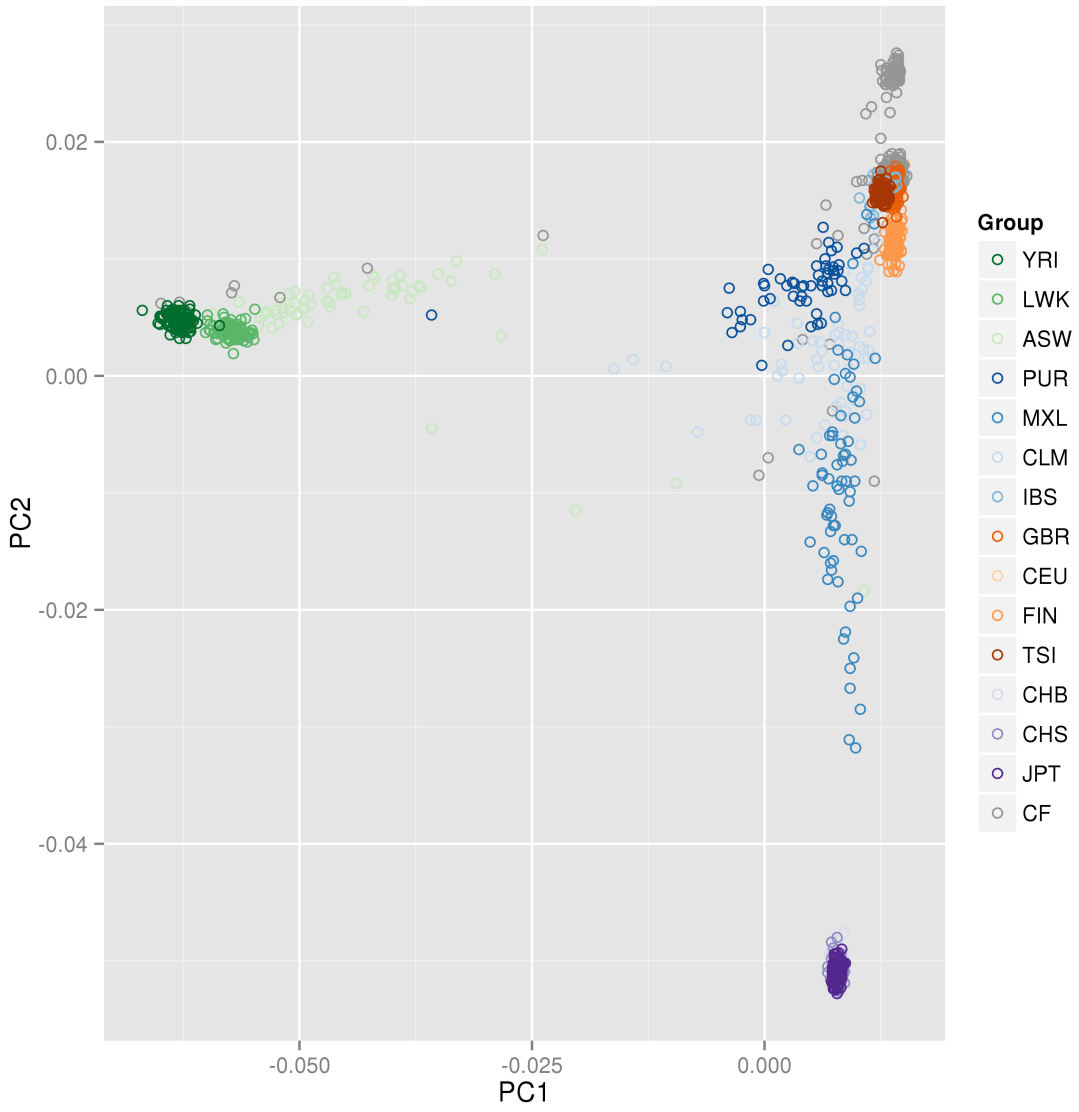


Figure 30. Principal component analysis (PCA) using merged samples from the CF analysis and 1000 Genomes Project Phase I data to identify ancestry.

Table 21. dbGaP cystic fibrosis clinical and demographic characteristics for two studies: recurrent pseudomonas infection (PA) and mild/severe pulmonary phenotype (PF).

Covariates	PA*		PF	
	Cases	Controls	Cases	Controls
Gender†				
Male	38	51	58	60
Female	49	43	56	62
Age‡				
Male	5.95 (2.89)	15.92 (7.50)	12.12 (3.45)	19.24 (6.83)
Female	5.68 (2.53)	14.28 (5.51)	13.56 (3.98)	19.52 (8.85)
FEV values*‡				
Male	1.33 (0.54)	2.15 (0.92)	1.27 (0.49)	3.71 (0.86)
Female	1.39 (0.53)	2.01 (0.59)	1.16 (0.40)	2.86 (0.57)
Height (cm)*‡				
Male	124.87 (11.73)	159.65 (19.15)	145.77 (17.75)	165.17 (14.13)
Female	127.30 (13.67)	150.77 (11.21)	148.8 (14.81)	155.17 (12.29)
Sequencer (HiSeq)†				
Male	7	6	26	30
Female	7	7	28	23
Sequencer(GAII)†				
Male	31	45	34	33
Female	42	36	33	41

† counts (N)

‡ calculated mean and standard deviation

* Values inclusive of imputed measures

clinical and demographic variables shown in Table 21 are stratified by study and gender.

To maximize the number of individuals in each statistical analysis, missing values were imputed for any covariates with missingness. FEV values and height variables were missing in 14 PA cases and are annotated with an "$*$" in Table 21. The R package "mi" was utilized to perform multiple imputations [120]. Although, these were not missing completely at random, the variables were highly correlated with age and gender, thus, this process provided a reasonable imputation. A fixed estimate for imputed values was used in the analysis.

Correlations between covariates shown in Table 20 and Table 21 were evaluated. The correlation plot shown in Figure 31 was generated using the R package "PerformanceAnalytics" to visualize the correlation between all of the study covariates, clinical covariates, demographic covariates, and calculated principal components in the PA study [121, 122]. On Figure 31, the bottom panel contains the scatterplots between pairwise variables, each

label and histogram are shown along the diagonal, and the top panel indicates the coefficient of correlation (in number and relative size) and the significance of the correlation in asterisks. As shown in Figure 31, the first principal component was highly correlated with both sequencing variables. Covariates age, FEV measurements, height, and gender were correlated. This is not surprising since lung capacity is anatomically related to individual size.

The comprehensive correlations from the PA analysis shown in Figure 31 are somewhat hard to study in detail, pertinent correlations shown in Figure 31 are shown again in Figure 32.

Methods

As healthcare providers continue to adopt electronic medical record systems and build collaborations with researchers, genomic study designs will increasingly include a multitude of genomic and clinical data. Many of these variables are highly correlated measures that can complicate analytical results. Using the CF dataset, the following approaches were taken with the CF data: traditional binning approach with logistic regression using maximum-likelihood estimators (ML) and logistic regression using penalized likelihood estimators (Firth), pre-processing variable selection approach with ML logistic and Firth logistic regression, pathway analysis using ML logistic and Firth logistic regression, and a machine learning elastic net analysis.

Statistical analyses

Firth logistic regression performs better than ML logistic regression in analyses with "separation." Separation occurs when the outcome is very rare, very common, or when there are several correlated predictors. In small data sets, separation occurs when both outcomes are not observed in a cell defined by two covariates, e.g., the outcome is perfectly predicted by a non-trivial combination of covariates. Perfect separation refers to a combination of predictors that perfectly predict the outcome. For example, zero controls (outcome) are

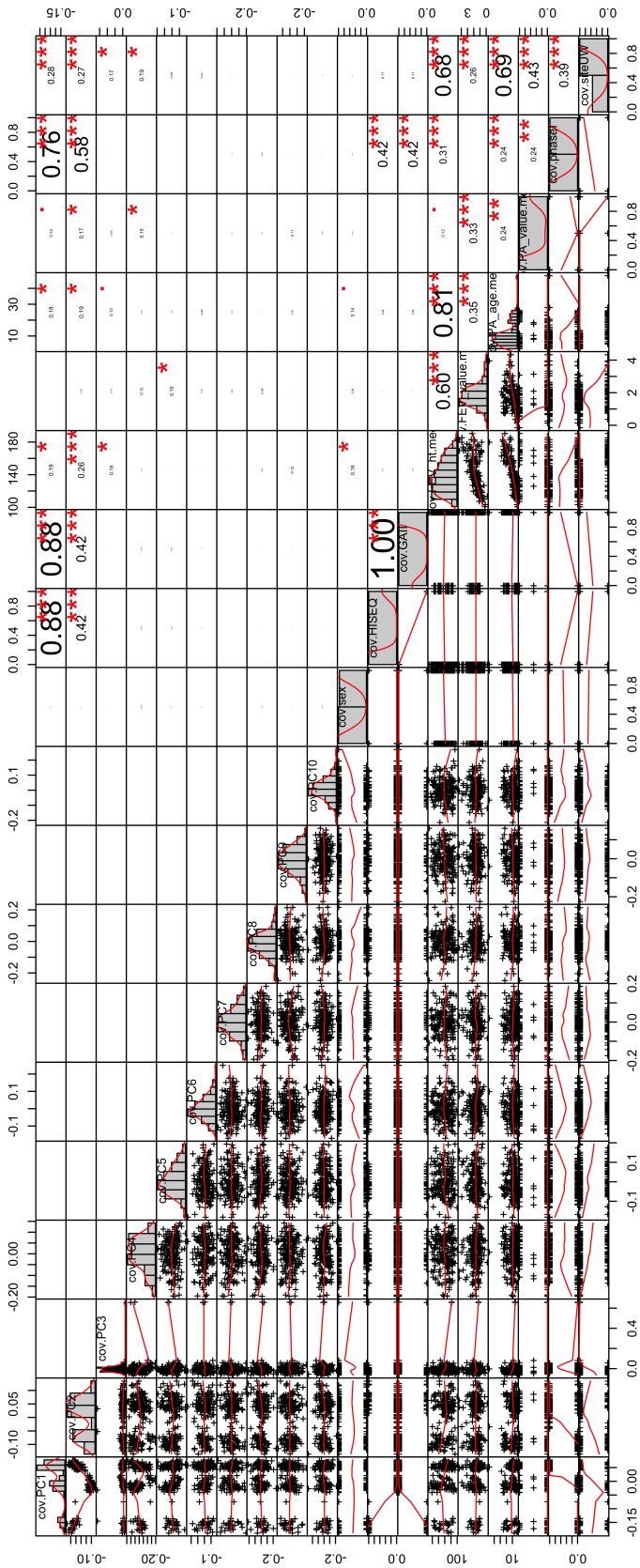


Figure 31. Correlation matrix for all variables considered in PA analysis. The data shown are the correlations in the PA analysis (181 individuals). The variables are labeled diagonally. The numbers on the x-axis and y-axis correspond to the distribution of values for each variable (see histograms).

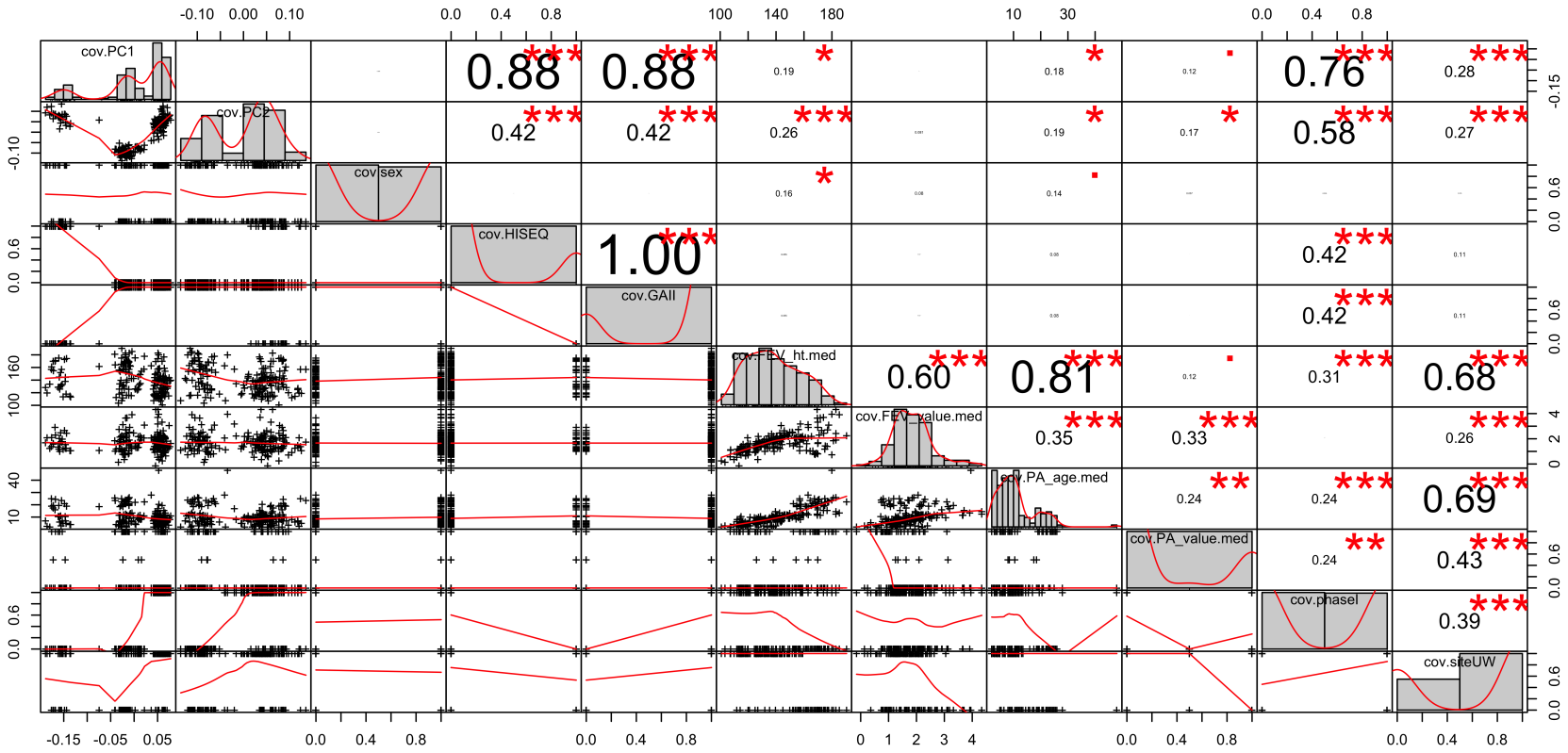


Figure 32. Correlation matrix for highly correlated variables in the PA analysis (181 individuals). The variables are labeled diagonally. The numbers on the x-axis and y-axis correspond to the distribution of values for each variable (see histograms).

both female (independent variable) and ever-smokers (independent variable). Separation can also occur if the outcome variable can be perfectly separated by a single independent variable. In logistic regression using maximum likelihood estimators, this causes problems fitting the model and results in at least one of the parameter estimates diverging to infinity (infinite log p-values). Occasionally this can cause an error; however, most of the time, the only diagnostic clue is unreliable standard error estimates, i.e., Wald test confidence intervals of infinite width. Firth introduced a modified score equation to include a penalty function which reduces bias and resolves the issue of separation. It is generally preferred to exact conditional regression because it can be implemented in cases of continuous and categorical variables and requires less computational power than exact conditional regression [123, 124, 125]. Firth regression was implemented using the R package "logistf" [126]. For each analysis, logistic regression used a maximum likelihood estimator (ML) and Firth logistic regression used a penalized likelihood estimator.

Elastic net analysis refers to a regularized regression method that can perform variable selection and prediction using a variety of parameter optimizations, bootstrapping, and/or cross validation. It is ideal when there are more predictors or modeling parameters than samples. Elastic nets use two penalty parameters, alpha (α) and lambda (λ), to shrink coefficients of inconsequential predictors and optimize coefficients for relevant predictors [127]. Although using a machine learning approach to select variables introduces bias to the analysis, it helps maintain reasonable error by building parsimonious models.

To perform variable selection, an elastic net algorithm was used to rank variables that model the phenotype; genetic data was not used to perform variable selection [128, 129]. For each phenotype, PA and PF, subsets of the covariate data were used in repeated cross validations to estimate optimal alpha and lambda parameters with minimal cross validation error. Once the optimal alpha and lambda values were obtained, 1000 models were built from bootstrap samples of the data. Variables that appeared in over 75% of the generated models were considered high priority and likely important predictors for the phenotype of interest. A threshold of 75% was an arbitrary designation to reduce noise from unimportant variables but still be fairly inclusive of any variables that consistently showed up across 1000 models. Depending on the alpha and lambda parameters chosen, the elastic net algorithm

can perform variable selection and also allow for correlated variables to remain in the model. For example, two highly correlated variables can be present in 85% and 90% of the 1000 bootstrapped models. As discussed before, correlations in covariates can lead to unstable error estimates in a logistic regression analysis. Therefore, the last statistical iteration included pruning the selected covariates using correlation measures to produce a minimally correlated subset of variables. Adjusting first for the subset of covariates and secondarily the selected and pruned covariates, statistical analyses were performed to detect an association between the outcome and each bin from BioBin.

Lastly, elastic net models were generated using all available covariates and low frequency bin data. In this more traditional application of elastic net, 10 cross validations were performed. First, subsets of the data were split into training (0.632) and testing sets. For each training set, 500 bootstrapped samples were used to estimate optimal alpha and lambda parameters with minimal cross validation error and to build a final model. Prediction accuracy, sensitivity, and specificity measurements were obtained by applying the final model to the testing set. This process was completed independently over 10 cross validations. The final models, selected variables, coefficients, and accuracy from each cross validation were compared.

In addition to varying binning parameters and statistical tests, various subsets of covariates were used to adjust the analyses. Each analysis was tested without adjusting for any covariates, adjusting only with significant principal components (with respect to explaining genetic variance), adjusting with all available covariates, adjusting for covariates based on variable selection, and adjusting for covariates based on variable selection and then pruned using correlation between covariates. Significant principal components were calculated using the Tracy-Widom statistic available in the Eigensoft package [130].

BioBin parameters

BioBin has a number of configurable parameters when performing analyses. These are described in Chapter III. In the following analysis examples, a binning threshold minor allele frequency of 5% was chosen. Gene bin analyses were performed using "no weight" and "minimum weight" approaches. Several of the analyses described use filters to separate

variants based on function. Nonsynonymous variants and predicted damaging variants were calculated using the Ensembl Variant Effect Predictor (VEP), which includes results from SIFT and PolyPhen-2 [72, 73, 74, 75]. A variant was labelled "damaging" if SIFT or Polyphen-2 identified the variant as damaging at any level (SIFT: "deleterious", PolyPhen-2: "possibly damaging", "probably damaging").

The CF exome data were binned based on gene regions determined by the NCBI Entrez gene source [61]. Depending on the analysis (PA or PF), approximately 15,000 gene bins were created. Several binning strategies were used: minimum weights or no weights, nonsynonymous variant filters, deleterious variant filters, and functional variant filters. Functional variant filters combined all nonsynonymous variants in a bin, but assigned custom weights to increase the influence of nonsynonymous variants (1.1) and predicted deleterious variants (1.2).

Pathway analyses were performed using multiple sources available in LOKI (see Chapter III). Similar parameter sweeps and statistical methods were applied to pathway analyses as previously described for the gene feature analyses.

Results and discussion

Pulmonary function phenotype

The pulmonary phenotype analyses were performed on 114 cases (severely affected pulmonary function) and 122 (mildly affected pulmonary function) controls. Several of the covariate measures were highly correlated with the phenotype, including FEV measurement, age, and height. The impact of covariate correlation is very important to result interpretation. For instance, when one of the covariates is highly predictive of the phenotype, separation can occur. Correlations between two or more "independent" variables can lead to a non-uniform p-value distribution and possibly an increase of false-positive results. In Figure 33, quantile-quantile plots (QQ plots) are shown for a few results from the pulmonary function analysis. In this analysis, no variant weight was used and only

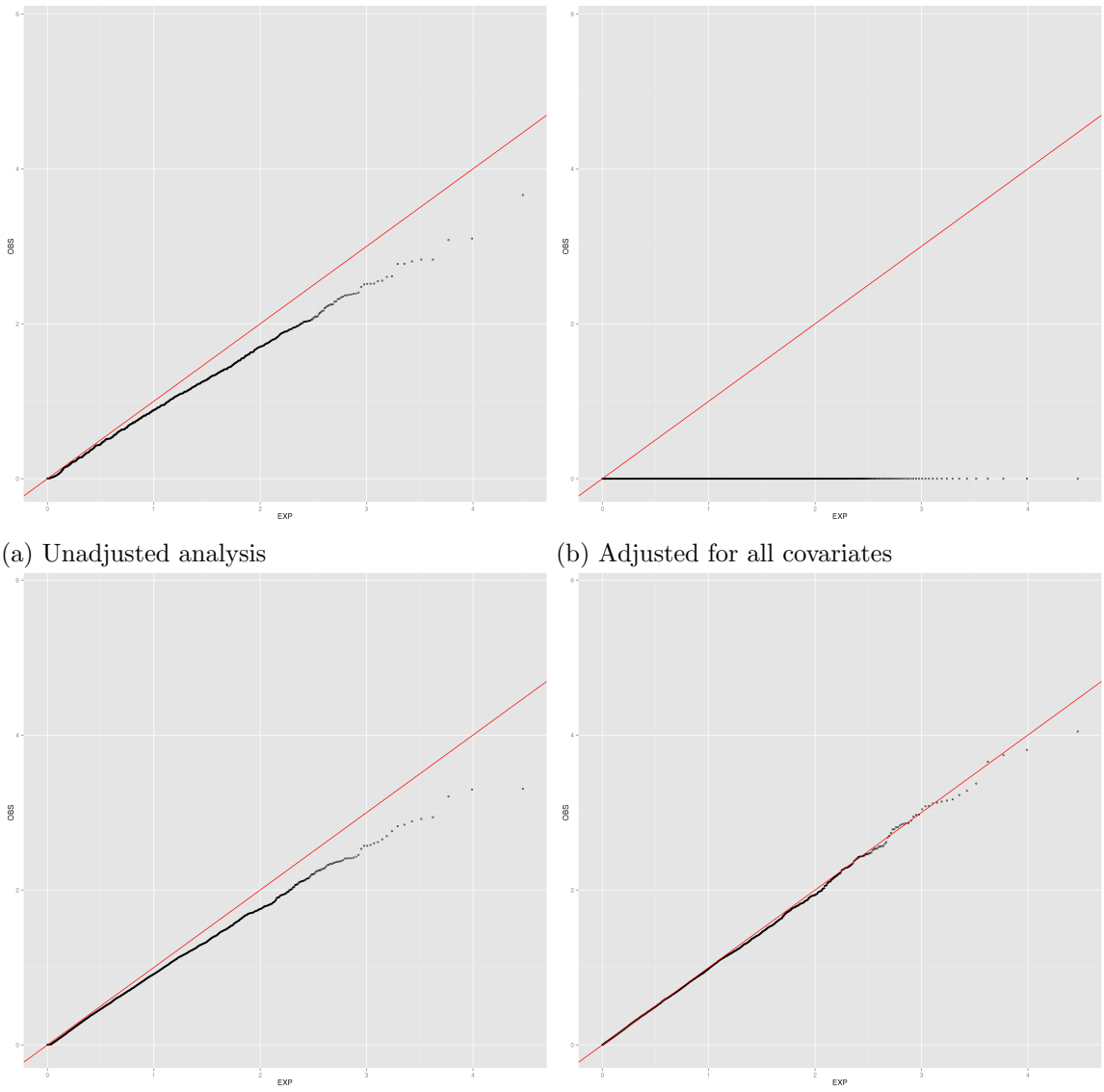
nonsynonymous variants were binned.

Figure 33a shows the unadjusted QQ plot, which does not quite fit the expected uniform distribution. Figure 33b shows the QQ plot adjusting for covariates listed in the caption. This plot is highly abnormal; all of the expected log p-values are 0. In the results, these are actually infinite log p-values. Figure 33c shows the QQ plot adjusted for ONLY covariates using variable selection (without correlation pruning). Lastly, Figure 33d shows the same data as Figure 33c except using a Firth logistic regression instead of a maximum likelihood estimated logistic regression which is sensitive to separation.

In the example shown in Figure 33b, the offending covariate is FEV, which was correlated to the measure used to phenotype individuals as severe or mild phenotypes and highly correlated to height and age. If a correlated predictor variable is critical to the analysis, one can use Firth regression, a penalized maximum likelihood estimator which better estimates coefficients and standard errors in cases of perfect or near perfect separation. As shown in Figure 33d, the penalized regression manages error much better than standard logistic regression and produces more uniform log p-values.

After evaluating the p-value distribution of models which used the ML logistic regression, variable selection with correlation pruning that included only significant principal components best resolved the issue of separation, i.e., two or more correlated predictors were not perfectly predicting the outcome. There were no highly significant hits from the PF gene bin analysis, which might be expected given the small sample size and data stratification present, e.g. different sequencing technologies, study phase, study sites, etc. However, Firth regression provided the most reliable results. A few of the top ranked results were biologically relevant and interesting (results are shown in Table 22). The p-values provided for each bin are uncorrected for multiple testing, such as Bonferroni correction. This analysis adjusted for selected covariates pruned for correlation or significant principal components, using no variant weights, and binning only nonsynonymous variants.

None of the results shown in Table 22 are significant after Bonferroni correction. A few of the results *PPP1R9A*, *EPS8L1*, and *FERMT1* are categorically related to cytoskeletal structure, which is potentially pertinent since severely affected lung phenotypes are likely to be fibrotic. Two other results involve ion channels, *SGK3* is a protein kinase involved



(a) Unadjusted analysis

(b) Adjusted for all covariates

(c) Adjusted for select subset of uncorrelated covariates

(d) Adjusted for all covariates using FIRTH regression

Figure 33. Quantile-quantile plots for pulmonary function (PF) analysis illustrating effect of perfect separation in PF data. (a) QQ plot for unadjusted gene bin analysis using ML logistic regression. (b) QQ plot for adjusted gene bin analysis using ML logistic regression, analysis adjusted for top 10 principal components (PC), gender, sequence technology, FEV measurement, age, height, and site (c) QQ plot for adjusted gene bin analysis using ML logistic regression, analysis adjusted for selected variables from elastic net ranking and post-ranking elimination of highly correlated variables: PC1, PC8, PC9, PC10 and height (d) QQ plot for adjusted gene bin analysis using FIRTH logistic regression, analysis adjusted for top 10 principal components (PC), gender, sequence technology, FEV measurement, age, height, and site

Table 22. Top results from PF analysis using no variant weights and binning only nonsynonymous variants. The study refers to one of two analyses: analysis adjusted for elastic net selected covariates and pruned for correlation or the analysis adjusted for significant principal components. The location refers to the variants contained in the respective bin: chromosome and first and last base pair location of binned variants. The p-value is a naive p-value pre-multiple test correction.

Gene	Study	Analysis	Location	Number of Loci	Variants (Cases)	Variants (Controls)	P-value* (Firth)	Notes
FERMT1	PF	VarSel	chr7:43982418-102309418	3	17	3	$9.0e^{-05}$	Cell adhesion and contributes to integrin activation, related to matrix invasion in colon cancer, mutations cause Kindler syndrome
SGK3	PF	VarSel	chr8:67590042-67748231	2	13	2	$1.6e^{-04}$	Serine/threonine-protein kinase involved in ion channel regulation modulates influenza virus replication [131]
ANO4	PF	VarSel	chr12:101333155-101493469	8	37	12	$1.8e^{-04}$	Transport cytosolic Cl- to extracellular space, previously associated in GWAS with heart ailure, BMI and cholesterol
COPS5	PF	varSel	chr8:67955495-67971452	3	13	2	$2.2e^{-04}$	Subunit of COP9 which regulates various signaling pathways
ALS2CL	PF	varSel	chr3:46712490-46729610	18	33	13	$4.2e^{-04}$	Mediates endosome dynamics, candidate gene for schizophrenia [132]
POLR2J	PF	sigPC	chr7:43982418-102309418	6	35	10	$5.2e^{-06}$	Subunit of RNA polymerase II, very involved in HIV replication and life cycle
EPS8L1	PF	sigPC	chr19:55587841-55598793	9	16	1	$1.3e^{-04}$	Signal transduction leading to actin cytoskeleton remodeling
SLAMF9	PF	sigPC	chr1:159922092-159923904	2	11	35	$1.7e^{-04}$	Signaling lymphocyte activation molecules, likely associated with macrophage inflammation
ANO4	PF	sigPC	chr12:101333155-101493469	8	37	12	$2.9e^{-04}$	Transport cytosolic Cl- to extracellular space, previously associated in GWAS with heart failure, BMI and cholesterol
PPP1R9A	PF	sigPC	chr7:94740687-94917894	7	10	0	$4.1e^{-04}$	Imprinted, cytoskeleton reorganization

Gene annotations were found in UCSC Genome Browser and the NCBI database [71, 133]

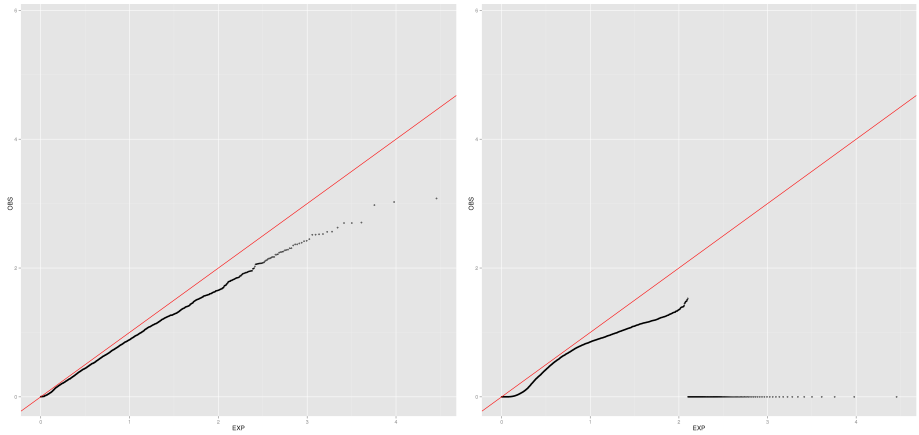
with ion channel regulation. However, another member of the *SGK* family, *SGK1*, regulates *CFTR* conductance [134, 135]. Lastly, *ANO4* is a calcium-mediated chloride ion transport, essential for Cl- secretion in epithelial cells, smooth muscle peristalsis, and olfactory signaling. The relationship between *ANO4* and *CFTR* has been tested in mice but requires further research [136].

Recurrent Pseudomonas aeruginosa phenotype

The PA phenotype analyses were performed on 87 cases and 94 controls. Again, many of the covariates were highly correlated. In Figure 34, quantile-quantile plots (QQ plots) are shown for few of the results from the PA analysis. The expected log p-values from a uniform distribution are shown by the red line, observed log p-values are shown in black. In the particular analysis shown, all nonsynonymous variants were binned with a functional weight applied, i.e., nonsynonymous variants received a weight of 1.1 and predicted deleterious variants received a weight of 1.2.

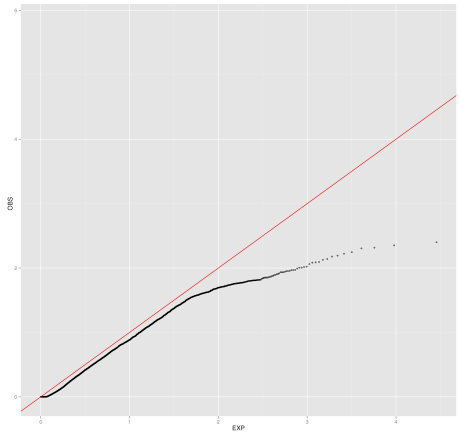
As shown in Figure 32, median age, median FEV measurement, and median height are correlated in the PA data. Figure 34a shows the unadjusted QQ plot, which does not quite fit the expected uniform distribution. Figure 34b shows the QQ plot adjusting for all possible covariates (listed in the caption). This plot has an unusual observed p-value distribution with several observed infinite log p-values. Figure 34c shows the QQ plot adjusted for ONLY covariates using variable selection (without correlation pruning). Figure 34d shows the QQ plot adjusted for covariates using variable selection and correlation pruning. In this case, site and height variables were dropped due to high correlation with other selected variables. Figure 34e shows the QQ plot after adjusting for only significant principal components using the Tracy-Widom statistic [130]. Lastly, Figure 34f shows the same data as Figure 34d except using a Firth logistic regression instead of a maximum likelihood estimated logistic regression which is sensitive to separation.

After evaluating the p-value distribution of models which used the ML logistic regression, variable selection with correlation pruning and using only significant principal components best resolved the issue of separation. However, Firth regression provided the most reliable results. A few of the top results using Firth regression are listed in Table 23.

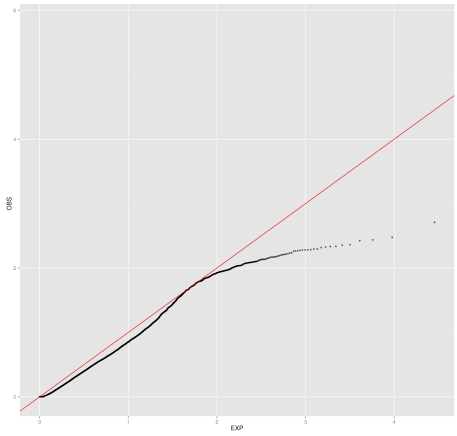


(a) Unadjusted analysis

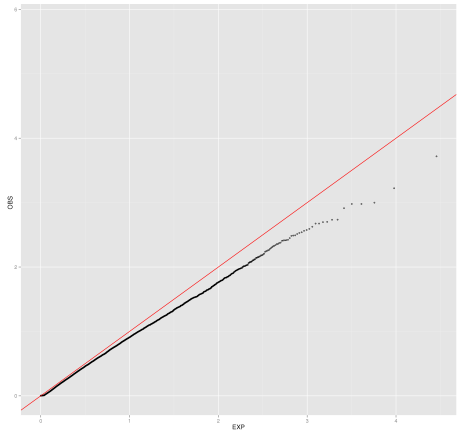
(b) Adjusted for all covariates



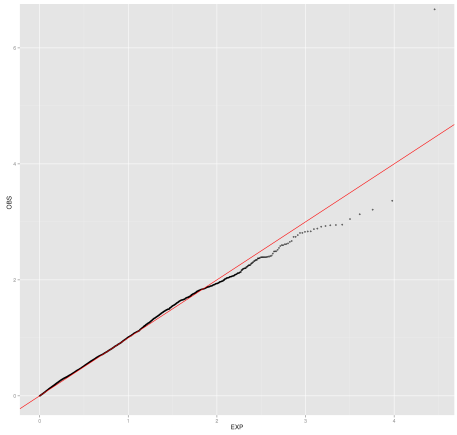
(c) Adjusted for select covariates



(d) Select covariates and pruned



(e) Significant principal components



(f) Select covariates-FIRTH regression

Figure 34. Quantile-quantile plots for *Pseudomonas aeruginosa* infection (PA) analysis (a) unadjusted gene bin analysis using ML logistic regression, (b) adjusted for top 10 principal components (PC), gender, sequence technology, FEV measurement, age, height, site, and PA culture using ML logistic regression, (c) gene bin analysis adjusted for selected variables from elastic net ranking using ML logistic regression, (d) adjusted for selected variables from elastic net ranking and post-ranking elimination of highly correlated variables using ML logistic regression,(e) ML logistic regression adjusted for significant principal components (PC), (f) FIRTH logistic regression adjusted for same variables as Figure 34c.

There is only one significant result from the PA gene bin analysis after Bonferroni correction, which might be expected given the small sample size and the aforementioned data stratification present (e.g., different sequence technology variables, phase variables, site variables, etc). However, some of the top ranked results were biologically relevant and interesting. Results from a PA analysis adjusting for selected covariates pruned for correlation or significant principal components, using minimum variant weights, and binning only nonsynonymous variants are shown in Table 23. The p-values for each bin are unadjusted for multiple test correction, i.e. prior to a Bonferroni correction.

Several factors are important to consider when interpreting these results; significance, bin size, and direction of effect. Only *ACER3* meets significance after a multiple test correction. With only two loci in the bin, and particularly because one group (cases) does not have any variants in this bin, it is important to consider stratification. This could be explained with sequence technology differences or differences in variant calling.

Many results appear protective of recurrent PA infection, i.e., there are more variants in controls than cases. The number of variants in each group shown in Table 23 is the sum of variants in that particular group. In the PA analysis, the sample size is uneven, there are seven more controls than cases. Therefore, the sum of variants shown in Table 23 does not immediately translate to effect size.

Some results listed in Table 23 do not appear to have any extensive relationships with infection, *CFTR*, or cystic fibrosis in the literature but do have some relevancy. For example, *SMG6* encodes a protein responsible for nonsense mediated decay (NMD), which is the mechanism used to degrade defective *CFTR* proteins with premature stops. However, there is no literature supporting a direct association of *SMG6* with cystic fibrosis or infection [138]. Another example is *FAM120A*. According to IntAct, a database of interacting proteins, *FAM120A* physically interacts with the *CFTR* protein according to an "anti bait coimmunoprecipitation" experiment, but there is no literature further defining this relationship [139]. The last example is *MLLT4*, which does not contain any direct links in the literature to CF, but is associated with elevated states of immune response. This is potentially interesting since much of the damage to lung epithelial cells is caused by chronic inflammation and bacterial response.

Table 23. Top results from PA analysis using no variant weights and binning only nonsynonymous variants. The study refers to one of two analyses: analysis adjusted for elastic net selected covariates and pruned for correlation or the analysis adjusted for significant principal components. The location refers to the variants contained in the respective bin: chromosome and first and last base pair location of binned variants. The p-value is a naive p-value pre-multiple test correction.

Gene	Study	Analysis	Location	Number of Loci	Variants (Cases)	Variants (Controls)	p-value* (Firth)	Notes
ACER3	PA	VarSel	chr11:76726103-76731362	2	0	2	$4.7e^{-08}$**	Located in ER and Golgi apparatus membranes. Hydrolyzes sphingoceramide into sphingosine and free fatty acid
GPT	PA	VarSel	chr8:145729726-145732151	8	3	8	$2.8e^{-04}$	Key role in the intermediary metabolism of glucose and amino acids
SMG6	PA	VarSel	chr17:1964797-2203527	17	16	30	$1.3e^{-04}$	Plays a role in nonsense-mediated mRNA decay. GWAS associations with CAD and T2D
CYBA	PA	varSel	chr16:88709869-88713533	4	16	4	$2.6e^{-04}$	Encodes light chain subunit of cytochrome B, primary component of the microbicidal oxidase system of phagocytes, important in inflammatory response
MLLT4	PA	VarSel	chr6:168281125-168352727	13	28	17	$3.1e^{-04}$	Contain GWAS hits for systemic lupus eyrthematosus, triglycerides, and ALS
EIF3K	PA	sigPC	chr19:9109975-39125671	2	9	33	$8.20e^{-06}$	Subunit of translation initiation factor 3
DHCR7	PA	sigPC	chr11:71146481-71155234	2	2	19	$1.80e^{-04}$	Adaptative locus in European populations affecting vitamin D metabolism [137], cholesterol biosynthesis
PRB4	PA	sigPC	chr12:11461444-11463283	10	17	51	$3.30e^{-04}$	Encodes proline-rich salivary protein
SLCO2B1	PA	sigPC	chr11:74862426-74915566	11	12	34	$7.30e^{-04}$	Encodes member of anion transporting membrane proteins
FAM120A	PA	sigPC	chr9:96214890-96326824	10	6	36	$5.00e^{-04}$	constitutive coactivator of PPAR-gamma-like protein 1

Gene annotations were found in UCSC Genome Browser and the NCBI database [71, 133]

* P-values presented prior to multiple testing correction

** The only result significant AFTER Bonferroni correction

A few of the results in Table 23 have more relevant relationships with cystic fibrosis, *CFTR*, or infection. Three are discussed in further detail below.

The *ACER3* bin contains only two variant sites and was the only bin to remain significant after Bonferroni correction. *ACER3* encodes an alkaline ceramidase that metabolizes ceramide to form lysolipid sphingosine. Both ceramide and sphingosine are signaling lipids; ceramide is involved in inflammation and apoptosis and sphingosines primarily induce cell proliferation and differentiation. Severe mutations in ceramidases can lead to lysosomal storage disease, but an imbalance between ceramide and sphingosine can lead to an accumulation of ceramide and cell death on lung epithelial cells [140, 141]. Teichgräber et al. published the pioneer observation of ceramide accumulation. In *Cftr*-deficient mice, an accumulation of ceramide in pulmonary epithelial cells resulted in age-dependent pulmonary inflammation, death of respiratory epithelial cells, deposits of DNA in bronchi and high susceptibility to severe *Pseudomonas aeruginosa* infection. In their particular mouse model, heterozygous genetic deficiency of acid sphingomyelinase (*Asm*) or administration of an *Asm* blocker, such as amitriptyline, normalized pulmonary ceramide and prevents all pathological findings, including susceptibility to infection [140]. In humans, acid sphingomyelinase (*SMPD1*) proteins convert sphingomyelin to ceramide [133].

In Table 23, *DHCR7* has more variants in controls than cases. *DHCR7* encodes the enzyme 7-dehydrocholesterol reductase that catalyzes the conversion of 7-dehydrocholesterol to vitamin D_3 using sunlight. Recent results suggest that *DHCR7* has been under selection in recent evolutionary history to allow European populations to avoid severe vitamin D deficiency as populations moved away from Africa and the equator. In the setting of cystic fibrosis, which is predominant in Caucasian populations, this is quite interesting. However, even more interesting are the implications of vitamin D deficiency, which include: cancer, autoimmune disease, infection, and cardiovascular disease [137]. Active vitamin D is a immune modulator; in an example by Holick, immune cells exposed to *Mycobacterium tuberculosis* up-regulate the vitamin D receptor gene. Increased production of active vitamin D produces cathelicidin, a peptide that can destroy many infectious agents. Low serum levels of active vitamin D prevent this innate immune response, which results in increased and more aggressive infections [142].

Lastly, *CYBA* is considered. Inflammation in the lung epithelial cells in patients with CF begins at an early age and is predominated by neutrophils. In CF patients, the persistent inflammation fails to resolve the recurrent bacterial infections. It is thought that the damage caused by neutrophils in CF is due to excessive reactive oxygen species production, thus inducing airway damage and promoting bronchiectasis and fibrosis in the lungs [143]. In Table 23, *CYBA* is shown to have more variants in cases than controls. *CYBA* has been previously associated with autosomal recessive chronic granulomatous disease (CGD), characterized by a lack of reactive oxygen production in neutrophils to kill pathogens, resulting in a diminished immune system and excessive bacterial and fungal infections [144]. Excessive or lack of reactive oxygen species production by neutrophils can cause chronic and severe infection.

Pathway and elastic net analyses

Lastly, a pathway burden analysis was compared to an elastic net approach using PA data. Again, none of the pathway results were significant after Bonferroni correction for multiple testing. Most of the top ranked results were BioGrid interaction groups; many had at least one bin from the top hits listed in Table 23. A few were unique signals, for example, BioGrid ID:637466 represents a physical interaction between *CKS2* and *LDHB*. Cyclin kinase 2 plays a critical regulatory role in meiosis and mitosis. Abnormal expression has been associated with many types of cancer [145]. Lactase dehydrogenase B encodes a subunit of lactase dehydrogenase, which converts pyruvate to lactate and NAD to NADH. The expression of *LDHB* can also be unregulated in cancers [133]. Another top hit included BioGrid ID:120309, which represents a physical interaction between *CTPS2* and *SPG21*. *CTPS2* encodes a catalyzing enzyme to convert CTP to UTP and deamination of glutamine to glutamate. Cells with increased cell proliferation also exhibit increased activity in *CTPS2* [71]. The protein encoded by *SPG21* binds to *CD4* to repress T cell activation. A mouse knockout of *SPG21* had almost two fold difference of *CFTR* expression in brain tissue [146], but a clear relationship for how these interactions might be linked with chronic PA infection is

unclear. *SPG21* could regulate *CFTR* expression or this could be a false positive result in the expression data, future studies are needed to confirm.

In the 10 cross-validations of the elastic net, only two yielded models with non-zero coefficients. The accuracy of both models in the respective test data set was 89%, but there was only one coefficient common to both models, *PPAP2C*. *PPAP2C* is a member of the phosphatidic acid phosphatase family (PAP). The function of this PAP is to convert phophatidic acid to diacylglycerol, participate in de novo synthesis of glycerolipids, and phospholipase mediated signal transduction [71]. It is also involved in the sphingolipid metabolism described in the previous section for *ACER3*.

In one of the two models formed during cross validations, there were six non-zero coefficients. Three of which could be distantly mapped using IMP [147] to genes related to cell cycle. The network for *CDK2*, *PPAP2C*, and *DNAJC9* is shown in Figure35. *CDK2* is a member of the Ser/thr protein kinase family, it is important for G1/S phase transition in the cell cycle. *DNAJC9* encodes a heat shock protein that can also map back to cell cycle regulation.

Conclusions

Care must be take to exclude or manage highly correlated covariates if performing ML logistic regression analyses. As shown in Figure 33 for the PF analysis, the log p-values were essentially unreliable using maximum likelihood logistic regression. FEV values were correlated with the diagnosis of severe pulmonary function because the FEV measures part of the diagnostic criteria. Firth regression requires considerably more computational resources, but is highly recommended in analyses with covariate correlation.

Clinical and demographic covariates were used to build a model predicting the phenotype; therefore, the bias generated from unaccounted degrees of freedom in the secondary regression analysis was likely unsubstantial. The results from this variable selection analysis, which served to reduce error in the resulting models, were more consistent with expected distribution of p-values than ML regression. A penalized likelihood estimator better controls

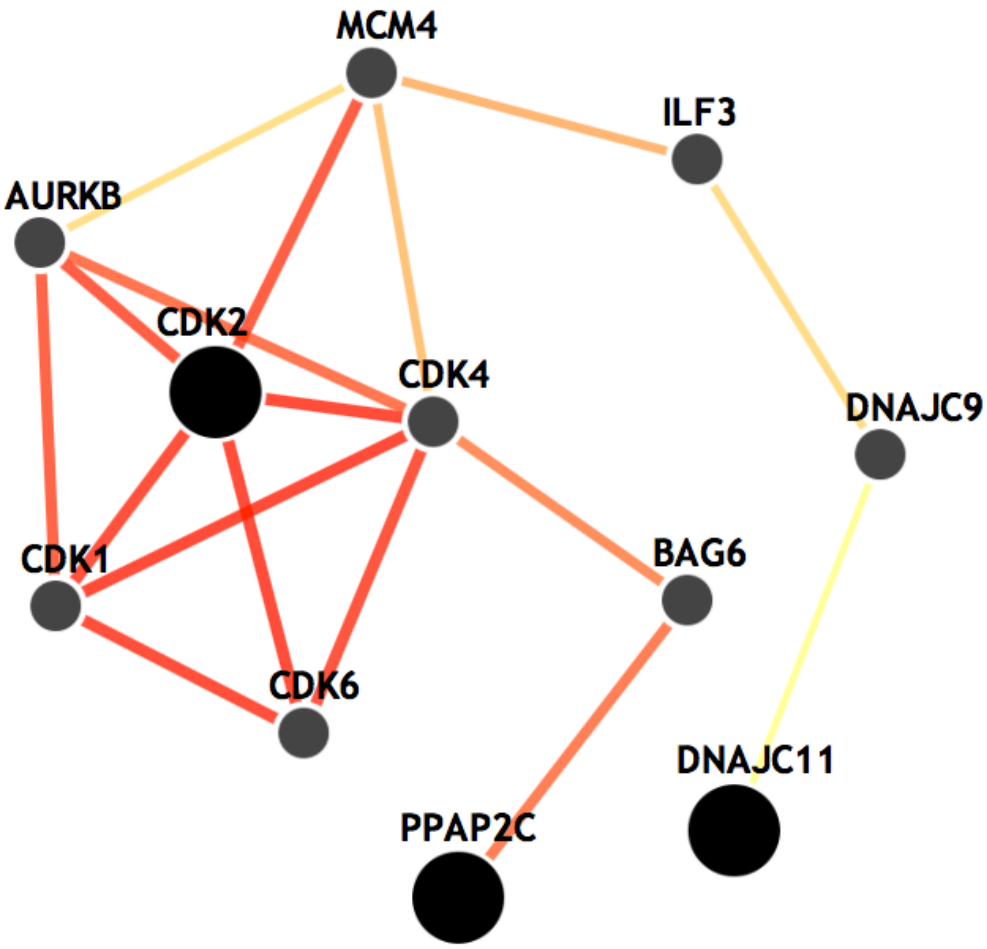


Figure 35. Gene set network for one of the top models. Three out of the six nonzero coefficients were connected using IMP.

errors when any correlation is present in the data.

Previously, authors of the dbGaP cystic fibrosis dataset published a study identifying *DCTN4* as a likely candidate gene modifier of PA infection using only the phase I data from this study. The authors used an additive rare variant threshold approach (RVT1) approach with a minor allele frequency threshold of 12.5% [119, 148]. After Bonferroni correction, *DCTN4* was significantly associated with chronic infection (adjusted p-value 0.025). Twelve out of 43 cases in the phase I data had a missense mutation in *DCTN4* [119]. In the PA analysis using only nonsynonymous variants described in this thesis and all phase I and phase II individuals, *DCTN4* had 4 loci. Sixteen variants were present in cases and 10 variants were present in controls (p-value = 0.3349, unadjusted for multiple testing). The results are likely not replicated in our study because the sample size was doubled by including phase 2 individuals, different minor allele frequency cutoffs were used, individuals were excluded that did not cluster with European descent populations, and potential noise contribution from neutral variants.

The pathway analysis and elastic net analysis did not yield any exceptional results. However, an important lesson can be learned. Noise contribution and variants of opposite effect have been discussed many times in this thesis (Chapter III and Chapter IV) and ultimately reduce power to identify true associations. Pathway analyses, such as the one presented above, have increased noise and most likely have increased numbers of variants with opposite effect. There are more variants included in the bin analysis, since multiple gene bins are collapsed into one pathway bin. But more importantly, pathways have known positive and negative regulators that can impact downstream function. A missense variant upstream in a negative regulator might increase overall function of a particular pathway, but a similar variant elsewhere could have a different effect. Therefore, until better variant selection methods are available for BioBin, pathway analyses utilizing burden tests are unlikely to contribute new information to low frequency variant analyses.

An alternate to the burden test or a machine learning technique is preferable for network analyses. The potential for pathway analyses using nonburden tests is somewhat better since these tests allow variants to be binned together even if the direction of effect is different among variants. Alternatively, variants binned in a similar way will allow tools such as

elastic net algorithms to build effective models from the data. This approach also benefits from the flexibility to add common variants to analyses. The elastic net models produced from the PA analysis do have some relationship with cell cycle regulation, but the fact that this model was not replicated in each cross validation decreases the likelihood it has any significance to the PA phenotype.

Conservative correction for the total results of all conducted analyses would eliminate all significant results. In previous sections where multiple test correction has been mentioned, the p-values were corrected for only the test performed within the context of a single analysis. For example, if a single bin analysis resulted in 15,000 bins, the Bonferroni correction would only account for those bins, not for every 15,000 bins created for additional parameters (weight testing, etc).

Lastly, most of the results discussed in this chapter were not significant after a multiple test correction. This could be related to sample size and/or the stratification introduced by multiple sequencing technologies, three separate exome capture kits, and multiple collection/sequence sites. However, the type I errors were well-controlled and the results could be prioritized for follow-up studies. Several of the bins in the PA gene analysis were particularly interesting in the context of cystic fibrosis disease, the *CFTR* gene, and chronic *Pseudomonas aeruginosa* infection.

CHAPTER VII

CONCLUSIONS

To explain additional phenotypic variation in common complex disease, it is imperative to consider genetic variation beyond common single nucleotide polymorphisms. Rare variant analyses are appealing since effect sizes are potentially larger and increasingly available in next generation data sets. However, to improve power, one must consider groups of rare variants with similar properties. The most powerful application of collapsing methods groups detrimental variants with other detrimental variants to effectively "build" a detectable signal. Alternatively, the least powerful application of a collapsing method would group variants with opposite directions of effect or include a significant number of variants that contribute to noise but no meaningful signal. BioBin is a novel collapsing method that uses allele frequency data and biological information to bin rare variants.

BioBin is unique because it is driven by a powerful database of biological/computational knowledge, does not require any one explicit statistical paradigm, and provides bins with direct interpretations of biology. The user can easily test complicated and interesting hypotheses on many features. LOKI provides access to integrated biological knowledge (pathways, groups, interactions, ECRs, regulatory regions, etc.), which is valuable to researchers that do not want to spend considerable effort to combine this knowledge manually. Additionally, the output of BioBin can be subsequently analyzed using the association test most appropriate for their specific scientific question. For any given bin analysis, many statistical tests including those from other published collapsing methods can be applied to BioBin output.

In this thesis, the benefits for studying low frequency variants and the advantages of using biological knowledge from BioBin have been outlined. In Chapter II, the properties of low frequency variants and their potential contributions to heritability are described. In addition, multiple published methods to study low frequency variants are reviewed.

Chapter III provides a thorough overview of the BioBin software and relevant analysis parameters or options. Extensive simulation studies testing BioBin and its various weights and statistical tests are presented in Chapter IV. Chapter V chronicles a study of population stratification in low frequency variant bins using Phase I 1000 Genomes data. Lastly, Chapter VI contains demonstration analyses from two applications of BioBin to natural data sets.

Because of the technological revolution resulting in increased available sequence data and available biological knowledge, BioBin will become an even more useful analysis tool. The ability to quickly form and test unique, interesting, and biologically relevant hypotheses using aggregated low frequency variation will aide scientists in revealing hidden heritability for common complex disease.

Evaluation of the analyses presented

Strengths of this approach

BioBin provides a tool for researchers with insurmountable data sets to utilize vast biological knowledge for hypothesis generation. BioBin is a novel automation of binning analyses that provides limitless flexibility. Users are able to bin at many levels of information: introns, exons, genes, pathways, regulatory regions, and multiple other combinations. The Library of Knowledge Integration (LOKI) database provides an assortment of established publicly available databases to guide binning. The structure of LOKI also allows users to import additional or external knowledge sources to satisfy individual binning needs.

Binning methods increase power over single low frequency variant association tests. BioBin reliably identifies regions or collections of variants with differences in low frequency variant distributions associated with the outcome of interest. It also affords users the ability to perform complex filtering and weighting to increase the power to detect association for any given bin.

Unlike other available methods, BioBin software is not restricted to any particular sta-

tistical test. Users are able to apply burden and nonburden tests and machine learning methods to BioBin output. This flexibility encourages users to identify and utilize statistical tests that best fit their data and hypothesis.

Limitations of this approach

In many cases, binning approaches are underpowered to detect true associations. Although filtering and weighting approaches are available in BioBin, more work is needed to overcome noise contributed by neutral variants or diminished signals from many variants of opposite effect (i.e.., detrimental versus protective effects). In the implementation described in this thesis, BioBin is limited by known biological knowledge. For example, the user is not able to filter variants beyond lists of variants in known databases (e.g. 1000 Genomes Project database), currently available prediction algorithms (e.g., SIFT or PolyPhen-2), or custom knowledge files. Nonburden tests or additional methods of filtering should be added to the binning analysis pipeline to overcome this loss of power.

Another limitation of using sequence data are genome reference. Variant calling format (VCF) files are a common output from next-generation sequencing studies and are the standard input file for BioBin. These files essentially contain all variation with respect to the human reference genome. In theory, variation with respect to some standard is exactly what should be compared in a bin analysis. In reality, the absence of a variant with regard to the reference sequence and thus absence from a VCF file can also be due to missingness (failed to meet quality standard to call variant with confidence) or off-target sites (not within target region of exome capture kit). This problem will only be alleviated when projects routinely sequence high depth of coverage ($>$ 30x) and consistent capture technology is used throughout the study.

The foundation of the BioBin approach is ultimately based on base pair proximity and biological knowledge. Using BioBin implies that a collection of rare variants from a predefined region confer an increased odds or risk of the phenotype of interest. For example, causative low frequency variants within a gene increase the odds of developing schizophre-

nia. Regardless of feature, every bin generated collapses variants based on contiguous genomic coordinates. Of course, this limits the user to identifying collections of variants defined proximity in genomic space, though this may not correspond to biological function (interregion bins).

Lastly, the biggest limitation of any binning approach is data quality and size. In order to reliably identify meaningful associations, artifact signals causing false positive associations must be smaller than the signal of interest (i.e., favorable signal to noise ratio). Current study designs have not afforded this opportunity. In efforts to progress genomic research and be considerate of costs, many studies combine samples from multiple platforms/sites/ancestral backgrounds to increase power to study rare variants. Unfortunately, these studies suffer from stratification due to ancestry, imprecise phenotyping, site-specific sequence technology, differing target capture kits, and ambiguous/inconsistent data processing pipelines. In addition, larger sample sizes are needed to establish reliable allele frequency calculations and increase the number of binnable variants.

BioBin is reliable at detecting bins with differences in low frequency variant burden; however, many times those differences are because a disproportionate number of controls were sequenced with a different technology or capture kits. Due to the process of locus selection, BioBin is sensitive to data quality and stratification. Statistical results from BioBin output can be misleading if confounding factors are not considered and/or prevented.

Binning considerations

Study design

Good study design is critical in low frequency variant studies. Do et al. considers sample selection to be the most critical component in exome-sequencing studies [6]. It is important to catalogue potential samples and consider the most relevant study design:

1. Population samples

2. Case/control studies

3. Families segregating Mendelian traits

Since the cost of sequencing is still quite expensive, researchers must carefully select the study design and samples to be sequenced. In most cases, it is beneficial to focus on extreme phenotypes [6].

Another consideration in sample selection is ancestry. The geographic distribution of low frequency variants is highly dependent on ancestry; refer to Chapter V for more information about ancestral differences in low frequency variant burden. Extreme care must be taken to accurately match case and control groups in low frequency variant studies [6]. It is relatively easy to incur false positive results due to population stratification and convenience control samples most often are not the best matched controls. Unfortunately, it is unclear that principal components can reliably and adequately adjust for population stratification in low frequency variant studies.

In addition to carefully matching cases and controls for ancestry, it is imperative to apply strict guidelines for the handling of case/control samples. Bias and type I errors can easily result from study designs that sequence and process samples differently: sequence technology, sequence site, and variant calling pipelines. Although it is currently prohibitively expensive, sequence studies will be more reliable when samples can be sequenced in duplicate and perhaps on more than one technology or at more than one site. Concordant data will enhance the accuracy of the analysis results.

Considering all of these recommendations, the ideal study for a BioBin analysis would include at least 4000 extreme phenotypic cases (N=2000) and controls (N=2000) from a single cohort study of similar ancestry. The phenotype should have well established heritability. Thousands of samples are needed to establish reliable allele frequency estimates generalizable to the sample population. Careful sample selection with regard to phenotype and ancestry will increase the power of finding an association and decrease potential error inducing noise. To identify the majority of variants with high specificity, filtered raw reads should have at least 20x coverage at 80-95% of the sequence [6]. In an ideal study, each sample would have whole-genome sequence data with a mean depth of coverage greater than

or equal to 50x. At this level of coverage, the number of variants identified is less correlated to depth of coverage [6]. Samples would be sequenced and processed at the same center with nominal error rates. Whole genome sequence is preferred because it provides many more avenues of discovery outside of coding regions. Higher mean coverage and controlled sequencing and processing reduces the number of errors and potential stratification in the data. The larger sample size and carefully phenotyped samples will increase the power of finding true associations.

Replication in rare variant analyses

Statistical replication is a critical step to validate genomic study results. Most commonly, replication is performed in a similar but independent dataset and the replicating signal must have the same direction of effect in the same SNP or a SNP in very high LD with the originally associated SNP [149]. This is often referred to as a "signal replication" since association with any tag SNP represents the same signal.

Others hold the stringent view that a signal must be detected in the same SNP to be considered a true replication; in this text defined as a "strict replication." In common variant association tests, which hinge on tag SNPs being in high LD with other common variants, strict replication is not practical. For example, SNP1 and SNP2 are in very high LD with each other. In a published study, authors only investigated SNP1 and found it to be associated with phenotype X. If a second group of researchers published an association between phenotype X and SNP2, this would not be considered a replication using "strict" replication guidelines. Since common variant studies utilize SNP-tagging for genotyping and association analyses, it is impractical to believe that results from a statistical test prove definitive causation. Most putative SNPs with a common allele frequency can be accurately tagged by more than one tag SNP. It is also impractical from a genotyping standpoint. Platforms vary in genomic coverage and strict replication across two studies may not be possible based on available genotypes. If constrained to strict replicate identification, many signal replications will be missed. However, in rare variant analyses, which are most often

in low LD or in very low frequency haplotypes, strict replication is more applicable.

In current literature, rare or low frequency single variant replications generally adhere to the definition for strict replication. Guey et al. detail steps for replication in low frequency analyses. The authors suggest first detecting an association in extreme phenotype samples, and then replicating in a similar population or cohort using random sampling. The extreme phenotypic sampling allows for variant prioritization, but also suffers from "winner's curse" similar to observations made from GWAS [149, 150], where the resulting association signals in an extreme phenotypic sample study are often overestimated. To assess the true effect size, the prioritized variants must be assessed in a randomly sampled population of similar ancestry to the original sample. In addition, to accurately state that a variant was replicated or not replicated, the replication dataset must be adequately powered to detect an association [149]. For researchers seeking strict replication of rare variant associations, data sets must be quite large [48, 150].

The concept of strict replication can be applied to common and rare single variant analyses; however, for binned analyses the interpretation of replication is somewhat less clear. To illustrate, consider the following example: GeneA bin is significant after primary analysis. Researchers want to replicate the finding in a similar independent data set of adequate size for single variant associations. If there were seven variants in the GeneA bin, how many would have to be significant in single variant association tests to be considered a replication of the GeneA signal? If only one variant replicated, could this still be considered a replication of the GeneA signal? If the direction of effect changed for one variant compared to the original direction of GeneA, would this diminish support of replication? If a bin analysis was repeated and GeneA was significant with the same direction of effect, is this adequate evidence for replication? What if the significant GeneA bin in the replication set had only a subset of variants found in the original sample? Lastly, how could strict replication be interpreted if there were epistatic interactions between variants in the bin?

Replication for bin analyses do not the follow the same interpretation as single variant analyses. Variants with opposing directions of effect can (unfortunately) fall into the same bin. Completely neutral variants are quite commonly binned with variants that contribute to an association signal. Therefore, it would be illogical to require every variant in a

significant bin to have an individual significant association in a replication study. It would also be insensible to determine that every significantly associated signal must be the same direction of effect as the original associated bin. This would require the exact same variants to be found across multiple sequencing studies, no more and no less. This contradicts a major benefit of binning studies, allelic heterogeneity. The idea behind rare variant binning analyses is not that a single variant causes every instance of a studied phenotype, it is that a collection or group of variants occurring in a functionally similar manner can manifest as similar or the same phenotype. Secondly, variants contributing only noise to the original signal should not necessarily be present again in the replicating signal. There may be new neutral variants or only a subset of the original neutral variants in the replication results. This change does not affect the interpretation of the replication.

Another complicating factor is linkage disequilibrium. Some variants in bins will be a part of rare haplotypes in the same bin. A single bin could contain multiple haplotypes and/or independent variants. Therefore, the application of signal replication can not be the gold standard in binning results replication.

One last definition for replication could be considered in low frequency bin analyses, functional replication. Essentially, what evidence can be found in a second independent sample to support replication that results in a similar functional outcome? In the example above, GeneA was significant in the original data set and researchers were seeking statistical replication of this signal. In a replication set, any subset or expanded set of loci in GeneA with a significant association supports functional replication. This could be from strict replication from single variant associations, signal replication from haplotypes found within the bin, or from bin replication. One particular challenge of functional replication is data quality and interpretation. For example, functional replication of a single low frequency variant requires larger sample sizes. Functional replication of a bin analysis requires a cohort of similar ancestry, careful phenotyping, and adequate sequencing (depth of coverage, consistent technology, etc). These are challenging for a research team, but are becoming more common in the literature.

As larger data sets become available, replication will be important to validate binning studies which are currently underpowered and potentially suffer from "winner's curse."

While strict validation is most likely the gold standard (second only to molecular validation), it is important to consider the context of binning analyses. Binning analyses focus on functional units of information, e.g. low frequency variation in GeneA leads to phenotype X. Variation within a bin can be neutral and variants with opposite effects can be binned together. Follow-up analyses can include: single variant associations in large data sets, haplotype analyses in large data sets, further binning analyses, or biological replication in in vitro or in vivo systems. Each of these provide useful evidence for functional replication.

Future improvements to BioBin

Binning

Accurately and precisely binning variants that contribute to a signal, while minimizing the inclusion of variants that only contribute noise, will increase the likelihood of detecting an association. This is currently performed using filtering methods, such as removing variants found in publicly available databases collected from "healthy" individuals or binning only nonsynonymous or predicted damaging variants. Other methods use "step-up" approaches or "sliding windows" to intermediately test the strength of the association signal, and prune out variants that do not positively impact the strength of association [49, 54].

Improved binning should be more stochastic in nature, less reliant on databases of "healthy" individuals and ultimately not limited to coding regions. A future step could include using machine learning techniques such as random forest or evaporative cooling implementations to prioritize variants that are most likely to contribute to an association signal. Machine learning algorithms can guide the user to rank and ultimately select which variants in a bin to keep and which should be dropped before performing a statistical test [127]. The obvious potential hurdle is the computational resources needed to perform this analysis, and deciding whether it should be performed on all low frequency variants or separately for each bin of variants. Likely, the latter will be more informative since a single variant can affect genes and pathways differently. Future low frequency variant analysis

pipelines should creatively consider new methods to filter in/out variants to build the most powerful association signal.

Novel statistical tests

Lastly, new statistical tests are needed to broaden the options available for low frequency variant testing. Currently, if successful filters are applied and variants contributing to the signal are in the same direction, burden tests are most powerful. If variants have different directions of effect or varying effect sizes, nonburden tests are more powerful. Some published statistical tests are designed for case-control studies, others do not allow for covariate adjustment. There is a need for unbiased testing to critically evaluate the effectiveness of each test on multiple versions of simulated data. It is unlikely one test is superior in every scenario. In addition, new statistical tests are needed to better evaluate small sample sizes and ignore variants that contribute to noise. Given the variety of tests available and lack of analytical standards, users should carefully test type I error and power in simulations for new statistical methods.

Future of rare variant analyses

Data integration and complex modeling

Although low frequency variants have frequently been published to explain additional heritability and resolve loci for Mendelian traits, low frequency variants most likely do not act independently in the genomic variation spectrum. Similar to the era of GWAS, after single variant association testing did not explain all of the heritability of a trait, researchers began to search for epistatic interactions and build more complex models in an effort to glean more information from the data. The same will be done in next-generation sequencing data. Once the novelty of significant low frequency variant associations has been exhausted, researchers will reconsider the complexity of biology and variation in the genome. Methods for rare variant analysis should be easily adaptable to data integration techniques.

Rare variant interactions

Statistical interactions between rare variant hits are also likely to contribute to heritability. In low frequency analyses, the user often focuses on a genomic unit, usually a gene. If binned together, epistatic interactions between rare variants can be captured within the same bin (in many burden and nonburden tests). However, low frequency variants between genes in a pathway and/or other types of "-omic" data can act in a non-additive manner to affect the trait or phenotype of interest. Additional work should be done to implement a pipeline to search for such interactions.

Combining common and rare variants

Currently, a few methods allow users to combine common and rare variants for analysis [21, 37, 49]. Models can be built to expand bins to include both common and rare variants or create rare variant bins and combine with common single variant analyses [5]. Similarly, algorithms such as Lasso could be used to select potentially important parameters in a model [127]. Further work should be done in developing these pipelines since it is likely that rare and common variants act in concert to affect or potentially cause a phenotype.

"Omic" modules

The most important future direction for low frequency variant analyses is data integration. Data integration is the structural foundation needed to combine common and rare variants, test interactions, and perform many other analyses that more closely resemble the true mechanisms of biology. Biofilter, a tool developed in the Ritchie lab is currently being expanded to build "omic" modules, an algorithm that will enable multiple types of genetic data to be integrated. For example, if a user has methylome data, transcriptome data, sequence data, and environmental variables, the extension of Biofilter will be able to build modules based on relationships in the data. Consider the tumor suppressor gene *CDH4* as an example, the center of the module is established by the DNA genomic location of the gene, *CDH4* is located at chr20:60,074,477-60,515,673 [71]. All other links between data and *CDH4* will map back to these coordinates. Links would be established to include

common variants and low frequency variant bins in or nearby *CDH4*. Biofilter would map relevant data to the root of the module (genomic location), i.e. hypermethlyation patterns on the CpG island overlapping *CDH4* promoter [151], pertinent expression data (for *CDH4* and known regulatory elements), and environmental variables that relate to function in this region.

The *CDH4* "omic" module describes the data integration component of future analyses. The next step will be to use these relationships to build testable models with machine learning or regression techniques. To build models, one must consider relationships within a single "omic" module or alternatively expand the search space using a variety of LOKI sources. For example, a user might want to create pairwise interaction models between all of the elements in the *CDH4* module as well as models between all of the omic modules involved in the cell adhesion molecule pathway [62]. This could be accomplished using neural networks that will initially include all of the data in the relevant "omic" modules, build multiple neural networks, and evaluate phenotype prediction power from generated models [152]. This data integration/machine learning pipeline is advantageous because it reduces the search space for model creation, allows for nonlinear models to be built, and better estimates the complexity of biology.

Summary

BioBin is a novel method to collapse low frequency variants based on biological knowledge. There are many available options in BioBin that can be configured to answer a variety of scientific questions. Although, there are still challenges in the field of genomic research, particularly with regard to sequence data consistency and quality, there are exciting opportunities to use collapsing methods such as BioBin to explore more complex biological process using data integration and advanced modeling.

BIBLIOGRAPHY

[1] Daniel L. Hartl and Andrew G. Clark. *Principles of Population Genetics, Fourth Edition.* Sinauer Associates, Inc., 4th edition, December 2006.

[2] T.A. Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, J.H. Cho, A.E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C.N. Rotimi, M. Slatkin, D. Valle, A.S. Whittemore, M. Boehnke, A.G. Clark, E.E. Eichler, G. Gibson, J.L. Haines, T.F. Mackay, S.A. McCarroll, and P.M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

[3] Elizabeth T. Cirulli and David B. Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet*, 11(6):415–425, June 2010.

[4] Matthew R. Nelson, Daniel Wegmann, Margaret G. Ehm, Darren Kessner, Pamela St Jean, Claudio Verzilli, Judong Shen, Zhengzheng Tang, Silviu-Alin Bacanu, Dana Fraser, Liling Warren, Jennifer Aponte, Matthew Zawistowski, Xiao Liu, Hao Zhang, Yong Zhang, Jun Li, Yun Li, Li Li, Peter Woollard, Simon Topp, Matthew D. Hall, Keith Nangle, Jun Wang, Gonalo Abecasis, Lon R. Cardon, Sebastian Zllner, John C. Whittaker, Stephanie L. Chissoe, John Novembre, and Vincent Mooser. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104, July 2012.

[5] John S Witte. Rare genetic variants and treatment response: sample size and analysis issues. *Statistics in medicine*, 31(25):3041–3050, November 2012. PMID: 22736504.

[6] Ron Do, Sekar Kathiresan, and Gonalo R Abecasis. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Human molecular genetics*, 21(R1):R1–9, October 2012. PMID: 22983955.

[7] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, NHLBI GO Exome Sequencing ProjectESP Lung Project Team, David C Christiani, Mark M Wurfel, and Xihong Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American journal of human genetics*, 91(2):224–237, August 2012. PMID: 22863193.

[8] Carrie C. Buchanan, John R. Wallace, Alex T. Frase, Eric S. Torstenson, Sarah A. Pendergrass, and Marylyn D. Ritchie. A biologically informed method for detecting associations with rare variants. In Mario Giacobini, Leonardo Vanneschi, and William S. Bush, editors, *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, number 7246 in Lecture Notes in Computer Science, pages 201–210. Springer Berlin Heidelberg, January 2012.

[9] Carrie B Moore, John R Wallace, Alex T Frase, Sarah A Pendergrass, and Marylyn D Ritchie. BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. *BMC Medical Genomics*, 6(Suppl 2):S6, May 2013. PMID: 23819467 PMCID: PMC3654874.

[10] Carrie B. Moore, John R. Wallace, Alex T. Frase, Sarah A Pendergrass, and Marylyn D. Ritchie. Using BioBin to explore rare variant population stratification. *Pacific Symposium on Biocomputing.*, 2013.

[11] Richard G. H. Cotton. Communicating "mutation:" modern meanings and connotations. *Human Mutation*, 19(1):2, January 2002.

[12] Charles Darwin. *On the Origin of the Species by Natural Selection.* Murray, 1859.

[13] J.G. Mendel. Experiments in plant hybridization. *Verhandlungen des naturforschenden Vereines in Brnn*, Bd. IV fr das Jahr:Abhandlungen:347, 1866.

[14] W. Bateson. *Mendel's Principles of Heredity.* Cambridge University Press, Cambridge, 1909.

[15] ArchibaldE. Garrod. THE INCIDENCE OF ALKAPTONURIA : A STUDY IN CHEMICAL INDIVIDUALITY. *The Lancet*, 160(4137):1616–1620, December 1902.

[16] E.E. Eichler, J. Flint, G. Gibson, A. Kong, S.M. Leal, J.H. Moore, and J.H. Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*, 11(6):446–450, June 2010.

[17] L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, and T.A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc.Natl.Acad.Sci.U.S.A*, 106(23):9362–9367, June 2009.

[18] L.A. Hindorff, J. MacArthur, J. Morales, H.A. Junkins, P.N. Hall, A.K. Klemm, and T.A. Manolio. Catalog of published genome-wide association studies.

[19] Shashaank Vattikuti, Juen Guo, and Carson C. Chow. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet*, 8(3):e1002637, March 2012.

[20] Ivan P. Gorlov, Olga Y. Gorlova, Shamil R. Sunyaev, Margaret R. Spitz, and Christopher I. Amos. Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 82(1):100–112, January 2008.

[21] B. Li and S.M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics*, 83(3):311–321, 2008.

[22] V. Bansal, O. Libiger, A. Torkamani, and N.J. Schork. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet*, 11(11):773–785, November 2010.

[23] D. Botstein and N. Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat.Genet.*, 33 Suppl:228–237, March 2003.

[24] Bat-sheva Kerem, Johanna M. Rommens, Janet A. Buchanan, Danuta Markiewicz, Tara K. Cox, Aravinda Chakravarti, Manuel Buchwald, and Lap-Chee Tsui. Identification of the cystic fibrosis gene: Genetic analysis. *Science*, 245(4922):1073–1080, September 1989. ArticleType: research-article / Full publication date: Sep. 8, 1989 / Copyright 1989 American Association for the Advancement of Science.

[25] Michael J. Bamshad, Sarah B. Ng, Abigail W. Bigham, Holly K. Tabor, Mary J. Emond, Deborah A. Nickerson, and Jay Shendure. Exome sequencing as a tool for mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755, November 2011.

[26] Jocelyn Kaiser. Affordable 'Exomes' fill gaps in a catalog of rare diseases. *Science*, 330(6006):903–903, November 2010. PMID: 21071642.

[27] G. Bhatia, V. Bansal, O. Harismendy, N.J. Schork, E.J. Topol, K. Frazer, and V. Bafna. A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS.Comput Biol*, 6(10):e1000954, 2010.

[28] C.T. Johansen, J. Wang, M.B. Lanktree, H. Cao, A.D. McIntyre, M.R. Ban, R.A. Martins, B.A. Kennedy, R.G. Hassell, M.E. Visser, S.M. Schwartz, B.F. Voight, R. Elosua, V. Salomaa, C.J. O'Donnell, G.M. Dallinga-Thie, S.S. Anand, S. Yusuf, M.W. Huff, S. Kathiresan, and R.A. Hegele. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet*, 42(8):684–687, 2010.

[29] T. Walsh, H. Shahin, T. Elkan-Miller, M.K. Lee, A.M. Thornton, W. Roeb, Rayyan A. Abu, S. Loulus, K.B. Avraham, M.C. King, and M. Kanaan. Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. *American Journal of Human Genetics*, 87(1):90–94, July 2010.

[30] Manuel A Rivas, Mlissa Beaudoin, Agnes Gardet, Christine Stevens, Yashoda Sharma, Clarence K Zhang, Gabrielle Boucher, Stephan Ripke, David Ellinghaus, Noel Burtt, Tim Fennell, Andrew Kirby, Anna Latiano, Philippe Goyette, Todd Green, Jonas

Halfvarson, Talin Haritunians, Joshua M Korn, Finny Kuruvilla, Caroline Lagac, Benjamin Neale, Ken Sin Lo, Phil Schumm, Leif Trkvist, National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC), United Kingdom Inflammatory Bowel Disease Genetics Consortium, International Inflammatory Bowel Disease Genetics Consortium, Marla C Dubinsky, Steven R Brant, Mark S Silverberg, Richard H Duerr, David Altshuler, Stacey Gabriel, Guillaume Lettre, Andre Franke, Mauro D'Amato, Dermot P B McGovern, Judy H Cho, John D Rioux, Ramnik J Xavier, and Mark J Daly. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics*, 43(11):1066–1073, November 2011. PMID: 21983784.

[31] Hilma Holm, Daniel F Gudbjartsson, Patrick Sulem, Gisli Masson, Hafdis Th Helgadottir, Carlo Zanon, Olafur Th Magnusson, Agnar Helgason, Jona Saemundsdottir, Arnaldur Gylfason, Hrafnhildur Stefansdottir, Solveig Gretarsdottir, Stefan E Matthiasson, Gu Mundur Thorgeirsson, Aslaug Jonasdottir, Asgeir Sigurdsson, Hreinn Stefansson, Thomas Werge, Thorunn Rafnar, Lambertus A Kiemeney, Babar Parvez, Raafia Muhammad, Dan M Roden, Dawood Darbar, Gudmar Thorleifsson, G Bragi Walters, Augustine Kong, Unnur Thorsteinsdottir, David O Arnar, and Kari Stefansson. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nature genetics*, 43(4):316–320, April 2011. PMID: 21378987.

[32] Gosia Trynka, Karen A Hunt, Nicholas A Bockett, Jihane Romanos, Vanisha Mistry, Agata Szperl, Sjoerd F Bakker, Maria Teresa Bardella, Leena Bhaw-Rosun, Gemma Castillejo, Emilio G de la Concha, Rodrigo Coutinho de Almeida, Kerith-Rae M Dias, Cleo C van Diemen, Patrick C A Dubois, Richard H Duerr, Sarah Edkins, Lude Franke, Karin Fransen, Javier Gutierrez, Graham A R Heap, Barbara Hrdlickova, Sarah Hunt, Leticia Plaza Izurieta, Valentina Izzo, Leo A B Joosten, Cordelia Langford, Maria Cristina Mazzilli, Charles A Mein, Vandana Midah, Mitja Mitrovic, Barbara Mora, Marinita Morelli, Sarah Nutland, Concepcin Nez, Suna Onengut-Gumuscu, Kerra Pearce, Mathieu Platteel, Isabel Polanco, Simon Potter, Carmen Ribes-Koninckx, Isis Ricao-Ponce, Stephen S Rich, Anna Rybak, Jos Luis Santiago,

Sabyasachi Senapati, Ajit Sood, Hania Szajewska, Riccardo Troncone, Jezabel Varad, Chris Wallace, Victorien M Wolters, Alexandra Zhernakova, Spanish Consortium on the Genetics of Coeliac Disease (CEGEC), PreventCD Study Group, Wellcome Trust Case Control Consortium (WTCCC), B K Thelma, Bozena Cukrowska, Elena Urcelay, Jose Ramon Bilbao, M Luisa Mearin, Donatella Barisani, Jeffrey C Barrett, Vincent Plagnol, Panos Deloukas, Cisca Wijmenga, and David A van Heel. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature genetics*, 43(12):1193–1201, December 2011. PMID: 22057235.

[33] Julius Gudmundsson, Patrick Sulem, Daniel F Gudbjartsson, Gisli Masson, Bjarni A Agnarsson, Kristrun R Benediktsdottir, Asgeir Sigurdsson, Olafur Th Magnusson, Sigurjon A Gudjonsson, Droplaug N Magnusdottir, Hrefna Johannsdottir, Hafdis Th Helgadottir, Simon N Stacey, Adalbjorg Jonasdottir, Stefania B Olafsdottir, Gudmar Thorleifsson, Jon G Jonasson, Laufey Tryggvadottir, Sebastian Navarrete, Fernando Fuertes, Brian T Helfand, Qiaoyan Hu, Irma E Csiki, Ioan N Mates, Viorel Jinga, Katja K H Aben, Inge M van Oort, Sita H Vermeulen, Jenny L Donovan, Freddy C Hamdy, Chi-Fai Ng, Peter K F Chiu, Kin-Mang Lau, Maggie C Y Ng, Jeffrey R Gulcher, Augustine Kong, William J Catalona, Jose I Mayordomo, Gudmundur V Einarsson, Rosa B Barkardottir, Eirikur Jonsson, Dana Mates, David E Neal, Lambertus A Kiemeney, Unnur Thorsteinsdottir, Thorunn Rafnar, and Kari Stefansson. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature genetics*, 44(12):1326–1329, December 2012. PMID: 23104005.

[34] Thorlakur Jonsson, Hreinn Stefansson, Stacy Steinberg, Ingileif Jonsdottir, Palmi V Jonsson, Jon Snaedal, Sigurbjorn Bjornsson, Johanna Huttenlocher, Allan I Levey, James J Lah, Dan Rujescu, Harald Hampel, Ina Giegling, Ole A Andreassen, Knut Engedal, Ingun Ulstein, Srdjan Djurovic, Carla Ibrahim-Verbaas, Albert Hofman, M Arfan Ikram, Cornelia M van Duijn, Unnur Thorsteinsdottir, Augustine Kong, and Kari Stefansson. Variant of TREM2 associated with the risk of alzheimer's disease. *The New England journal of medicine*, 368(2):107–116, January 2013. PMID:

23150908.

[35] Samuel P. Dickson, Kai Wang, Ian Krantz, Hakon Hakonarson, and David B. Goldstein. Rare variants create synthetic genome-wide associations. *PLoS Biol*, 8(1):e1000294, January 2010.

[36] Greg Gibson. Rare and common variants: twenty arguments. *Nature reviews. Genetics*, 13(2):135–145, February 2011. PMID: 22251874.

[37] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics*, 89(1):82–93, July 2011. PMID: 21737059.

[38] Molly Przeworski, Richard R. Hudson, and Anna Di Rienzo. Adjusting the focus on human variation. *Trends in Genetics*, 16(7):296–302, July 2000.

[39] Jacob A. Tennessen, Abigail W. Bigham, Timothy D. OConnor, Wenqing Fu, Eimear E. Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, Hyun Min Kang, Daniel Jordan, Suzanne M. Leal, Stacey Gabriel, Mark J. Rieder, Goncalo Abecasis, David Altshuler, Deborah A. Nickerson, Eric Boerwinkle, Shamil Sunyaev, Carlos D. Bustamante, Michael J. Bamshad, and Joshua M. Akey. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69, July 2012.

[40] Ferran Casals and Jaume Bertranpetit. Human genetic variation, shared and private. *Science*, 337(6090):39–40, July 2012.

[41] Philipp W. Messer. Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics*, 182(4):1219–1232, August 2009.

[42] Montgomery Slatkin and Bruce Rannala. E STIMATING a LLELE a GE. *Annual Review of Genomics and Human Genetics*, 1(1):225–249, September 2000.

[43] Naomi R. Wray, Shaun M. Purcell, and Peter M. Visscher. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biol*, 9(1):e1000579, January 2011.

[44] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat.Genet.*, 22(2):139–144, June 1999.

[45] Jonathan L. Haines and Margaret A. Pericak-Vance. *Genetic Analysis of Complex Disease*. Wiley-Liss, 2 edition, May 2006.

[46] Benjamin M. Neale and Pak C. Sham. The future of association studies: Gene-based analysis and replication. *The American Journal of Human Genetics*, 75(3):353–362, September 2004.

[47] Dianne Keen-Kim, Carol A. Mathews, Victor I. Reus, Thomas L. Lowe, Luis Diego Herrera, Cathy L. Budman, Varda Gross-Tsur, Ann E. Pulver, Ruth D. Bruun, Gerald Erenberg, Allan Naarden, Chiara Sabatti, and Nelson B. Freimer. Overrepresentation of rare variants in a specific ethnic group may confuse interpretation of association analyses. *Human Molecular Genetics*, 15(22):3324 –3328, November 2006.

[48] Sergey Nejentsev, Neil Walker, David Riches, Michael Egholm, and John A Todd. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science (New York, N.Y.)*, 324(5925):387–389, April 2009. PMID: 19264985.

[49] T.J. Hoffmann, N.J. Marini, and J.S. Witte. Comprehensive approach to analyzing rare genetic variants. *PLoS.One.*, 5(11):e13584, 2010.

[50] Soumya Raychaudhuri, Oleg Iartchouk, Kimberly Chin, Perciliz L Tan, Albert K Tai, Stephan Ripke, Sivakumar Gowrisankar, Soumya Vemuri, Kate Montgomery, Yi Yu, Robyn Reynolds, Donald J Zack, Betsy Campochiaro, Peter Campochiaro, Nicholas Katsanis, Mark J Daly, and Johanna M Seddon. A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. *Nat Genet*, 43(12):1232–1236, December 2011.

[51] Carmen Dering, Claudia Hemmelmann, Elizabeth Pugh, and Andreas Ziegler. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genetic Epidemiology*, 35(S1):S12–S17, January 2011.

[52] T.B. Haack, K. Danhauser, B. Haberberger, J. Hoser, V. Strecker, D. Boehm, G. Uziel, E. Lamantea, F. Invernizzi, J. Poulton, B. Rolinski, A. Iuso, S. Biskup, T. Schmidt, H.W. Mewes, I. Wittig, T. Meitinger, M. Zeviani, and H. Prokisch. Exome sequencing identifies ACAD9 mutations as a cause of complex i deficiency. *Nat Genet*, 42(12):1131–1134, December 2010.

[53] S. Morgenthaler and W.G. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res*, 615(1-2):28–56, February 2007.

[54] S. Basu and W. Pan. Comparison of statistical tests for association with rare variants, November 2010.

[55] B.E. Madsen and S.R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS.Genet*, 5(2):e1000384, February 2009.

[56] A.L. Price, G.V. Kryukov, P.I. de Bakker, S.M. Purcell, J. Staples, L.J. Wei, and S.R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics*, 86(6):832–838, June 2010.

[57] Hongyan Fang, Bo Hou, Qi Wang, and Yaning Yang. Rare variants analysis by risk-based variable-threshold method. *Computational Biology and Chemistry*, 46:32–38, October 2013.

[58] F. Han and W. Pan. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered*, 70(1):42–54, 2010.

[59] Benjamin M. Neale, Manuel A. Rivas, Benjamin F. Voight, David Altshuler, Bernie Devlin, Marju Orho-Melander, Sekar Kathiresan, Shaun M. Purcell, Kathryn Roeder, and Mark J. Daly. Testing for an unusual distribution of rare variants. *PLoS Genet*, 7(3):e1001322, March 2011.

[60] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, November 2012.

[61] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39(Database):D38–D51, November 2010.

[62] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, November 2011.

[63] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Database):D691–D697, November 2010.

[64] E. C. Dimmer, R. P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M. J. Martin, B. Bely, P. Browne, W. Mun Chan, R. Eberhardt, M. Gardner, K. Laiho, D. Legge, M. Magrane, K. Pichler, D. Poggioli, H. Sehra, A. Auchincloss, K. Axelsen, M.-C. Blatter, E. Boutet, S. Braconi-Quintaje, L. Breuza, A. Bridge, E. Coudert, A. Estreicher, L. Famiglietti, S. Ferro-Rojas, M. Feuermann, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, J. James, S. Jimenez, F. Jungo, G. Keller, P. Lemercier, D. Lieberherr, P. Masson, M. Moinat, I. Pedruzzi, S. Poux, C. Rivoire, B. Roechert, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, L. Bougueleret, G. Argoud-Puy,

I. Cusin, P. Duek-Roggli, I. Xenarios, and R. Apweiler. The UniProt-GO annotation database in 2011. *Nucleic Acids Research*, 40(D1):D565–D570, November 2011.

[65] Marco Punta, Penny C. Coggill, Ruth Y. Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, Andreas Heger, Liisa Holm, Erik L. L. Sonnhammer, Sean R. Eddy, Alex Bateman, and Robert D. Finn. The pfam protein families database. *Nucleic Acids Research*, 40(D1):D290–D301, January 2012.

[66] Kumaran Kandasamy, S Sujatha Mohan, Rajesh Raju, Shivakumar Keerthikumar, Ghantasala S Sameer Kumar, Abhilash K Venugopal, Deepthi Telikicherla, J Daniel Navarro, Suresh Mathivanan, Christian Pecquet, Sashi Kanth Gollapudi, Sudhir Gopal Tattikota, Shyam Mohan, Hariprasad Padhukasahasram, Yashwanth Subbannayya, Renu Goel, Harrys K C Jacob, Jun Zhong, Raja Sekhar, Vishalakshi Nanjappa, Lavanya Balakrishnan, Roopashree Subbaiah, Y L Ramachandra, B Abdul Rahiman, T S Keshava Prasad, Jian-Xin Lin, Jon C D Houtman, Stephen Desiderio, Jean-Christophe Renauld, Stefan N Constantinescu, Osamu Ohara, Toshio Hirano, Masato Kubo, Sujay Singh, Purvesh Khatri, Sorin Draghici, Gary D Bader, Chris Sander, Warren J Leonard, and Akhilesh Pandey. NetPath: a public resource of curated signal transduction pathways. *Genome biology*, 11(1):R3, 2010. PMID: 20067622.

[67] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, Luisa Castagnoli, and Gianni Cesareni. MINT, the molecular interaction database: 2012 update. *Nucleic acids research*, 40(Database issue):D857–861, January 2012. PMID: 22096227.

[68] Chris Stark, Bobby-Joe Breitkreutz, Andrew Chatr-Aryamontri, Lorrie Boucher, Rose Oughtred, Michael S Livstone, Julie Nixon, Kimberly Van Auken, Xiaodong Wang, Xiaoqi Shi, Teresa Reguly, Jennifer M Rust, Andrew Winter, Kara Dolinski, and

Mike Tyers. The BioGRID interaction database: 2011 update. *Nucleic acids research*, 39(Database issue):D698–704, January 2011. PMID: 21071413.

[69] Ellen M McDonagh, Michelle Whirl-Carrillo, Yael Garten, Russ B Altman, and Teri E Klein. From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomarkers in medicine*, 5(6):795–806, December 2011. PMID: 22103613.

[70] O. L. Griffith, S. B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M. C. Sleumer, M. Bilenky, M. Haeussler, M. Griffith, S. M. Gallo, B. Giardine, B. Hooghe, P. Van Loo, E. Blanco, A. Ticoll, S. Lithwick, E. Portales-Casamar, I. J. Donaldson, G. Robertson, C. Wadelius, P. De Bleser, D. Vlieghe, M. S. Halfon, W. Wasserman, R. Hardison, C. M. Bergman, S. J.M. Jones, and The Open Regulatory Annotation Consortium. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Research*, 36(Database):D107–D113, December 2007.

[71] Pauline A. Fujita, Brooke Rhead, Ann S. Zweig, Angie S. Hinrichs, Donna Karolchik, Melissa S. Cline, Mary Goldman, Galt P. Barber, Hiram Clawson, Antonio Coelho, Mark Diekhans, Timothy R. Dreszer, Belinda M. Giardine, Rachel A. Harte, Jennifer Hillman-Jackson, Fan Hsu, Vanessa Kirkup, Robert M. Kuhn, Katrina Learned, Chin H. Li, Laurence R. Meyer, Andy Pohl, Brian J. Raney, Kate R. Rosenbloom, Kayla E. Smith, David Haussler, and W. James Kent. The UCSC genome browser database: update 2011. *Nucleic Acids Research*, October 2010.

[72] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, April 2010. PMID: 20354512.

[73] S.R. Sunyaev, W.C. Lathe III, V.E. Ramensky, and P. Bork. SNP frequencies in human genes. *Trends Genet.*, 16:335–337, 2000.

[74] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols*, 4(7):1073–1081, 2009. PMID: 19561590.

[75] William McLaren, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. Deriving the consequences of genomic variants with the ensembl API and SNP effect predictor. *Bioinformatics*, 26(16):2069–2070, August 2010. PMID: 20562413.

[76] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature reviews. Genetics*, 11(7):459–463, July 2010. PMID: 20548291.

[77] Mathieu Lemire. Defining rare variants by their frequencies in controls may increase type i error. *Nature Genetics*, 43(5):391–392, May 2011.

[78] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly, and P.C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575, 2007.

[79] F. Anthony San Lucas, Gao Wang, Paul Scheet, and Bo Peng. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics*, 28(3):421–422, February 2012. PMID: 22138362.

[80] Bo Peng, Christopher I Amos, and Marek Kimmel. Forward-time simulations of human populations with complex diseases. *PLoS genetics*, 3(3):e47, March 2007. PMID: 17381243.

[81] Biao Li, Gao Wang, and Suzanne M. Leal. SimRare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits. *Bioinformatics*, 28(20):2703–2704, October 2012. PMID: 22914216.

[82] G.V. Kryukov, A. Shpunt, J.A. Stamatoyannopoulos, and S.R. Sunyaev. Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci U S A*, 106(10):3871–3876, March 2009.

[83] Carrie B. Moore, John R. Wallace, Daniel J. Wolfe, Alex T. Frase, Sarah A. Pendergrass, Kenneth M. Weiss, and Marylyn D. Ritchie. Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data.

[84] SharonR. Grossman, KristianG. Andersen, Ilya Shlyakhter, Shervin Tabrizi, Sarah Winnicki, Angela Yen, DanielJ. Park, Dustin Griesemer, ElinorK. Karlsson, SunnyH. Wong, Moran Cabili, RichardA. Adegbola, RameshwarN.K. Bamezai, AdrianV.S. Hill, FredrikO. Vannberg, JohnL. Rinn, EricS. Lander, StephenF. Schaffner, and PardisC. Sabeti. Identifying recent adaptations in large-scale genomic data. *Cell*, 152(4):703–713, February 2013.

[85] Luis B. Barreiro, Guillaume Laval, Hlne Quach, Etienne Patin, and Llus Quintana-Murci. Natural selection has driven population differentiation in modern humans. *Nature Genetics*, 40(3):340, 2008.

[86] R Development Core Team. R: A language and environment for statistical computing, 2011.

[87] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[88] R.M. Durbin, G.R. Abecasis, D.L. Altshuler, A. Auton, L.D. Brooks, R.M. Durbin, R.A. Gibbs, M.E. Hurles, and G.A. McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010.

[89] Sharon R. Browning and Elizabeth A. Thompson. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, 190(4):1521–1531, April 2012. PMID: 22267498 PMCID: PMC3316661.

[90] Annick Joelle Nembot-Simo, Jinko Graham McNeney, and Brad. CrypticIBDcheck: identifying cryptic relatedness in genetic association studies, April 2012.

[91] Elisha D. O. Roberson and Jonathan Pevsner. Visualization of shared genomic regions and meiotic recombination in high-density SNP data. *PLoS ONE*, 4(8):e6711, August 2009.

[92] Gonalo R. Abecasis, Stacey S. Cherny, W. O. C. Cookson, and Lon R. Cardon. GRR: graphical representation of relationship errors. *Bioinformatics*, 17(8):742–743, August 2001.

[93] Scott H. Williamson, Ryan Hernandez, Adi Fledel-Alon, Lan Zhu, Rasmus Nielsen, and Carlos D. Bustamante. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences*, 102(22):7882–7887, May 2005.

[94] Benjamin F Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. A map of recent positive selection in the human genome. *PLoS Biol*, 4(3):e72, March 2006.

[95] Joseph K Pickrell, Graham Coop, John Novembre, Sridhar Kudaravalli, Jun Z Li, Devin Absher, Balaji S Srinivasan, Gregory S Barsh, Richard M Myers, Marcus W Feldman, and Jonathan K Pritchard. Signals of recent positive selection in a worldwide sample of human populations. *Genome research*, 19(5):826–837, May 2009. PMID: 19307593.

[96] David Lpez Herrez, Marc Bauchet, Kun Tang, Christoph Theunert, Irina Pugach, Jing Li, Madhusudan R Nandineni, Arnd Gross, Markus Scholz, and Mark Stoneking. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PloS one*, 4(11):e7888, 2009. PMID: 19924308.

[97] Bingshan Li and Suzanne M Leal. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS genetics*, 5(5):e1000481, May 2009. PMID: 19436704.

[98] Jeffrey A Rosenfeld, Christopher E Mason, and Todd M Smith. Limitations of the human reference genome for personalized genomics. *PloS one*, 7(7):e40294, 2012. PMID: 22811759.

[99] Ruiqiang Li, Yingrui Li, Hancheng Zheng, Ruibang Luo, Hongmei Zhu, Qibin Li, Wubin Qian, Yuanyuan Ren, Geng Tian, Jinxiang Li, Guangyu Zhou, Xuan Zhu, Honglong Wu, Junjie Qin, Xin Jin, Dongfang Li, Hongzhi Cao, Xueda Hu, Hlne Blanche, Howard Cann, Xiuqing Zhang, Songgang Li, Lars Bolund, Karsten Kristiansen, Huanming Yang, Jun Wang, and Jian Wang. Building the sequence map of the human pan-genome. *Nature Biotechnology*, 28(1):57–63, 2010.

[100] Jeffrey M. Kidd, Nick Sampas, Francesca Antonacci, Tina Graves, Robert Fulton, Hillary S. Hayden, Can Alkan, Maika Malig, Mario Ventura, Giuliana Giannuzzi, Joelle Kallicki, Paige Anderson, Anya Tsalenko, N. Alice Yamada, Peter Tsang, Rajinder Kaul, Richard K. Wilson, Laurakay Bruhn, and Evan E. Eichler. Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature Methods*, 7(5):365–371, 2010.

[101] Trevor J. Pemberton, Chaolong Wang, Jun Z. Li, and Noah A. Rosenberg. Inference of unexpected genetic relatedness among individuals in HapMap phase III. *American Journal of Human Genetics*, 87(4):457–464, October 2010. PMID: 20869033 PMCID: PMC2948801.

[102] Alan Hodgkinson and Adam Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, 12(11):756–766, November 2011.

[103] Hans Ellegren, Nick GC Smith, and Matthew T Webster. Mutation rate variation in the mammalian genome. *Current Opinion in Genetics & Development*, 13(6):562–568, December 2003.

[104] Sandra Beleza, Antnio M. Santos, Brian McEvoy, Isabel Alves, Cludia Martinho, Emily Cameron, Mark D. Shriver, Esteban J. Parra, and Jorge Rocha. The timing of pigmentation lightening in europeans. *Molecular Biology and Evolution*, 30(1):24–35, January 2013. PMID: 22923467.

[105] Nina G. Jablonski and George Chaplin. Human skin pigmentation, migration and disease susceptibility. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1590):785–792, March 2012. PMID: 22312045.

[106] Rebecca L. Lamason, Manzoor-Ali P. K. Mohideen, Jason R. Mest, Andrew C. Wong, Heather L. Norton, Michele C. Aros, Michael J. Jurynec, Xianyun Mao, Vanessa R. Humphreville, Jasper E. Humbert, Soniya Sinha, Jessica L. Moore, Pudur Jagadeeswaran, Wei Zhao, Gang Ning, Izabela Makalowska, Paul M. McKeigue, David O'Donnell, Rick Kittles, Esteban J. Parra, Nancy J. Mangini, David J. Grunwald, Mark D. Shriver, Victor A. Canfield, and Keith C. Cheng. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, 310(5755):1782–1786, December 2005. PMID: 16357253.

[107] ENCODE Project Consortium, Ian Dunham, Anshul Kundaje, Shelley F Aldred, Patrick J Collins, Carrie A Davis, Francis Doyle, Charles B Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R Lajoie, Stephen G Landt, Burn-Kyu Lee, Florencia Pauli, Kate R Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M Simon, Lingyun Song, Nathan D Trinklein, Robert C Altshuler, Ewan Birney, James B Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Ian Dunham, Jason Ernst, Terrence S Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C Hardison, Robert S Harris, Javier Herrero, Michael M Hoffman, Sowmya Iyer, Manolis Kelllis, Jainab Khatun, Pouya Kheradpour, Anshul Kundaje, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K Marinov, Angelika Merkel, Ali Mortazavi, Stephen C J Parker, Timothy E Reddy, Joel Rozowsky, Felix Schlesinger, Robert E Thurman, Jie Wang, Lucas D Ward, Troy W Whitfield, Steven P Wilder, Weisheng Wu, Hualin S Xi, Kevin Y Yip, Jiali Zhuang, Bradley E Bernstein, Ewan Birney, Ian Dunham, Eric D Green, Chris Gunter, Michael Snyder, Michael J Pazin, Rebecca F Lowdon, Laura A L Dillon, Leslie B Adams, Caroline J Kelly, Julia Zhang, Judith R Wexler, Eric D Green, Peter J Good, Elise A Feingold, Bradley E Bernstein, Ewan Birney, Gregory E Crawford, Job Dekker, Laura Elinitski, Peggy J Farnham, Mark Gerstein, Morgan C Giddings, Thomas R Gingeras,

Eric D Green, Roderic Guig, Ross C Hardison, Tomothy J Hubbard, Manolis Kellis, W James Kent, Jason D Lieb, Elliott H Margulies, Richard M Myers, Michael Snyder, John A Starnatoyannopoulos, Scott A Tennebaum, Zhiping Weng, Kevin P White, Barbara Wold, Jainab Khatun, Yanbao Yu, John Wrobel, Brian A Risk, Harsha P Gunawardena, Heather C Kuiper, Christopher W Maier, Ling Xie, Xian Chen, Morgan C Giddings, Bradley E Bernstein, Charles B Epstein, Noam Shoresh, Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J Coyne, Timothy Durham, Manching Ku, Thanh Truong, Lucas D Ward, Robert C Altshuler, Matthew L Eaton, Manolis Kellis, Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K Marinov, Jainab Khatun, Brian A Williams, Chris Zaleski, Joel Rozowsky, Maik Rder, Felix Kokocinski, Rehab F Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha P Gunawardena, Cdric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J Luo, Eddie Park, Jonathan B Preall, Kimberly Presaud, Paolo Ribeca, Brian A Risk, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandu, Lorain Schaeffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Timothy J Hubbard, Alexandre Reymond, Stylianos E Antonarakis, Gregory J Hannon, Morgan C Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guig, Thomas R Gingeras, Kate R Rosenbloom, Cricket A Sloan, Katrina Learned, Venkat S Malladi, Matthew C Wong, Galt P Barber, Melissa S Cline, Timothy R Dreszer, Steven G Heitner, Donna Karolchik, W James Kent, Vaness M Kirkup, Laurence R Meyer, Jeffrey C Long, Morgan Maddren, Brian J Raney, Terrence S Furey, Lingyun Song, Linda L Grasfeder,

Paul G Giresi, Bum-Kyu Lee, Anna Battenhouse, Nathan C Sheffield, Jeremy M Simon, Kimberly A Showers, Alexias Safi, Darin London, Akshay A Bhinge, Christopher Shestak, Matthew R Schaner, Seul Ki Kim, Zhuzhu Z Zhang, Piotr A Mieczkowski, Joanna O Mieczkowska, Zheng Liu, Ryan M McDaniell, Yunyun Ni, Naim U Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Ewan Birney, Vishwanath R Iyer, Jason D Lieb, Gregory E Crawford, Guoliang Li, Kljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Oscar J Luo, Atif Shahab, Melissa J Fullwood, Xiaoan Ruan, Yijun Ruan, Richard M Myers, Florencia Pauli, Brian A Williams, Jason Gertz, Georgi K Marinov, Timothy E Reddy, Jost Vielmetter, E Christopher Partridge, Diane Trout, Katherine E Varley, Clarke Gasper, Anita Bansal, Shirley Pepke, Preti Jain, Henry Amrhein, Kevin M Bowling, Michael Anaya, Marie K Cross, Brandon King, Michael A Muratet, Igor Antoshechkin, Kimberly M Newberry, Kenneth McCue, Amy S Nesmith, Katherine I Fisher-Aylor, Barbara Pusey, Gilberto DeSalvo, Stephanie L Parker, Sreeram Balasubramanian, Nicholas S Davis, Sarah K Meadows, Tracy Eggleston, Chris Gunter, J Scott Newberry, Shawn E Levy, Devin M Absher, Ali Mortazavi, Wing H Wong, Barbara Wold, Matthew J Blow, Axel Visel, Len A Pennachio, Laura Elnitski, Elliott H Margulies, Stephen C J Parker, Hanna M Petrykowska, Alexej Abyzov, Bronwen Aken, Daniel Barrell, Gemma Barson, Andrew Berry, Alexandra Bignell, Veronika Boychenko, Govanni Bussotti, Jacqueline Chrast, Claire Davidson, Thomas Derrien, Gloria Despacio-Reyes, Mark Diekhans, Iakes Ezkurdia, Adam Frankish, James Gilbert, Jose Manuel Gonzalez, Ed Griffiths, Rachel Harte, David A Hendrix, Cdric Howald, Toby Hunt, Irwin Jungreis, Mike Kay, Ekta Khurana, Felix Kokocinski, Jing Leng, Michael F Lin, Jane Loveland, Zhi Lu, Deepa Manthravadi, Marco Mariotti, Jonathan Mudge, Gaurab Mukherjee, Cedric Notredame, Baikang Pei, Jose Manuel Rodriguez, Gary Saunders, Andrea Sboner, Stephen Searle, Cristina Sisu, Catherine Snow, Charlie Steward, Andrea Tanzer, Electra Tapanari, Michael L Tress, Marijke J van Baren, Nathalie Walters, Stefan Washieti, Laurens Wilming, Amonida Zadissa, Zhang Zhengdong, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Mark Gerstein, Alexandre Raymond, Roderic Guig, Jennifer Harrow, Timothy J Hubbard,

Stephen G Landt, Seth Frietze, Alexej Abyzov, Nick Addleman, Roger P Alexander, Raymond K Auerbach, Suganthi Balasubramanian, Keith Bettinger, Nitin Bhardwaj, Alan P Boyle, Alina R Cao, Philip Cayting, Alexandra Charos, Yong Cheng, Chao Cheng, Catharine Eastman, Ghia Euskirchen, Joseph D Fleming, Fabian Grubert, Lukas Habegger, Manoj Hariharan, Arif Harmanci, Susma Iyenger, Victor X Jin, Konrad J Karczewski, Maya Kasowski, Phil Lacroute, Hugo Lam, Nathan Larnarre-Vincent, Jing Leng, Jin Lian, Marianne Lindahl-Allen, Renqiang Min, Benoit Miotto, Hannah Monahan, Zarmik Moqtaderi, Xinmeng J Mu, Henriette O'Geen, Zhengqing Ouyang, Dorrelyn Patacsil, Baikang Pei, Debasish Raha, Lucia Ramirez, Brian Reed, Joel Rozowsky, Andrea Sboner, Minyi Shi, Cristina Sisu, Teri Slifer, Heather Witt, Linfeng Wu, Xiaoqin Xu, Koon-Kiu Yan, Xinqiong Yang, Kevin Y Yip, Zhengdong Zhang, Kevin Struhl, Sherman M Weissman, Mark Gerstein, Peggy J Farnham, Michael Snyder, Scott A Tenebaum, Luiz O Penalva, Francis Doyle, Subhradip Karmakar, Stephen G Landt, Raj R Bhanvadia, Alina Choudhury, Marc Domanus, Lijia Ma, Jennifer Moran, Dorrelyn Patacsil, Teri Slifer, Alec Victorsen, Xinqiong Yang, Michael Snyder, Kevin P White, Thomas Auer, Lazaro Centarin, Michael Eichenlaub, Franziska Gruhl, Stephan Heerman, Burkard Hoeckendorf, Daigo Inoue, Tanja Kellner, Stephan Kirchmaier, Claudia Mueller, Robert Reinhardt, Lea Schertel, Stephanie Schneider, Rebecca Sinn, Beate Wittbrodt, Jochen Wittbrodt, Zhiping Weng, Troy W Whitfield, Jie Wang, Patrick J Collins, Shelley F Aldred, Nathan D Trinklein, E Christopher Partridge, Richard M Myers, Job Dekker, Gaurav Jain, Bryan R Lajoie, Amartya Sanyal, Gayathri Balasundaram, Daniel L Bates, Rachel Byron, Theresa K Canfield, Morgan J Diegel, Douglas Dunn, Abigail K Ebersol, Abigail K Ebersol, Tristan Frum, Kavita Garg, Erica Gist, R Scott Hansen, Lisa Boatman, Eric Haugen, Richard Humbert, Gaurav Jain, Audra K Johnson, Ericka M Johnson, Tattayana M Kutyavin, Bryan R Lajoie, Kristin Lee, Dimitra Lotakis, Matthew T Maurano, Shane J Neph, Fiedencio V Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Eric Rynes, Peter Sabo, Minerva E Sanchez, Richard S Sandstrom, Amartya Sanyal, Anthony O Shafer, Andrew B Stergachis, Sean Thomas, Robert E Thurman, Benjamin Vernot, Jeff Vierstra, Shinny Vong,

Hao Wang, Molly A Weaver, Yongqi Yan, Miaohua Zhang, Joshua A Akey, Michael Bender, Michael O Dorschner, Mark Groudine, Michael J MacCoss, Patrick Navas, George Stamatoyannopoulos, Rajinder Kaul, Job Dekker, John A Stamatoyannopoulos, Ian Dunham, Kathryn Beal, Alvis Brazma, Paul Flicek, Javier Herrero, Nathan Johnson, Damian Keefe, Margus Lukk, Nicholas M Luscombe, Daniel Sobral, Juan M Vaquerizas, Steven P Wilder, Serafim Batzoglou, Arend Sidow, Nadine Hussami, Sofia Kyriazopoulou-Panagiotopoulou, Max W Libbrecht, Marc A Schaub, Anshul Kundaje, Ross C Hardison, Webb Miller, Belinda Giardine, Robert S Harris, Weisheng Wu, Peter J Bickel, Balazs Banfai, Nathan P Boley, James B Brown, Haiyan Huang, Qunhua Li, Jingyi Jessica Li, William Stafford Noble, Jeffrey A Bilmes, Orion J Buske, Michael M Hoffman, Avinash O Sahu, Peter V Kharchenko, Peter J Park, Dannon Baker, James Taylor, Zhiping Weng, Sowmya Iyer, Xianjun Dong, Melissa Greven, Xinying Lin, Jie Wang, Hualin S Xi, Jiali Zhuang, Mark Gerstein, Roger P Alexander, Suganthi Balasubramanian, Chao Cheng, Arif Harmanci, Lucas Lochovsky, Renqiang Min, Xinmeng J Mu, Joel Rozowsky, Koon-Kiu Yan, Kevin Y Yip, and Ewan Birney. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. PMID: 22955616.

[108] Qianying Liu, Dan L. Nicolae, and Lin S. Chen. Marbled inflation from population structure in gene-based association studies with rare variants. *Genetic Epidemiology*, 37(3):286292, 2013.

[109] Hua He, Xue Zhang, Lili Ding, Tesfaye M Baye, Brad G Kurowski, and Lisa J Martin. Effect of population stratification analysis on false-positive rates for common and rare variants. *BMC Proceedings*, 5(Suppl 9):S116, 2011.

[110] Mark C. Hannibal, Kati J. Buckingham, Sarah B. Ng, Jeffrey E. Ming, Anita E. Beck, Margaret J. McMillin, Heidi I. Gildersleeve, Abigail W. Bigham, Holly K. Tabor, Heather C. Mefford, Joseph Cook, Koh-ichiro Yoshiura, Tadashi Matsumoto, Naomichi Matsumoto, Noriko Miyake, Hidefumi Tonoki, Kenji Naritomi, Tadashi Kaname, Toshiro Nagai, Hirofumi Ohashi, Kenji Kurosawa, Jia-Wei Hou, Tohru Ohta, Deshung Liang, Akira Sudo, Colleen A. Morris, Siddharth Banka, Graeme C. Black,

Jill Clayton-Smith, Deborah A. Nickerson, Elaine H. Zackai, Tamim H. Shaikh, Dian Donnai, Norio Niikawa, Jay Shendure, and Michael J. Bamshad. Spectrum of MLL2 (ALR) mutations in 110 cases of kabuki syndrome. *American journal of medical genetics. Part A*, 155(7):1511–1516, July 2011. PMID: 21671394 PMCID: PMC3121928.

[111] Sarah B. Ng, Abigail W. Bigham, Kati J. Buckingham, Mark C. Hannibal, Margaret J. McMillin, Heidi I. Gildersleeve, Anita E. Beck, Holly K. Tabor, Gregory M. Cooper, Heather C. Mefford, Choli Lee, Emily H. Turner, Joshua D. Smith, Mark J. Rieder, Koh-ichiro Yoshiura, Naomichi Matsumoto, Tohru Ohta, Norio Niikawa, Deborah A. Nickerson, Michael J. Bamshad, and Jay Shendure. Exome sequencing identifies MLL2 mutations as a cause of kabuki syndrome. *Nature Genetics*, 42(9):790–793, 2010.

[112] Radoje Drmanac, Andrew B Sparks, Matthew J Callow, Aaron L Halpern, Norman L Burns, Bahram G Kermani, Paolo Carnevali, Igor Nazarenko, Geoffrey B Nilsen, George Yeung, Fredrik Dahl, Andres Fernandez, Bryan Staker, Krishna P Pant, Jonathan Baccash, Adam P Borcherding, Anushka Brownley, Ryan Cedeno, Linsu Chen, Dan Chernikoff, Alex Cheung, Razvan Chirita, Benjamin Curson, Jessica C Ebert, Coleen R Hacker, Robert Hartlage, Brian Hauser, Steve Huang, Yuan Jiang, Vitali Karpinchyk, Mark Koenig, Calvin Kong, Tom Landers, Catherine Le, Jia Liu, Celeste E McBride, Matt Morenzoni, Robert E Morey, Karl Mutch, Helena Perazich, Kimberly Perry, Brock A Peters, Joe Peterson, Charit L Pethiyagoda, Kaliprasad Pothuraju, Claudia Richter, Abraham M Rosenbaum, Shaunak Roy, Jay Shafto, Uladzislau Sharanhovich, Karen W Shannon, Conrad G Sheppy, Michel Sun, Joseph V Thakuria, Anne Tran, Dylan Vu, Alexander Wait Zaranek, Xiaodi Wu, Snezana Drmanac, Arnold R Oliphant, William C Banyai, Bruce Martin, Dennis G Ballinger, George M Church, and Clifford A Reid. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science (New York, N.Y.)*, 327(5961):78–81, January 2010. PMID: 19892942.

[113] W.D. Dupont and W.D. Plummer. Power and sample size calculations. a review and computer program. *Control Clin.Trials*, 11(2):116–128, April 1990.

[114] Joy Armistead, Sunita Khatkar, Britta Meyer, Brian L Mark, Nehal Patel, Gail Coghlan, Ryan E Lamont, Shuangbo Liu, Jill Wiechert, Peter A Cattini, Peter Koetter, Klaus Wrogemann, Cheryl R Greenberg, Karl-Dieter Entian, Teresa Zelinski, and Barbara Triggs-Raine. Mutation of a gene essential for ribosome biogenesis, EMG1, causes bowen-conradi syndrome. *American journal of human genetics*, 84(6):728–739, June 2009. PMID: 19463982.

[115] Cystic fibrosis foundation - 2011 patient registry annual data report - 2011-patient-registry.pdf.

[116] C. Castellani, H. Cuppens, M. Macek, J.J. Cassiman, E. Kerem, P. Durie, E. Tullis, B.M. Assael, C. Bombieri, A. Brown, T. Casals, M. Claustres, G.R. Cutting, E. Dequeker, J. Dodge, I. Doull, P. Farrell, C. Ferec, E. Girodon, M. Johannesson, B. Kerem, M. Knowles, A. Munck, P.F. Pignatti, D. Radojkovic, P. Rizzotti, M. Schwarz, M. Stuhrmann, M. Tzetis, J. Zielenski, and J.S. Elborn. Consensus on the use and interpretation of cystic fibrosis mutation analysis in clinical practice. *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society*, 7(3):179–196, May 2008. PMID: 18456578 PMCID: PMC2810954.

[117] Zhanhai Li, Michael R Kosorok, Philip M Farrell, Anita Laxova, Susan E H West, Christopher G Green, Jannette Collins, Michael J Rock, and Mark L Splaingard. Longitudinal development of mucoid pseudomonas aeruginosa infection and lung disease progression in children with cystic fibrosis. *JAMA: the journal of the American Medical Association*, 293(5):581–588, February 2005. PMID: 15687313.

[118] Deanna M Green, J Michael Collaco, Kathryn E McDougal, Kathleen M Naughton, Scott M Blackman, and Garry R Cutting. Heritability of respiratory infection with pseudomonas aeruginosa in cystic fibrosis. *The Journal of pediatrics*, 161(2):290–295.e1, August 2012. PMID: 22364820.

[119] Mary J Emond, Tin Louie, Julia Emerson, Wei Zhao, Rasika A Mathias, Michael R Knowles, Fred A Wright, Mark J Rieder, Holly K Tabor, Deborah A Nickerson, Kathleen C Barnes, National Heart, Lung, and Blood Institute (NHLBI) GO Ex-

ome Sequencing Project, Lung GO, Ronald L Gibson, and Michael J Bamshad. Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic pseudomonas aeruginosa infection in cystic fibrosis. *Nature genetics*, 44(8):886–889, August 2012. PMID: 22772370.

[120] Andrew Gelman, Jennifer Hill, Yu-Sung Su, Masanao Yajima, and Maria Grazia Pittau. mi: Missing data imputation and model checking, September 2012.

[121] Peter Carl and Brian G. Peterson. PerformanceAnalytics: econometric tools for performance and risk analysis, January 2013.

[122] Stephen Turner. Getting genetics done: More on exploring correlations in r, August 2012.

[123] David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, March 1993.

[124] Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16):2409–2419, August 2002. PMID: 12210625.

[125] Georg Heinze. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in medicine*, 25(24):4216–4226, December 2006. PMID: 16955543.

[126] Georg Heinze, Meinhard Ploner, Daniela Dunkler (former versions), and Harry Southworth (former versions). logistf: Firth's bias reduced logistic regression, June 2013.

[127] T Hastie, R Tibshirani, and JH Friedman. *The Elements of Statistical Learning*. Springer, July 2003.

[128] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models, March 2013.

[129] Max Kuhn Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt Cooper, and Tony. caret: Classification and regression training, June 2013.

[130] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat.Genet.*, 38(8):904–909, 2006.

[131] Abhijeet Bakre, Lauren E. Andersen, Victoria Meliopoulos, Keegan Coleman, Xiuzhen Yan, Paula Brooks, Jackelyn Crabtree, S. Mark Tompkins, and Ralph A. Tripp. Identification of host kinase genes required for influenza virus replication and the regulatory role of MicroRNAs. *PLoS ONE*, 8(6):e66796, June 2013.

[132] Loubna Jouan, Simon L Girard, Sylvia Dobrzeniecka, Amirthagowri Ambalavanan, Marie-Odile Krebs, Ridha Joober, Julie Gauthier, Patrick A Dion, and Guy A Rouleau. Investigation of rare variants in LRP1, KPNA1, ALS2CL and ZNF480 genes in schizophrenia patients reflects genetic heterogeneity of the disease. *Behavioral and brain functions: BBF*, 9:9, 2013. PMID: 23425335.

[133] The NCBI handbook [internet], 2002.

[134] Florian Lang, Guido Henke, Hamdy Embark, Siegfried Waldegger, Monica Palmada, Christoph Böhmer, and Volker Vallon. Regulation of channels by the serum and glucocorticoid-inducible kinase - implications for transport, excitability and cell proliferation. *Cellular Physiology and Biochemistry*, 13(1):41–50, 2003.

[135] J Denry Sato, M Christine Chapline, Renee Thibodeau, Raymond A Frizzell, and Bruce A Stanton. Regulation of human cystic fibrosis transmembrane conductance regulator (CFTR) by serum- and glucocorticoid-inducible kinase (SGK1). *Cellular physiology and biochemistry: international journal of experimental cellular physiology, biochemistry, and pharmacology*, 20(1-4):91–98, 2007. PMID: 17595519.

[136] Yuemin Tian, Rainer Schreiber, and Karl Kunzelmann. Anoctamins are a family of ca2+-activated cl- channels. *Journal of cell science*, 125(Pt 21):4991–4998, November 2012. PMID: 22946059.

[137] Valerie Kuan, Adrian R. Martineau, Chris J. Griffiths, Elina Hyppnen, and Robert Walton. DHCR7 mutations linked to higher vitamin d status allowed early human migration to northern latitudes. *BMC Evolutionary Biology*, 13(1):144, July 2013. PMID: 23837623.

[138] Jill A. Holbrook, Gabriele Neu-Yilik, Matthias W. Hentze, and Andreas E. Kulozik. Nonsense-mediated decay approaches the clinic. *Nature Genetics*, 36(8):801–808, August 2004.

[139] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob. The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40(D1):D841–D846, November 2011.

[140] Volker Teichgrber, Martina Ulrich, Nicole Endlich, Joachim Riethmller, Barbara Wilker, Cheyla Conceio De OliveiraMunding, Anna M. van Heeckeren, Mark L. Barr, Gabriele von Krthy, Kurt W. Schmid, Michael Weller, Burkhard Tmmler, Florian Lang, Heike Grassme, Gerd Dring, and Erich Gulbins. Ceramide accumulation mediates inflammation, cell death and infection susceptibility in cystic fibrosis. *Nature Medicine*, 14(4):382–391, April 2008.

[141] Krishna P. Bhabak, Denny Proksch, Susanne Redmer, and Christoph Arenz. Novel fluorescent ceramide derivatives for probing ceramidase substrate specificity. *Bioorganic & Medicinal Chemistry*, 20(20):6154–6161, October 2012.

[142] Michael F. Holick. Vitamin d deficiency. *New England Journal of Medicine*, 357(3):266–281, 2007. PMID: 17634462.

[143] The cystic fibrosis neutrophil: A specialized yet potentially defective cell - springer.

[144] M Y Kker, K van Leeuwen, M de Boer, F Celmeli, A Metin, T T Ozgr, I Tezcan, O Sanal, and D Roos. Six different CYBA mutations including three novel mutations in ten families from turkey, resulting in autosomal recessive chronic granulomatous disease. *European journal of clinical investigation*, 39(4):311–319, April 2009. PMID: 19292887.

[145] Dong-Yan Shen, Yi-Hong Zhan, Qian-Ming Wang, Gang Rui, and Zhi-Ming Zhang. Oncogenic potential of cyclin kinase subunit-2 in cholangiocarcinoma. *Liver international: official journal of the International Association for the Study of the Liver*, 33(1):137–148, January 2013. PMID: 23121546.

[146] Cynthia Soderblom, Julia Stadler, Henri Jupille, Craig Blackstone, Oleg Shupliakov, and Michael C Hanna. Targeted disruption of the mast syndrome gene SPG21 in mice impairs hind limb function and alters axon branching in cultured cortical neurons. *Neurogenetics*, 11(4):369–378, October 2010. PMID: 20661613.

[147] Aaron K Wong, Christopher Y Park, Casey S Greene, Lars A Bongo, Yuanfang Guan, and Olga G Troyanskaya. IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic acids research*, 40(Web Server issue):W484–490, July 2012. PMID: 22684505.

[148] Andrew P Morris and Eleftheria Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, 34(2):188–193, February 2010. PMID: 19810025 PMCID: PMC2962811.

[149] William S. Bush and Jason H. Moore. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, December 2012.

[150] Lin T. Guey, Jasmina Kravic, Olle Melander, Nol P. Burtt, Jason M. Laramie, Valeriya Lyssenko, Anna Jonsson, Eero Lindholm, Tiinamaija Tuomi, Bo Isomaa, Peter Nilsson, Peter Almgren, Sekar Kathiresan, Leif Groop, Albert B. Seymour, David Altshuler, and Benjamin F. Voight. Power in the phenotypic extremes: a simulation

study of power in discovery and replication of rare variants. *Genetic Epidemiology*, 35(4):236246, 2011.

[151] Nathan D. VanderKraats, Jeffrey F. Hiken, Keith F. Decker, and John R. Edwards. Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. *Nucleic Acids Research*, June 2013. PMID: 23748561.

[152] Emily R Holzinger, Scott M Dudek, Alex T Frase, Ronald M Krauss, Marisa W Medina, and Marylyn D Ritchie. ATHENA: a tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 385–396, 2013. PMID: 23424143.