

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Genetics and Complex Disease Epidemiology in Diverse Populations

**Permalink**

<https://escholarship.org/uc/item/4mq1w04x>

**Author**

Drake, Katherine Anne

**Publication Date**

2012

Peer reviewed|Thesis/dissertation

Genetics and Complex Disease Epidemiology in Diverse Populations

by

Katherine Anne Drake

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Copyright 2012

by

Katherine Anne Drake

## ACKNOWLEDGEMENTS

Even though I'm the one coming away with the degree, I owe the fact that I got through grad school to a large group of people who have been extremely supportive and encouraging. Knowing that I had these wonderful people standing in my corner allowed me to finish.

I don't want to forget to thank the scientists that helped me get started at U of M in the Olsen lab. In particular, my undergraduate mentor, Laura Olsen, took me under her wing when I had no idea what I was doing or what I wanted to do. She has always been an inspiration and I consider myself lucky to still also count her as a friend.

My experience doing a MS at UC Berkeley was also hugely influential on my science and I owe that to a number of people there. First, Ira Tager took the time to mentor me while I was there and work with me on a project outside of his expertise because it was what I was interested in – I can't thank him enough for his encouragement, enthusiasm for mentoring, and for stressing the importance of always doing things the right way. Maureen Lahiff first encouraged me to take the leap into the biostatistically focused work that has been very enjoyable for me. Sandrine Dudoit and Eran Halperin both provided insight on a project that started for Sandrine's class and is now the fourth chapter of this dissertation. My biostatistics classmates at Berkeley- Karen McKeown, Jordan Brooks, Ashley Olson, and Elise Ruark- were a tremendous help to the start of the haplotype project, and they were also great company for dinner parties.

I would also like to thank the faculty who allowed me to rotate in their labs during my first year at UCSF – Jennifer Puck and Alan Wu. I appreciate the array of experience,

especially in the wetlab, that I gained in both of their labs and their patience with me as I figured out how to deal with mice, meconium, and all sorts of other strange laboratory things. My rotation experiences, along with my experience in the PSPG program as a whole, have given me a much bigger perspective than I would have gained only working in one lab.

PSPG wouldn't exist without our program director Deanna Kroetz and our program administrator Debbie Acoba. Both Deanna and Debbie have always been willing to help me figure things out, particularly when I went to do the MS in the middle of my PhD. In addition, Deanna has been particularly supportive and encouraging through my grad school struggles- I appreciate her willingness to hear me out at any time and her support of my career goals.

Steve Hamilton and John Witte were the members of my thesis committee from outside of my lab. I appreciate that both Steve and John gave me very insightful comments on my research and writing, and my science is better as a result.

Of course, I wouldn't be where I am today without the members and affiliates of the Burchard lab, both past and present. Each and every one of them has helped me both professionally and personally and I would like to thank all of them – you know who you are. In particular, Dara Torgerson has helped me immensely in the last couple of years, especially as I wrote my thesis. In addition to her scientific advice and her willingness to read the first draft of my entire thesis, I appreciate her listening to me and her pep talks and encouragement. I think I would have quit or might still be trying to get everything done without them. Josh Galanter, Marc Via, Melinda Aldrich, Sam Oh, Lindsey Roth, Chris Gignoux, and Laura Fejerman have all been wonderful friends who have all seen

me cry way too many times and are miraculously still always willing to listen. I also greatly appreciate the brainstorming sessions with all of these people and other members and affiliates of the lab. In particular, Josh and Chris helped me conceptualize the prediction project while at a conference in Denver. Saunak Sen was wonderfully patient and kind in discussing the finer points of how to execute the prediction project. In addition, although his data is not included here, Keoki Williams has kindly allowed me to visit his lab in Detroit several times in order to use his data for replication of the prediction project. The BDR GWAS took an army, of course, and I am indebted to everyone – the participants; the GALA II site PIs and recruiters; Sandra Salazar, Lindsey, Elizabeth Nguyen, and Celeste Eng for their fabulous management of the data, blood and genotyping; Chris, Josh, Scott Huntsman, and Donglei Hu for QCing the genotypes and running all of the ancestry estimation and imputation; and Dara for helping me organize my story and put it all together at the end. As I mentioned above, the haplotype project was largely aided by my professors and classmates at UC Berkeley. In addition, Chris, Josh, and Marc entertained many discussions about the project and both Elad Ziv and Ryan Hernandez took the time to meet with me about it at various times and help me see the bigger picture around the story. Finally, of course, the Burchard lab wouldn't be as it is without Esteban, my PI and the chair of my thesis committee. He has persevered and brought into his lab not only the massive amounts of data I used in my dissertation, but also the fabulous, smart and collaborative people in the lab. I wouldn't have been able to do this without either of those things.

Outside of school and the scientific community there, I have made many fabulous friends in the bay area who have kept me sane. Luckily, there are too many to name each

of them, but you know who you are. All my friends have all made me laugh, listened to me cry, and gone with me on ridiculous and exciting adventures– and seriously, I couldn't have done it without such a great support system. One of these friends, Jeremy Davis-Turak, also became my boyfriend in the later part of grad school, and I am so fortunate for that. He has been a great listener, wonderfully supportive of me as I decide and then decide again what I want to do, and even loans me his R code on occasion. I look forward to our continued adventures together and supporting him as finishes his PhD.

Finally, I only appreciate my family more and more as time passes. My brother is a wonderfully genuine person and it means so much that I know he will always support and love me no matter what I do. My relationship with my mom continues to grow stronger with every minute we spend together, and I can't get enough of that. Knowing that she's standing behind me no matter what has helped me immensely when I struggled over the last few years. My dad has always encouraged me to set high standards for myself and push myself to achieve them, and has also recognized that I need to be encouraged to relax sometimes, too. In addition, he's answered countless questions like 'Is there such a thing as too long for an acknowledgments section?' throughout my time in grad school. I wouldn't have been able to do this without the support of my family.

Thank you, to all of my colleagues, friends, and family who have been instrumental in my grad school experience.

# Genetics and Complex Disease Epidemiology in Diverse Populations

Katherine Anne Drake

## Abstract

Asthma, like other common diseases, has both genetic and environmental causes.

Understanding the heterogeneity in asthma, the genetics of associated traits, and how we introduce error by using certain methods are critical to determining the causes of asthma and other complex diseases.

We examined two ways to define asthma heterogeneity: using statistical clustering methods and using principal components analysis. We compared the fit of these variables and how well they predicted asthma exacerbations in data from 1,085 Latino and African American children with asthma. We found that principal components both fit the data better and predicted exacerbations better than cluster groups. These variables need to be compared to other known predictors of exacerbations.

In addition, we conducted a genome-wide association study and admixture mapping study of bronchodilator response (BDR) in 1,782 Latino children with asthma. Four of the genome-wide significant SNPs were promising rare variants. All four had good dose-response relationships with BDR and two were in promising candidate genes. Our admixture mapping found five regions where a specific ancestry was significantly associated with BDR. Since rare variants are often present on specific ancestral backgrounds, this result supports the hypothesis that rare variants are important for



BDR. Unfortunately, replication of individual rare variants is difficult. Future efforts should focus on sequencing the regions we identified to find other rare variants and better understand their function.

Finally, we compared the accuracy of haplotype inference error between four populations from HapMap Phase 3. We found that haplotype inference error was highest in the African populations, intermediate in the Mexican population, and lowest in the European population. In addition, some regions had higher haplotype inference error than others and this was not explained by several measured features of the regions. Comparisons between haplotype association studies across populations should account for possible differences in haplotype inference error between populations.

## TABLE OF CONTENTS

### **Preface**

|                   |      |
|-------------------|------|
| Copyright         | ii   |
| Acknowledgements  | iii  |
| Abstract          | vii  |
| Table of Contents | ix   |
| List of Tables    | xiii |
| List of Figures   | xiv  |

### **Chapter 1**

#### **Introduction to Asthma and the Genetics of Complex Human Disease**

|   |    |
|---|----|
| 1.1. Asthma and Bronchodilator Response                               | 1  |
| 1.2. Genetics of Complex Human Disease                                |    |
| 1.2.1. Introduction to Genetics of Complex Human Disease              | 5  |
| 1.2.2. Methodological Considerations                                  | 8  |
| 1.2.2.1. Subject Recruitment and Phenotype Measurement                | 8  |
| 1.2.2.2. Statistical Inference Methods for Genetic Data               | 9  |
| 1.2.2.3. Burden of Proof: Multiple Testing Correction and Replication | 11 |
| 1.2.3. Genetics of Asthma and Bronchodilator Response                 | 12 |
| 1.3. Summary of Chapters  | 14 |
| 1.4. References   | 15 |

## **Chapter 2**

### **A Continuous Asthma Phenotype Predicts Exacerbations Better than Cluster**

#### **Groups**

|  |    |
|--|----|
| 2.1. Introduction  | 21 |
| 2.2. Methods   |    |
| 2.2.1. AGES Population   | 23 |
| 2.2.2. Variable Selection  | 25 |
| 2.2.3. Primary Analysis  | 25 |
| 2.2.4. Prediction from Fewer Input Variables   | 29 |
| 2.2.5. Analysis of Imputed Data  | 29 |
| 2.3. Results   |    |
| 2.3.1. Principal Components 1 and 2 Represent a Continuous Asthma Spectrum   | 30 |
| 2.3.2. Three Kmeans Clusters and SARP Groups Represent Clinical Subgroups of<br>Asthma and are Related to Principal Components | 33 |
| 2.3.3. Principal Components Predict All Outcomes Better than Clusters  | 36 |
| 2.3.4. Analysis of Imputed Data Supports Original Results  | 36 |
| 2.4. Discussion  | 39 |
| 2.5. References  | 44 |

## **Chapter 3**

### **Genome-Wide Association Study and Admixture Mapping of Bronchodilator**

#### **Response Implicates Rare Variants**

|                   |    |
|-------------------|----|
| 3.1. Introduction | 46 |
|-------------------|----|

|  |    |
|--|----|
| 3.2. Methods   |    |
| 3.2.1. Discovery Population: GALA II   | 48 |
| 3.2.2. Replication Population: GALA I  | 50 |
| 3.2.3. Genotyping and Genetic Ancestry Estimation in GALA II                             | 51 |
| 3.2.4. Genotyping, Genetic Ancestry Estimation and Imputation in GALA I                  | 52 |
| 3.2.5. Analysis of Allelic Associations  | 53 |
| 3.2.6. Analysis of Ancestry Associations   | 55 |
| 3.3. Results   |    |
| 3.3.1. Rare Variants are Associated with BDR in GALA II                                  | 56 |
| 3.3.2. Admixture Mapping Supports the Association of Rare Variants and BDR in<br>GALA II | 61 |
| 3.4. Discussion  | 65 |
| 3.5. References  | 70 |

## **Chapter 4**

### **Haplotype Inference Error Varies Across Populations**

|  |    |
|--|----|
| 4.1. Introduction  | 73 |
| 4.2. Methods   |    |
| 4.2.1. Data  | 74 |
| 4.2.2. Haplotype Inference and Gold Standard             | 75 |
| 4.2.3. Calculation of Error Proportion                   | 76 |
| 4.2.4. Calculation of Factors Related to Inference Error | 76 |
| 4.2.5. Statistical Analysis                              | 76 |

|  |    |
|--|----|
| 4.3. Results   |    |
| 4.3.1. Haplotype Inference Error Varies by Population  | 78 |
| 4.3.2. Measurable Factors and Certain Regions are Associated with Haplotype Inference Error                  | 81 |
| 4.3.3. Most Previously Significant Regions are Associated with Error Even After Adjustment for Other Factors | 81 |
| 4.4. Discussion  | 84 |
| 4.5. References  | 87 |
| <br>   |    |
| <b>Appendix</b>  |    |
| Appendix A: Genotype Clusters for Six SNPs in GALA II  | 89 |

## LIST OF TABLES

### Chapter 2

|   |    |
|---|----|
| Table 2.1. Eligibility criteria for participation for AGES asthma cases and healthy controls                    | 24 |
| Table 2.2. Number of cases from participating study centers and institutions in the GALA II and SAGE II studies | 24 |
| Table 2.3. Input Variables Used for Principal Components, Kmeans Clusters, and Random Forest in AGES            | 26 |
| Table 2.4. AUCs for all outcomes and predictors with non-missing or imputed data                                | 36 |

### Chapter 3

|   |    |
|---|----|
| Table 3.1. Participating study centers and institutions in the GALA II study            | 48 |
| Table 3.2. Eligibility criteria for participation of asthma cases in GALA II and GALA I | 49 |
| Table 3.3. Genome-wide significant hits from allelic associations with BDR in GALA II   | 60 |
| Table 3.4. Significant Admixture Mapping Peaks  | 64 |
| Table 3.5. Summary of Top SNPs Identified in GWAS and Admixture Mapping                 | 65 |

### Chapter 4

|   |    |
|---|----|
| Table 4.1. Factors that may be associated with haplotype inference error      | 77 |
| Table 4.2. Significant Differences in Haplotype Inference Error by Population | 80 |

## LIST OF FIGURES

### Chapter 2

- Figure 2.1. PCs 1 & 2 are both heavily loaded by spirometry variables in different directions, but other contributing variables differ 32
- Figure 2.2. A) Kmeans clusters and B) SARP groups separate PCs 1 and 2 by slicing a continuous set of points into groups 34
- Figure 2.3. Clustergrams indicate that three kmeans clusters are the most stable in AGES. A) Average of PC1 in each cluster vs. number of clusters (k). B) Average of PC1 in each cluster vs. repeated kmeans clustering runs with k=3 to check the stability of points 35
- Figure 2.4. Principal components predict A) hospitalization, B) ER visits, and C) oral steroid use better than kmeans cluster or SARP groups 38
- Figure 2.5. Most variables mean and standard deviation are not biased but do not converge with imputation 39
- Figure 2.6. Loadings of PCs 1 & 2 after imputation differ from original PCs 1 & 2 40

### Chapter 3

- Figure 3.1. Admixture proportions for GALA II cases 52
- Figure 3.2. QQ plots for allelic associations with bronchodilator response (BDR) show that signal is driven by rare variants in A) all of GALA II, B) GALA II Puerto Ricans and C) GALA II Mexicans 57
- Figure 3.3. Genome-wide allelic associations with bronchodilator response in A) all GALA II, B) GALA II Puerto Ricans and C) GALA II Mexicans 59

|  |    |
|--|----|
| Figure 3.4. Genome-wide ancestry associations with bronchodilator response in all GALA II for A) African ancestry, B) Native American ancestry, and C) European ancestry           | 62 |
| Figure 3.5. Genome-wide ancestry associations with bronchodilator response in GALA II Puerto Ricans for A) African ancestry, B) Native American ancestry, and C) European ancestry | 63 |
| Figure 3.6. Genome-wide ancestry associations with bronchodilator response in GALA II Mexicans for A) African ancestry, B) Native American ancestry, and C) European ancestry      | 64 |
| <br><b>Chapter 4</b>   |    |
| Figure 4.1. Haplotype Inference Error Varies by Population   | 79 |
| Figure 4.2. Seven measured factors, two populations, and 18 regions are associated with error  | 80 |
| Figure 4.3. Trends in p-values for significant regions after adjusting for increasing numbers of PCs for A) Phase, B) Beagle, C) Shape-IT and D) fastPhase                         | 82 |
| Figure 4.4. Most Regions Remain Associated with Error After Adjusting for Other Measurable Factors   | 83 |
| Fig 4.5. PCs 1, 2 and 3 are driven by number of SNPs / between-SNP distance, LD / MAF / number of heterozygous SNPs and genic region, respectively                                 | 84 |



# CHAPTER 1

## INTRODUCTION TO ASTHMA AND THE GENETICS OF COMPLEX HUMAN DISEASE

### 1.1. Asthma and Bronchodilator Response

Asthma is a common but complex respiratory disease with heterogeneous clinical expression. Asthma is defined as recurrent, reversible airway obstruction. However, the expression of this airway obstruction varies drastically between individuals. Between 2006 and 2008, asthma affected 7.8% of the US population<sup>1</sup>. Although the death rate from asthma is not high overall, both prevalence and mortality are higher in certain ethnic and racial groups<sup>2-4</sup>. Puerto Ricans have the highest prevalence of asthma in the US, followed by African Americans. Mexicans have the lowest prevalence, followed by Dominicans. Trends in mortality between populations follow the same pattern as trends in prevalence. The cause of these differences in prevalence and mortality between racial and ethnic groups is unknown but is likely related to differences in environmental and genetic factors that contribute to asthma susceptibility and the interaction between these factors.

Asthma is caused by both genetic and environmental factors. The way in which these causal factors work together to cause asthma and asthma disparities is not well understood. One theory is that a specific combination of genetic and environmental factors is required to produce asthma<sup>5</sup>. These gene-environment interactions may contribute to asthma disparities, as the frequency of both genetic factors and environmental exposures vary between racial and ethnic groups<sup>6</sup>. Thus, if a particular

group has a high prevalence of exposures and genetic variants that interact to cause asthma, the prevalence of asthma will be higher in this group.

Several gene-environment interactions that may cause asthma have been discovered in the last decade. Many of these interactions involve innate immunity genes that have biological interactions with environmental stimuli<sup>5</sup>. For example, the interaction of the *CD14* C-159T allele and levels of endotoxin exposure has been described in several studies<sup>7,8</sup>. In most studies, the T allele is associated with protection from asthma if endotoxin exposure is low. In contrast, the C allele is associated with protection from asthma if endotoxin exposure is high. The *CD14*-endotoxin exposure interaction is not yet fully understood, but is a promising example that has supporting evidence from several studies.

Unfortunately, gene-environment interactions are not easy to identify, partly because their genetic and environmental components are not always significant alone. Testing the association of these components in a subset of asthmatics may help identify them. In particular, if there are known environmental risk factors for a subset of asthmatics we might identify candidates for gene-environment interactions by looking for genetic risk factors in that subset. This was recently demonstrated in a meta-analysis of *CD14*<sup>9</sup>. In this meta-analysis, *CD14* was associated with asthma only in the subset of atopic asthmatics with other allergic disease. Thus, using a subset of atopic asthma patients may have helped to more clearly identify *CD14* as a risk factor in earlier studies. Then, researchers could have looked for gene-environment interactions with known environmental risk factors for atopy. One difficulty in using this type of approach is how to choose the subset of asthmatics to study. In order to use subsets of asthma patients to

identify gene-environment interactions and their components, we must better understand the heterogeneity of asthma itself.

Atopic and non-atopic asthma is probably the most common way that asthma heterogeneity has been described. Triggers, response to certain therapies, onset and severity are just a few of the other variables that vary widely among individuals with asthma<sup>10</sup>. Many overlapping subsets of asthma, or so-called asthma phenotypes, have been defined using each of these variables. As demonstrated in the *CD14* meta-analysis described above, the causes of asthma may vary depending on the asthma phenotype<sup>9</sup>.

Many studies of asthma restrict their analysis to specific asthma phenotypes, but the range of phenotypes used in these studies varies substantially<sup>11-13</sup>. Asthma phenotypes have been defined both based on expert opinion and more recently using statistical clustering methods. Statistical clustering methods split the individuals into a researcher-defined number of groups based on any number of input variables. Defining asthma phenotypes called ‘cluster groups’ using statistical clustering methods is appealing because it is more objective than expert opinion. It is still unclear which phenotypes should be used to search for risk factors. One suggestion is that asthma phenotypes might be defined based on their relevance to clinically meaningful outcomes like exacerbations and response to therapy<sup>10</sup>. However, no studies of cluster groups have examined their ability to predict these clinically meaningful outcomes as compared with other asthma phenotypes.

The ability to predict clinically meaningful outcomes should be an important factor in defining useful asthma phenotypes. Well-defined asthma phenotypes will let researchers identify genetic and environmental causes of each asthma phenotype without

having the results ‘washed out’ by noise from the other phenotypes. Understanding these causes, in turn, will lead to a better understanding of asthma physiology and an ability to develop treatments specific to each phenotype. Thus, researchers need a better understanding of these asthma phenotypes in order to fully understand the genetic and environmental causes of asthma.

Understanding asthma-related traits like bronchodilator response (BDR) to  $\beta_2$ -adrenergic receptor ( $\beta_2$ AR) agonists may also help us understand asthma and its causes. Although other bronchodilators have been used historically,  $\beta_2$ AR agonists are now the primary rescue medication for individuals having an asthma attack<sup>14</sup>. However, bronchodilator response, like asthma, varies between individuals and populations. Generally, individuals with lower baseline lung function have higher BDR because they have more room to improve than individuals with higher baseline lung function. However, in one study, Puerto Ricans had lower BDR than African Americans or Mexicans, despite having lower baseline lung function<sup>15,16</sup>. In another study, African Americans had lower BDR than white patients<sup>17</sup>. Since bronchodilators are the primary rescue medication for asthma, low BDR may lead to increased asthma severity and mortality. Thus, understanding the genetic and environmental contributions to secondary phenotypes like BDR is crucial to understanding and treating asthma.

Recently, technology for assaying genetic variation has improved drastically. In addition, recent studies on the genetics of asthma are collecting more environmental data. These facts will allow researchers to uncover more about the genetic and environmental causes of asthma and lead to a better understanding of asthma etiology.

## **1.2. Genetics of Complex Human Diseases**

### **1.2.1. Introduction to Genetics of Complex Human Disease**

In the past decade, there has been an explosion in genome-wide association studies (GWAS) that agnostically scan the genome for variation associated with complex diseases<sup>18,19</sup>. GWAS address the ‘common disease-common variant’ hypothesis, which suggests that a combination of several common genetic variants causes disease<sup>20</sup>. Most of these variants are expected to have modest effects. In contrast, linkage studies, which were available before GWAS and also agnostically scanned the genome, have low power to identify these types of associations<sup>21</sup>. Linkage studies successfully identified several variants associated with Mendelian disorders but have had limited success in identifying important variants for common disease<sup>20</sup>. Agnostic scans of the genome like GWAS and linkage studies are important because our understanding of the biological function of the human genome is still very limited. GWAS have very quickly identified hundreds of common variants associated with many common diseases<sup>22,23</sup>.

GWAS have become possible because of massive technological and scientific advances that allowed for the development of commercial arrays that ‘tag’ most of the common variation in the human genome. These arrays genotype ‘tag SNPs’ that are in linkage disequilibrium (LD) with many other SNPs in the genome<sup>24</sup>. SNPs in LD with each other are frequently transmitted from one generation to another together. Thus, a tag SNP can be used as a proxy for other SNPs in association tests. Consortia like the International HapMap Consortium and 1000 Genomes Consortium have identified patterns in genetic variation and linkage disequilibrium across populations<sup>25,26</sup>. These

consortia, along with advances in genotyping technology, were required for GWAS and related studies of haplotypes and imputed SNPs to become possible.

Although GWAS have successfully identified many previously unknown genes and regions associated with disease, it has become clear that these associations do not explain all of the heritability of complex diseases<sup>27</sup>. Recently, researchers have hypothesized that rare variation may play an important causal role in complex disease<sup>20,22</sup>. In fact, significant associations with common variants identified in GWAS might be driven by causal rare variants<sup>28</sup>. Thus, research to identify heritable factors associated with disease continues.

Much of the recent research has focused on epigenetics, structural variation like copy number variants, and rare variation. Sequencing-based approaches can identify these types of genetic variation. However, whole-genome sequencing is still cost-prohibitive in many large studies. To allow the study of rare variation without sequencing, some SNP arrays now include coverage of more rare variation than in the past<sup>29</sup>. In addition, admixture mapping can identify regions where variation that is likely to be rare is associated with disease.

Admixture mapping is a technique for identifying regions of the genome containing risk alleles that differ in frequency between the ancestral populations of a population with recent mixed ancestry<sup>e.g. 30</sup>. It can identify all types of variation specific to an ancestral background that are tagged by being in LD with a specific ancestry. Since rare variants are likely to be specific to an ancestral background<sup>31</sup>, admixture mapping is likely to pick up signals from rare variants on these backgrounds. Although admixture

mapping doesn't identify specific causal variants, sequencing candidate regions discovered by admixture mapping can be used to identify the causal variants.

Most researchers agree that like the common variants identified in GWAS, rare variants alone are unlikely to explain all of the heritability of a complex disease. Techniques like GWAS and admixture mapping should be used as tools to identify regions that contain clues to the heritability of a disease. Since these approaches identify different types of 'clues,' they are complimentary. Researchers should combine data from all approaches in order to understand the way that each disease is inherited and the likely sources of genetic variation that contribute to each disease.

Researchers should also continue to use innovative approaches to gather more information on the genetics of complex diseases. One limitation of genome-wide association studies is that they generally do not identify causal genetic variants<sup>22</sup>. Sequencing and experiments that help us understand the biological function of genes, regions, and the variants within them will be necessary to understand which variants are causal. Another limitation of many previous genome-wide association studies is that they do not take into account environmental exposures and their interaction with genetic factors<sup>20</sup>. Most complex diseases have genetic and environmental components, and understanding the way these factors work together is also necessary to understand the causes of complex disease. However, given the scope of variation in the human genome and our lack of understanding of its function, identifying causal variants and their interactions with the environment would be impossible without first identifying regions of importance through GWAS, admixture mapping and similar studies. Methodologically sound, genome-wide studies of well-characterized samples continue to be important

sources of information about the regions in the genome that may be associated with disease.

### **1.2.2. Methodological Considerations**

Before the GWAS era, published genetic studies of complex disease often produced inconsistent or irreproducible results<sup>23</sup>. There was a general lack of consistency in methods across these studies. The availability of genome-wide data has brought with it many new methods that are now used much more consistently across studies<sup>e.g. 32-35</sup>. In addition, good reviews of methodological considerations necessary for genetic studies are available<sup>24,36</sup>. However, as technology and datasets improve, methods will need to continue to improve and researchers will need to remain aware of the potential pitfalls of each method. Methodological considerations related to subject recruitment and phenotype measurement, statistical estimation of secondary genetic data like haplotypes or ancestry, and methods for multiple testing correction are particularly relevant to this thesis. Here, I will briefly review these considerations, the ways error can be introduced and relevant ongoing research into these areas.

#### **1.2.2.1. Subject Recruitment and Phenotype Measurement**

Selection bias is unlikely to be as important in genetic studies as in traditional epidemiology. In traditional epidemiology, selection bias is an important consideration in subject recruitment because it can lead to spurious associations between the exposure and the disease. These spurious associations happen because a factor important for recruitment causes both the exposure and the disease<sup>37</sup>. In genetic studies, the exposure is



the genetic variant. Since genetic variation is present at birth, it is impossible for a requirement for recruitment that happens after birth to cause the genetic variation. Thus, selection bias in the traditional sense is unlikely to be important for genetic studies. However, careful subject recruitment and phenotype measurement will improve an investigator's ability to pick up a signal in genetic studies<sup>20</sup>. Lack of careful phenotype measurement or recruitment of a non-specific population can create so much noise in the signal that a genetic study is unable to identify true associations. The meta-analysis of *CD14* mentioned above is an example where the signal was only present in atopic asthmatics and was washed out when all asthmatics were included<sup>9</sup>. Research into appropriate phenotype measurement for many diseases and related traits is ongoing<sup>10,38,39</sup>. This research will continue to improve our ability to recruit subjects for genetic studies so that we can identify regions where genetic variation is associated with disease.

#### **1.2.2.2. Statistical Inference Methods for Genetic Data**

GWAS often use several statistical inference methods leading up to the final analysis. These include methods for phasing haplotypes, imputing SNPs, and identifying population structure or estimating genetic ancestry. These methods have improved drastically since population genetics models were included in the inference process and large-scale genome-wide data became available<sup>40</sup>. Research on these methods is ongoing, and the current versions can be applied to genome-wide data in a relatively time-efficient manner. Furthermore, these methods are frequently evaluated in comparison with each other and the expected levels of error have been published as part of these evaluations<sup>e.g.</sup><sup>35, 41,42</sup>. However, published GWAS rarely discuss error introduced as a result of using

these methods and how this might explain differences between their results and others' results.

There are several ways that error introduced by these methods might cause differing results across studies. One example is in global ancestry estimation. The number of ancestry informative markers (AIMs) and the specific AIMs used for global ancestry estimation are chosen on a study-by-study basis. Several panels of AIMs have been proposed for various populations. However, systematic evaluations comparing estimates of global ancestry from these AIMs panels to estimates from genome-wide data have only been performed recently<sup>43,44</sup>. Many previous AIMs panels contained fewer than 200 markers<sup>45,46</sup>. The recent systematic evaluations have found that close to 500 markers are necessary to have a high degree of accuracy in global ancestry estimation. Since global ancestry estimates are used to correct for population stratification, error in these estimates may result in a lack of appropriate correction for population stratification and spurious false positive associations<sup>47</sup>. Alternatively, error in global ancestry estimates may simply add noise and cause investigators to miss true associations.

A second and less understood way that statistical inference methods might cause differing results across studies is if the error from these methods varies by population. In particular, to our knowledge there are no published studies detailing how haplotype inference error varies across populations. The assumptions of statistical inference methods may not hold equally across all populations and may therefore create more error in some populations than in others. For example, many of these methods assume random mating. However, people mate non-randomly and the level of this non-random mating may vary across populations. In fact, two studies have demonstrated that people mate

non-randomly based on genetic ancestry<sup>48,49</sup>. Variation in the validity of assumptions like non-random mating across populations may result in varying error across populations. This, in turn, would result in more bias of results in some populations than in others, making it difficult to compare results across populations. As statistical inference methods continually improve, researchers need to keep investigating and discussing how error in these methods affects results of genetic studies and the comparability of these results across studies. The increasing quantity and quality of publically available genetic data in diverse populations should aid this effort.

### **1.2.2.3. Burden of Proof: Multiple Testing Correction and Replication**

The burden of proof in GWAS has been defined by correcting for multiple testing appropriately and replicating the results. However, both of these are complex issues. The need to correct for multiple testing in genetic studies to avoid false positives is well understood. Despite the fact that Bonferonni correction is notoriously conservative when tests are not independent, many genetic studies use Bonferonni to correct for multiple testing<sup>24</sup>. When SNPs are in linkage disequilibrium, the tests of these SNPs will not be independent. One alternative strategy to help avoid false negatives is to use random permutations to determine an appropriate genome-wide significance level for each study and each type of test<sup>24</sup>. Random permutations are an unbiased way to determine the p-values that are expected by chance in a given population for a given test.

Random permutations are especially advantageous for studies like admixture mapping. Here, the number of independent tests in an admixture mapping study depends on the size of the ancestry blocks in the specific population and the number of loci tested.

Therefore, the number of independent tests varies greatly between studies making random permutations in each study necessary. Even with a study-appropriate significance cutoff determined by random permutations, replication is a commonly accepted way to show that results are not false positives. However, replication may be problematic for rare variants due to low power for detection and the fact that they are often population-specific<sup>50</sup>. Rare variants are not necessarily false positives if they fail to replicate. How to deal with assessing false positives and replication of rare variants is still an ongoing area of research that will undoubtedly require functional follow-up. Researchers should continue to scrutinize their own data and maintain high standards of proof in their own research.

### **1.2.3. Genetics of Asthma and Bronchodilator Response**

Over 100 candidate genes for asthma have been reported<sup>51</sup>. In addition, 23 GWAS have been published for asthma or related traits as of April 6<sup>th</sup>, 2012<sup>18</sup>. The first region identified in a GWAS of asthma covers several genes that are in tight LD, but appears to be caused by variation in *ORMDL3*<sup>52</sup>. This association has been replicated convincingly across many diverse populations<sup>53</sup>. Although this region is well replicated, only a handful of other genes and regions have been replicated convincingly and none have been replicated across all published studies<sup>51,53</sup>. Many of the genes that have strong evidence for association with asthma are involved in epithelial barrier function, environmental sensing and immune detection, T<sub>H</sub>2-mediated cell response, and tissue response<sup>53</sup>. Genes with these functions are sensible based on the known pathology of asthma. However, most of these well-replicated genes were chosen for study based on their function.

*ORMDL3* is clearly associated with asthma, yet it does not fit into these categories. Thus, the complete pathology of asthma is still not understood and GWAS and other agnostic scans of the genome will continue to add to our knowledge of asthma.

Gene-environment interaction studies and genetic studies within specific asthma phenotypes will also add to our knowledge of asthma pathology. Gene-environment interactions likely play a role in the lack of replication seen previously in asthma candidate gene studies<sup>5</sup>. Recent studies of *ORMDL3* provide additional evidence that the effects of risk alleles are modified under specific conditions. *ORMDL3* was found to be associated with asthma only in childhood-onset asthma and in cases where early life exposure to tobacco was present<sup>54</sup>. Researchers are now starting to sequence large populations of asthma patients to add to the wealth of information about genetic variants linked to asthma. These data will be combined with better asthma phenotype and environmental exposure information to help elucidate the causes of asthma.

The genetics of BDR have been much less studied than the genetics of asthma. Only five candidate genes have been reported for BDR and no GWAS have been published to date<sup>18,55-57</sup>. Most of the candidate gene studies for BDR have focused sensibly on the drug target, *ADRB2*. In addition, *ADCY9*, *CRHR2*, *ARG1* and *THRB* have been reported as candidate genes for BDR. Results for all of these candidate genes have been inconsistent. One reason for this may be inconsistent phenotype definitions across studies. BDR is affected by many factors including times of assessment, specific drugs used (*e.g.* long- or short-acting  $\beta_2$ AR agonists), and dose<sup>58</sup>. Genetic variants could be affecting BDR only under specific conditions based on these factors, like a high dose of a particular bronchodilator. Thus, the genetic risk factors that can be identified for BDR

have a lot of potential for variation across studies. Larger genome-wide genetic studies of BDR and more careful phenotyping will add to our understanding of the causes of variation in BDR.

### **1.3. Summary of Chapters**

In the study of the genetics of complex disease, it is important to consider measurement of the disease phenotype, methods used to conduct the genetic study, and to do a thorough job of the analysis itself. This thesis contains three distinct pieces of work that contribute to each of these areas involved in studying the genetics of complex disease.

In chapter 2, I examine ways to define subsets of asthma patients and their utility in predicting exacerbations in 2,743 Latino and African American children with asthma. Asthma has historically been divided into discrete subsets of patients based on many characteristics and more recently based on agnostic statistical clustering methods. My results show that a continuous definition of asthma predicts exacerbations better than subsets of asthma. I suggest that although these subsets are convenient, they may not be an appropriate definition of asthma. Therefore, they may not be a relevant phenotype to use for genetic studies.

In chapter 3, I conduct a genome-wide study of bronchodilator response in 1,782 Latino children with asthma. By examining allelic associations and admixture mapping, I identify several rare variants that may play a role in BDR. This study is, to my knowledge, the first genome-wide association or admixture mapping study of BDR to date. My results suggest that rare variants play an important role in BDR and imply that

further work using sequencing and experimental follow-up should be performed to identify and validate more rare variation involved with BDR.

Finally, in chapter 4, I examine how haplotype inference error can lead to bias in effect estimates in genetic association studies and how this error varies by population. I use publicly available data from four populations in the HapMap project to examine the distribution of error in 100 randomly sampled regions from Chromosome 1. I find that haplotype inference error indeed varies by population and that error in some regions is not explained by factors I hypothesized would be associated with error. My results imply that variation in effect estimates from studies of haplotypes or imputed SNPs may be due to differences in error rather than true differences in effect estimates. Differences in this error will require careful evaluation when comparing results of these studies across populations.

#### **1.4 References**

1. Moorman, J.E., Zahran, H., Truman, B.I., Molla, M.T., Centers for Disease Control and Prevention (CDC) (2011). Current asthma prevalence - United States, 2006-2008. *MMWR Surveill Summ* 60 *Suppl*, 84–86.
2. Moorman, J.E., Rudd, R.A., Johnson, C.A., King, M., Minor, P., Bailey, C., Scalia, M.R., Akinbami, L.J., Centers for Disease Control and Prevention (CDC) (2007). National surveillance for asthma--United States, 1980-2004. *MMWR Surveill Summ* 56, 1–54.
3. Association, A.L. (2010). State of Lung Disease in Diverse Communities 2010. 27–33.
4. Homa, D.M., Mannino, D.M., and Lara, M. (2000). Asthma mortality in U.S. Hispanics of Mexican, Puerto Rican, and Cuban heritage, 1990-1995. *American Journal of Respiratory and Critical Care Medicine* 161, 504–509.
5. Vercelli, D. (2010). Gene-environment interactions in asthma and allergy: the end of the beginning? *Current Opinion in Allergy and Clinical Immunology* 10, 145–148.

6. Drake, K.A., Galanter, J.M., and Burchard, E.G. (2008). Race, ethnicity and social class and the complex etiologies of asthma. *Pharmacogenomics* 9, 453–462.
7. Martinez, F.D. (2007). CD14, endotoxin, and asthma risk: actions and interactions. *Proceedings of the American Thoracic Society* 4, 221–225.
8. Martinez, F.D. (2007). Gene-environment interactions in asthma: with apologies to William of Ockham. *Proceedings of the American Thoracic Society* 4, 26–31.
9. Zhao, L., and Bracken, M.B. (2011). Association of CD14 -260 (-159) C>T and asthma: a systematic review and meta-analysis. *BMC Med Genet* 12, 93.
10. Siroux, V., and Garcia-Aymerich, J. (2011). The investigation of asthma phenotypes. *Current Opinion in Allergy and Clinical Immunology* 11, 393–399.
11. Benton, A.S., Wang, Z., Lerner, J., Foerster, M., Teach, S.J., and Freishtat, R.J. (2010). Overcoming Heterogeneity in Pediatric Asthma: Tobacco Smoke and Asthma Characteristics Within Phenotypic Clusters in an African American Cohort. *J Asthma* 47, 728–734.
12. Eriksson, J., Bjerg, A., Lötvall, J., Wennergren, G., Rönmark, E., Torén, K., and Lundbäck, B. (2011). Rhinitis phenotypes correlate with different symptom presentation and risk factor patterns of asthma. *Respir Med* 105, 1611–1621.
13. Fukutomi, Y., Taniguchi, M., Tsuburai, T., Tanimoto, H., Oshikata, C., Ono, E., Sekiya, K., Higashi, N., Mori, A., Hasegawa, M., et al. (2011). Obesity and aspirin intolerance are risk factors for difficult-to-treat asthma in Japanese non-atopic women. *Clin Exp Allergy* 42, 738-746.
14. Nelson, H.S. (1995). Beta-adrenergic bronchodilators. *N Engl J Med* 333, 499–506.
15. Burchard, E.G., Avila, P.C., Nazario, S., Casal, J., Torres, A., Rodriguez-Santana, J.R., Toscano, M., Sylvia, J.S., Alioto, M., Salazar, M., et al. (2004). Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. *American Journal of Respiratory and Critical Care Medicine* 169, 386–392.
16. Naqvi, M., Thyne, S., Choudhry, S., Tsai, H.-J., Navarro, D., Castro, R.A., Nazario, S., Rodriguez-Santana, J.R., Casal, J., Torres, A., et al. (2007). Ethnic-Specific Differences in Bronchodilator Responsiveness Among African Americans, Puerto Ricans, and Mexicans with Asthma. *J Asthma* 44, 639–648.
17. Blake, K., Madabushi, R., Derendorf, H., and Lima, J. (2008). Population Pharmacodynamic Model of Bronchodilator Response to Inhaled Albuterol in Children and Adults With Asthma. *Chest* 134, 981–989.
18. Hindorff, L., MacArthur, J., Wise, A., Junkins, H., Hall, P., Klemm, A., and Manolio, T. A Catalog of Published Genome-Wide Association Studies. *Genome.Gov*.



19. Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 363, 166–176.
20. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., Mccarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753.
21. Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* 273, 1516–1517.
22. Frazer, K.A., Murray, S.S., Schork, N.J., and Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. *Nat Rev Genet* 10, 241–251.
23. Goldstein, D.B. (2009). Common genetic variation and human traits. *N Engl J Med* 360, 1696–1698.
24. Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7, 781–791.
25. Consortium, T.I.H.3. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52.
26. Consortium, T.I.G.P., author, C., committee, S., Medicine, P.G.B.C.O., BGI-Shenzhen, Broad Institute of MIT and Harvard, Illumina, Technologies, L., Max Planck Institute for Molecular Genetics, Science, R.A., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073.
27. Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature* 456, 18–21.
28. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D.B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biol.* 8, e1000294.
29. Hoffmann, T., Zhan, Y., Kvale, M., and Hesselson, S. (2011). Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* 98, 422-430.
30. Torgerson, D.G., Gignoux, C.R., Galanter, J.M., Drake, K.A., Roth, L.A., Eng, C., Huntsman, S., Torres, R., Avila, P.C., Chapela, R., et al. (2012). Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. *J Allergy Clin Immunol*, in press.
31. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., 1000 Genomes Project, and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. *PNAS* 108, 11983–11988.

32. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81, 559–575.
33. Browning, B.L., and Browning, S.R. (2009). A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics* 84, 210–223.
34. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* 5, e1000529.
35. Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 28, 1359–1367.
36. Nsengimana, J., and Bishop, D.T. (2012). Design considerations for genetic linkage and association studies. *Methods Mol. Biol.* 850, 237–262.
37. Kleinbaum, D.G., Morgenstern, H., and Kupper, L.L. (1981). Selection bias in epidemiologic studies. *Am J Epidemiol* 113, 452–463.
38. Ganesalingam, J., Stahl, D., Wijesekera, L., Galtrey, C., Shaw, C.E., Leigh, P.N., and Al-Chalabi, A. (2009). Latent cluster analysis of ALS phenotypes identifies prognostically differing groups. *PLoS ONE* 4, e7107.
39. Wessman, J., Paunio, T., Tuulio-Henriksson, A., Koivisto, M., Partonen, T., Suvisaari, J., Turunen, J.A., Wedenoja, J., Hennah, W., Pietiläinen, O.P.H., et al. (2009). Mixture model clustering of phenotype features reveals evidence for association of DTNBP1 to a specific subtype of schizophrenia. *Biol. Psychiatry* 66, 990–996.
40. Stephens, M., Smith, N., and Donnelly, P. (2001). A New Statistical Method for Haplotype Reconstruction from Population Data. *The American Journal of Human Genetics* 68, 978–989.
41. Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., et al. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* 78, 437–450.
42. Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11, 499–511.
43. Galanter, J.M., Fernandez-Lopez, J.C., Gignoux, C.R., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., Hidalgo-Miranda, A., Contreras, A.V., Figueroa, L.U., Raska, P., et al. (2012). Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genetics* 8, e1002554.

44. Paschou, P., Lewis, J., and Javed, A. (2010). Ancestry informative markers for fine-scale individual assignment to worldwide populations. *Journal of Medical Genetics* 47, 835-847.
45. Kosoy, R., Nassir, R., Tian, C., White, P.A., Butler, L.M., Silva, G., Kittles, R., Alarcon-Riquelme, M.E., Gregersen, P.K., Belmont, J.W., et al. (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum. Mutat.* 30, 69-78.
46. Halder, I., Shriver, M., Thomas, M., Fernandez, J.R., and Frudakis, T. (2008). A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *Hum. Mutat.* 29, 648-658.
47. Rosenberg, N.A., Huang, L., Jewett, E.M., Szpiech, Z.A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nat Rev Genet* 11, 356-366.
48. Risch, N., Choudhry, S., Via, M., Basu, A., Sebro, R., Eng, C., Beckman, K., Thyne, S., Chapela, R., Rodriguez-Santana, J.R., et al. (2009). Ancestry-related assortative mating in Latino populations. *Genome Biol* 10, R132.
49. Sebro, R., Hoffman, T.J., Lange, C., Rogus, J.J., and Risch, N.J. (2010). Testing for non-random mating: evidence for ancestry-related assortative mating in the Framingham heart study. *Genet. Epidemiol.* 34, 674-679.
50. Stephens, J.C. (2001). Haplotype Variation and Linkage Disequilibrium in 313 Human Genes. *Science* 293, 489-493.
51. Ober, C., and Hoffjan, S. (2006). Asthma genetics 2006: the long and winding road to gene discovery. *Genes Immun* 7, 95-100.
52. Moffatt, M.F., Kabesch, M., Liang, L., Dixon, A.L., Strachan, D., Heath, S., Depner, M., Berg, von, A., Bufe, A., Rietschel, E., et al. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 448, 470-473.
53. Swarr, D.T., and Hakonarson, H. (2010). Unraveling the complex genetic underpinnings of asthma and allergic disorders. *Current Opinion in Allergy and Clinical Immunology* 10, 434-442.
54. Bouzigon, E., Corda, E., Aschard, H., Dizier, M.-H., Boland, A., Bousquet, J., Chateigner, N., Gormand, F., Just, J., Le Moual, N., et al. (2008). Effect of 17q21 variants and smoking exposure in early-onset asthma. *N Engl J Med* 359, 1985-1994.
55. Tantisira, K., and Weiss, S. (2008). The pharmacogenetics of asthma treatment. *Current Allergy and Asthma Reports* 9, 10-17.
56. Duan, Q.L., and Tantisira, K.G. (2009). Pharmacogenetics of Asthma Therapy. *Curr Pharm Des* 15, 3742-3753.

57. Duan, Q.L., Du, R., Lasky-Su, J., Klanderma, B.J., Partch, A.B., Peters, S.P., Irvin, C.G., Hanrahan, J.P., Lima, J.J., Blake, K.V., et al. (2012). A polymorphism in the thyroid hormone receptor gene is associated with bronchodilator response in asthmatics. *The Pharmacogenomics Journal* 1–7.

58. Contopoulos-Ioannidis, D.G., Alexiou, G.A., Gouvias, T.C., and Ioannidis, J.P.A. (2006). An empirical evaluation of multifarious outcomes in pharmacogenetics: beta-2 adrenoceptor gene polymorphisms in asthma treatment. *Pharmacogenet Genomics* 16, 705–711.

## CHAPTER 2

# A CONTINUOUS ASTHMA PHENOTYPE PREDICTS EXACERBATIONS BETTER THAN CLUSTER GROUPS

### 2.1. Introduction

Asthma is a clinically heterogeneous disease. Not all patients have the same symptoms or respond to therapy in the same way. Since asthma is so heterogeneous, clinical subgroups of asthma patients have been defined in many ways<sup>1</sup>. Recently, these clinical subgroups have been labeled ‘asthma phenotypes’<sup>1,2</sup>. Historically, asthma phenotypes have been defined based on expert opinion and using many potentially overlapping variables such as atopy, age of onset, severity and response to therapy. Recently, several studies have used statistical clustering methods to more objectively define asthma phenotypes<sup>3-7</sup>. In addition, multiple papers have suggested that asthma phenotypes should be clinically useful<sup>3,8</sup>. In other words, they should either predict clinical outcomes like exacerbations or response to therapy or they should be directly related to asthma pathology. Asthma phenotypes that are clinically useful will also be more useful in studies of the genetic and environmental risk factors for asthma because they will allow for more accurate identification of risk factors for each asthma phenotype.

Recent asthma phenotypes studies that used clustering have found that subjects were clustered together based on a number of important variables such as lung function, response to and use of medication, and age of onset or asthma duration. However, there are a number of limitations of clustering methods that have not been thoroughly discussed in the asthma phenotypes literature<sup>9</sup>. Two important limitations are that the

researcher must specify a pre-defined number of clusters and that clusters will always be generated whether they are meaningful or not. Thus, these methods are not entirely unbiased. Indeed, the first asthma phenotypes clustering study suggested that these clusters might be hypothesis generating rather than solving<sup>3</sup>. Although clustering has proven useful in identifying some axes of variation in asthma patients, asthma may be better described as a continuous spectrum than discrete clinical subgroups. Furthermore, a continuous spectrum may be more clinically useful, either in terms of predicting clinical outcomes or its relationship to asthma pathology.

To our knowledge, no published studies have compared the fit and clinical utility of asthma phenotypes defined by a continuous variable and defined by clusters. Principal components analysis (PCA) provides a set of principal components (PCs) that are continuous variables and can measure an asthma spectrum using the same set of input variables as clustering. Furthermore, since clustering and PCA are related, PCs can be used to evaluate the separation of clusters<sup>10</sup>. Many other fields use PCA to reduce data and to visualize and define discrete groups and continuums of data. One example is in human genetics. There, PCA is used to clearly separate individuals from different continents based on their genetic markers or to visualize the spread of alleles across a continent<sup>11-13</sup>.

PCA has two definite advantages over clustering when the data form a continuum rather than distinct clusters. First, PCA is more objective than clustering because the researcher is not required to define the number of clusters. Second, PCs have higher power than clusters to predict outcomes like exacerbations when the data are continuous. In this case, the clusters are essentially a categorization of the continuous PC variables

and therefore will result in a loss of power<sup>14</sup>. Thus, exploring the relationship of cluster groups and PCs and the clinical utility of both of these are important to understanding asthma phenotypes.

We hypothesized that a continuous measure of asthma from PCA would predict exacerbations better than cluster group membership. To test this hypothesis, we assessed the ability of PCA and two clustering methods to predict exacerbations in the AGES population of Latino and African American children with asthma. We compared the ability of PCA and two clustering methods to predict hospitalizations, ER visits, and oral steroid use in this population.

## **2.2. Methods**

### **2.2.1. AGES Population**

Subjects were recruited from the Asthma, Genes and Environment Studies (AGES). AGES is a combination of the Genes-Environments & Admixture in Latino Americans (GALA II) Study and the Study of African Americans: Asthma, Genes & Environments (SAGE II). Both studies began in 2008 and are parallel, ongoing, clinic-based case-control studies using similar protocols and questionnaires. Subjects are recruited from urban study centers across the mainland U.S. and Puerto Rico (Table 2.2). The current study includes 2,007 asthma cases from GALA II and 736 asthma cases from SAGE II who were recruited through November 2011.

All participants who met criteria for enrollment (Table 2.1) completed in-person questionnaires related to their medical, asthma, allergic, social, environmental and demographic histories. In addition, all participants provided blood for genetic analysis

and underwent spirometry and skin allergen testing. Each participant or parent was also required to self-identify as having all four grandparents of Latino (GALA II) or African American (SAGE II) ethnicity. Local institutional review boards approved the studies and all subjects or legal guardians provided written informed consent.

**Table 2.1. Eligibility criteria for participation for AGES asthma cases and healthy controls**

| <b>Criterion</b>   |
|--|
| Age between 8 and 21 years old   |
| Self-identified Latino/Hispanic (GALA II) or African American (SAGE II) origin |
| History of physician-diagnosed asthma  |
| Symptoms of coughing, wheezing or shortness of breath in the past 2 years      |
| No respiratory infections for $\geq 6$ weeks (clinical stability)              |
| No asthma exacerbations for $\geq 6$ weeks (clinical stability)                |
| Less than 10 pack year smoking history and no smoking in the last year         |
| If pregnant, < 3rd trimester   |
| No history of other lung diseases or other chronic illnesses                   |

**Table 2.2. Number of cases from participating study centers and institutions in the GALA II and SAGE II studies**

| <b>Study Center</b>            | <b>Institution</b>   | <b>GALA II</b> |                      | <b>SAGE II</b> |                      |
|--------------------------------|--|----------------|----------------------|----------------|----------------------|
|                                |  | <b>Total</b>   | <b>Complete Data</b> | <b>Total</b>   | <b>Complete Data</b> |
| Chicago                        | Northwestern University  | 101            | 23                   | -              | -                    |
|                                | Children's Memorial Hospital   | 227            | 47                   | -              | -                    |
| Houston                        | Baylor College of Medicine Ben Taub Children's Center, Texas Children's Hospital | 213            | 63                   | -              | -                    |
|                                | Jacobi Medical Center  | 308            | 131                  | -              | -                    |
| New York City                  | Jacobi Medical Center  | 308            | 131                  | -              | -                    |
| Puerto Rico                    | Centro de Neumologia Pediatrica (San Juan)                                       | 471            | 229                  | -              | -                    |
|                                | VA Medical Center (San Juan)   | 343            | 110                  | -              | -                    |
| San Francisco Bay Area         | Kaiser Permanente  |                |                      |                |                      |
|                                | Richmond Medical Center  | 8              | 2                    | 118            | 51                   |
|                                | Oakland Medical Center   | 8              | 1                    | 88             | 39                   |
|                                | Vallejo Medical Center   | 52             | 12                   | 221            | 66                   |
|                                | Bay Area Pediatrics (Oakland)  | 3              | 3                    | 83             | 50                   |
|                                | Children's Hospital (Oakland)  | 21             | 7                    | 222            | 59                   |
|                                | La Clinica de la Raza (Oakland)  | 103            | 76                   | -              | -                    |
|                                | Alta Vista (Oakland)   | 13             | 6                    | -              | -                    |
| San Francisco General Hospital | 136  | 108            | 4                    | 2              |                      |



### **2.2.2. Variable Selection**

We selected variables to use for Random Forest, PCA and kmeans clustering that might be indicative of the type of asthma and the status of the disease process in each individual. We selected 52 input variables related to health and family health history, symptoms, control, medication use, spirometry, age, gender, height and weight to use as input variables (Table 2.3). Then, we eliminated variables with >20% missing information. After this elimination, we used 40 input variables. Before implementing any of the methods described below, we scaled and centered all input variables. In addition, categorical variables were coded as dummy variables so that in total we used 67 input variables. For our secondary analysis of imputed data (described below), we included all 52 originally selected input variables. We attempted prediction of three outcomes related to exacerbations. These outcomes were hospitalizations, emergency room (ER) visits, and oral steroid use. These outcomes were self-reported retrospectively in the last 12 months.

### **2.2.3. Primary Analysis**

All statistical analyses were performed using R v2.11.1 and the packages mentioned below<sup>16</sup>. The methods we used required non-missing data, which included 1,085 cases with data for the final 40 input variables in AGES. For the primary analysis, we randomly split these cases into a training set and a validation set. This division resulted in 548 and 537 individuals in the training and validation sets, respectively. We also conducted a secondary analysis on 2,718 individuals using imputed data for 52 variables (described below).

**Table 2.3. Input Variables Used for Principal Components, Kmeans Clusters, and Random Forest in AGES**

| <b>Variable Type</b>                    | <b>Variable</b>  | <b>Percent Missing</b> |
|---|--|------------------------|
| <b>Exacerbation Outcomes</b>            | Hospitalized because of asthma in the last 12 months   | 3.5                    |
|   | Visited the ER because of asthma in the last 12 months   | 2.3                    |
|   | Used oral steroids for asthma in the last 12 months  | 3.5                    |
| <b>Demographic / Basic</b>              | Sex  | 0                      |
|   | Height   | 1                      |
|   | Weight   | 2.4                    |
|   | BMI category, determined from CDC growth charts and percentiles  | 5.3                    |
|   | Current age  | 0.3                    |
| <b>Medication Use in Last 12 months</b> | Short-acting Beta Agonist  | 0                      |
|   | Inhaled Corticosteroid   | 0                      |
|   | Over-the-counter allergy medication  | 2.4                    |
|   | Acetaminophen  | 1.1                    |
| <b>Asthma &amp; Symptoms</b>            | Asthma duration in years   | 15.9                   |
|   | Age of asthma onset  | 15.9                   |
|   | Has regular doctor who treats asthma   | 1.1                    |
|   | Has had trouble sleeping because of wheezing or coughing in last two weeks   | 13.9                   |
|   | How many days has child coughed or wheezed in last two weeks   | 17.5                   |
|   | In the last two weeks, has child:<br>Wheezed, had shortness of breath, or coughed so much he/she couldn't finish a sentence? | 27.5                   |
|   | Had trouble keeping up with others while playing a sport or exercising because of wheeze / shortness of breath / cough       | 24.5                   |
|   | In the last week:  |                        |
|   | • How often did child wake up from asthma during the night   | 0.9                    |
|   | • How bad were symptoms in the morning   | 0.9                    |
|   | • How limited were activities because of asthma  | 0.9                    |
|   | • How much shortness of breath did child have because of asthma  | 0.9                    |
|   | • How much of the time did child wheeze  | 1                      |
|   | In the last 12 months, have any of these made wheezing/coughing worse:   |                        |
|   | • Weather  | 15.1                   |
|   | • Pollen   | 19.7                   |
|   | • Cold or flu  | 15.1                   |
| • Physical activity                     | 16   |                        |
| • Housedust                             | 18.6   |                        |
| • Pets or animals                       | 19.5   |                        |
| • Windy conditions                      | 17.1   |                        |
| • Perfumes or odors                     | 18.3   |                        |
| • Pollution                             | 21.3   |                        |
| • Smoke                                 | 20.7   |                        |
| • Mold                                  | 22.7   |                        |
| • Wood smoke                            | 23.5   |                        |
| • Street dust                           | 22.2   |                        |
| • Food                                  | 20.9   |                        |

**Table 2.3. Continued**

| <b>Variable Type</b>                       | <b>Variable</b>  | <b>Percent Missing</b> |
|--|--|------------------------|
| <b>Health History &amp; Family History</b> | Ever diagnosed with hayfever   | 3.6                    |
|  | Ever had an itchy rash that came and went for at least six months                                    | 1.1                    |
|  | Ever diagnosed with eczema   | 1.6                    |
|  | Ever diagnosed with a sinus infection or sinusitis   | 1.9                    |
|  | Has problems with sneezing, runny or blocked nose, itchy or watery eyes without having a cold or flu | 0.3                    |
|  | Has had this problem in last 12 months   | 0.8                    |
|  | Family history of:   |                        |
|  | Asthma   | 20.5                   |
|  | Eczema   | 31.6                   |
|  | COPD or bronchitis   | 55.5                   |
| <b>Spirometry**</b>                        | Pre-bronchodilator Forced Expiratory Volume in 1 second (FEV <sub>1</sub> ), % of predicted          | 1.9                    |
|  | Pre-bronchodilator Forced Vital Capacity (FVC), % of predicted                                       | 1.9                    |
|  | Post-bronchodilator FEV <sub>1</sub> , % of predicted after 4 puffs of albuterol                     | 2                      |
|  | Post-bronchodilator FVC, % of predicted after 4 puffs of albuterol                                   | 2                      |
|  | % change in FEV <sub>1</sub> after 4 puffs of albuterol  | 2                      |
|  | Maximal FEV <sub>1</sub> after either 4 or 6-8* puffs of albuterol, % of predicted                   | 6.3                    |
| <b>Biomarkers</b>                          | Total serum IgE  | 2.3                    |

\*Second dose depends on age: Patients < 16 received two additional puffs and patients >16 received four additional puffs

\*\*FEV<sub>1</sub> and FVC predicted based on equations from Hankinson *et al.*<sup>15</sup>

In the training set, we built PCs, kmeans clusters, and Random Forest classifiers using all of the input variables selected. PCs were used as a continuous measure of asthma. Since PCA creates as many PCs as there are input variables, we used only the ones that explained the largest proportion of the variance in the sample for prediction. We chose to use PCs 1 and 2, which together represent a two-dimensional asthma spectrum, after looking for separation of the eigenvalues of each of the PCs on a scree plot. PCs 1, 2 and 3 explained 7.8%, 6.9%, and 4.9% of the variance in the sample, respectively.

We used two clustering methods to define clinical subgroups of patients with asthma. First, we built kmeans clusters using the R package `cclust`<sup>17</sup>. Kmeans clustering requires that the number of clusters,  $k$ , be pre-specified. To determine the appropriate number of clusters, we clustered each training set multiple times with the number of clusters ranging from  $k=2$  to  $k=10$ . Then, we used clustergrams to visually examine the stability of each point within each cluster<sup>18</sup>. The second clustering method we used was the algorithm developed in the Severe Asthma Research Program (SARP) study<sup>4</sup>. This algorithm was developed using Ward's minimum-variance hierarchical clustering followed by a discriminant analysis that identified variables important for assigning individuals to cluster groups. We assigned patients with asthma to one of five 'SARP groups' using this algorithm based on percent of predicted pre-bronchodilator FEV<sub>1</sub>, maximal percent of predicted post-bronchodilator FEV<sub>1</sub>, and age of onset.

We also built Random Forest classifiers in the training set using the R package `randomForest`<sup>19</sup> to estimate the maximum predictive ability of the set of input variables we used. Random Forest classifiers differ from PCs and clusters because they are built for the purpose of predicting a particular binary outcome. However, like PCs and clusters, Random Forest classifiers can be built in the training sets and assigned in the validation set to predict exacerbations.

In the validation set, we assigned the PCs, kmeans clusters, SARP groups, and Random Forest classifications. Then, we predicted each of the three exacerbation outcomes from each of these measures. For each prediction, we used a logistic regression of the outcome on a predictive measure. For example, we regressed whether or not an individual was hospitalized in the last 12 months on each individual's first and second PC

(PC1 and PC2) values. We performed regressions of each outcome on the Random Forest classification, PC1 and PC2, kmeans clusters, and SARP groups. We used the fitted values and observed outcomes from each logistic regression to make an ROC curve. Then, we used the area under an ROC curve (AUC) to compare the predictive ability of each predictor for each outcome.

#### **2.2.4. Prediction from Fewer Input Variables**

Measuring 40 variables in order to define clinical subgroups or predict exacerbations in the clinic might be difficult. Therefore, we repeated the analysis predicting exacerbations from PCs starting with fewer input variables. We used the 9 variables that were the union of the top 7 variables loading the original PCs 1 and 2 to create ‘reduced PCs’. We used the first two ‘reduced PCs’ to repeat the analysis in the training and validation sets described above.

#### **2.2.5. Analysis of Imputed Data**

To investigate the effect of dropping individuals with missing data from our primary analysis, we used multiple chained imputation as implemented in the R package mice to fill in missing data in the AGES population<sup>20</sup>. We started with all 52 variables selected for Random Forest, PCA, and clustering. We removed variables from this set that were derived from other variables in the set and recalculated them after imputation. The variables we removed and recalculated after imputation were BMI category, delta FEV<sub>1</sub>, asthma duration, and maximal FEV<sub>1</sub>. For each variable we imputed, we used all other variables with a minimum proportion of usable cases > 0.25 and a minimum

correlation with the variable being predicted  $> 0.1$  as predictors. For most of the variables, we used the default imputation methods: predictive mean matching for continuous variables, logistic regression for binary variables, and polytomous regression for factors with more than two levels. We ran 5 imputations with 20 iterations each. We assessed convergence and bias of the imputation chains by plotting the mean and standard deviation of each variable versus iteration number for each imputation chain.

We checked that the imputations were sensible by plotting the distribution of observed and imputed values for each variable and checking that they were similar. We could not get the distribution of the number of days of wheezing and coughing in the last two weeks to be similar in the observed and imputed data with any of the available imputation methods. Therefore, we used random sampling to fill in this variable from the observed data. Before using the imputed data to replicate our original results, we dropped 25 individuals who had imputed values that were inconsistent with their actual data. We dropped one individual for whom we could not calculate a BMI percentile, nine individuals whose delta FEV<sub>1</sub> was  $> 60\%$ , 14 individuals whose age of onset was greater than their current age, and one individual whose maximal FEV<sub>1</sub> was  $> 200$ . We used 2,718 individuals to repeat the analysis of prediction from Random Forest, PCs 1 & 2, kmeans clustering, and SARP groups as described above.

## **2.3. Results**

### **2.3.1. Principal Components 1 and 2 Represent a Continuous Asthma Spectrum**

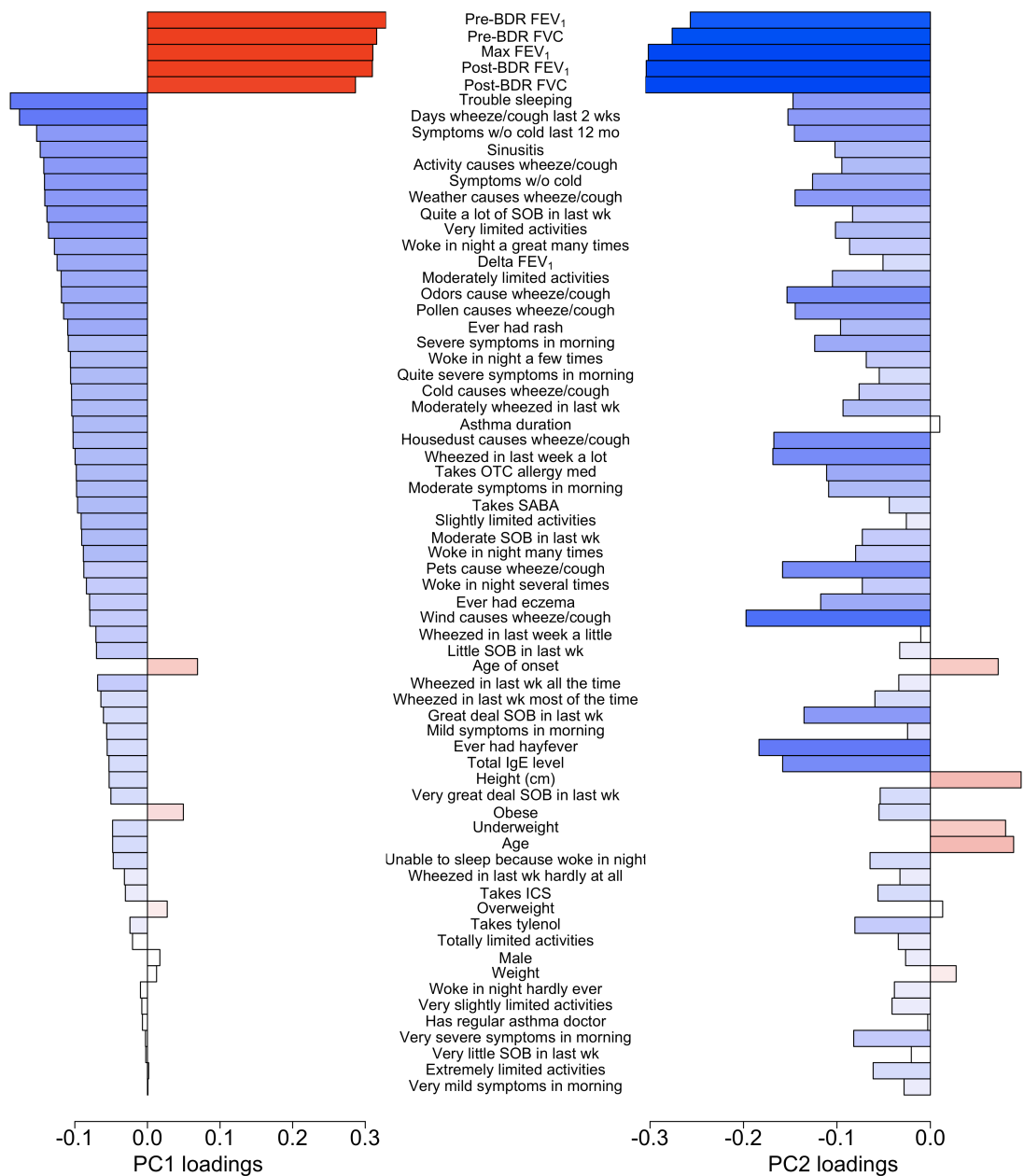
PCs 1 and 2 built in the AGES data capture the most and second-most variance in these data, respectively. The loadings on PCs 1 and 2 show which variables are the most

important contributors to these PCs (Figure 2.1). The most important variables for each PC are those with the largest absolute loadings. PC 1 and 2 values for a new individual are predicted by summing over, for each PC, the multiplication of the loading for each variable by the value of the individual's variable.

The most important variables contributing to PCs 1 and 2 are different, although spirometry is very important for both. The spirometry variables pre-bronchodilator FEV<sub>1</sub> and FVC, post-bronchodilator FEV<sub>1</sub> and FVC, and maximal FEV<sub>1</sub> contributed the most information to both PCs 1 and 2, although the direction of their loading and the order of importance differ. Trouble sleeping because of asthma symptoms and number of days wheezing or coughing in the last 2 weeks were also among the top variables contributing to PC1. Wheezing and coughing because of wind and hayfever were also among the top variables contributing to PC2. These top variables contribute most to the continuous asthma spectrum measured by PCs 1 and 2.

Since measuring 40 variables to predict exacerbations might be tedious, we used only the union of the top 7 variables from each of the original PCs 1 and 2 to create new PCs from only 9 variables. PCs 1 and 2 from these 'reduced PCs' had similar loading patterns to the original PCs 1 and 2 (data not shown). Prediction from these reduced PCs 1 and 2 is compared to prediction from the original PCs 1 and 2 below.

**Figure 2.1. PCs 1 & 2 are both heavily loaded by spirometry variables but other contributing variables differ.** The x-axis shows PC1 loading on the left and PC2 loading on the right. Bars are colored from darker blue to darker red based on increasing positive loading. The y-axis (down the middle) shows the variables names that contributed to the PCs. BDR= bronchodilator response; SOB = shortness of breath; OTC = over-the-counter; SABA = short-acting beta agonist; ICS = inhaled corticosteroid.



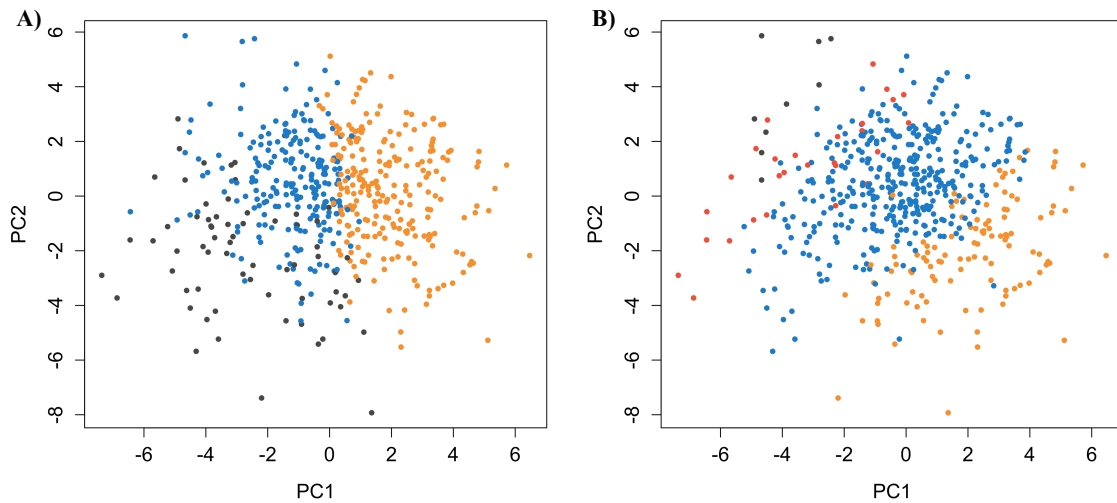


### **2.3.2. Three Kmeans Clusters and SARP Groups Represent Clinical Subgroups of Asthma and are Related to Principal Components**

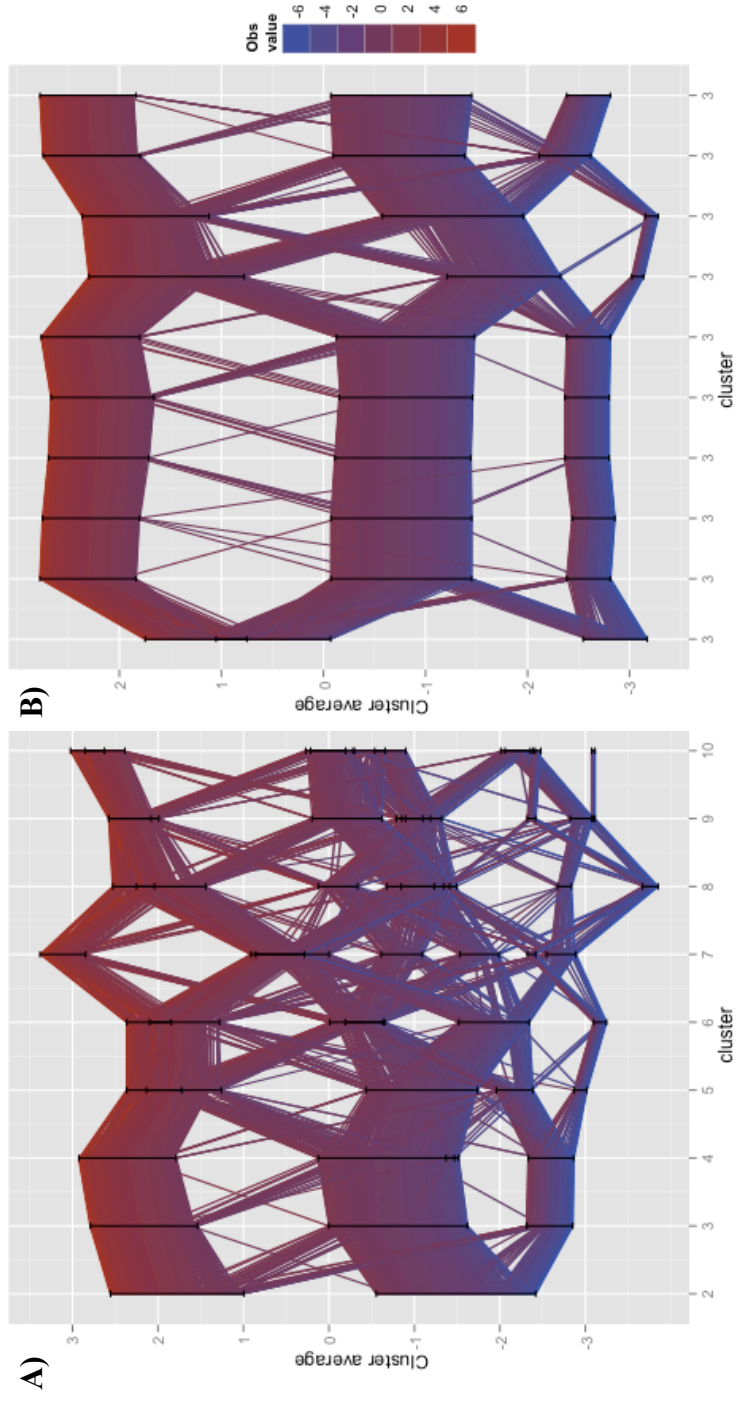
In kmeans clustering, the number of clusters is determined either arbitrarily or based on subject matter knowledge. Since we had no expected number of clusters in this case, we used clustergrams to determine the appropriate number of clusters in AGES (Figure 2.3). Three kmeans clusters were qualitatively the most stable in two ways. First, the points from these clusters stayed together more than the points from other clusters as the number of clusters changed (Figure 2.3A). Second, the most points from these clusters stayed together over repeated formation of three clusters (Figure 2.3B). Therefore, we used three kmeans clusters to predict exacerbations in our study.

The SARP study developed an algorithm to assign patients to five SARP groups<sup>4</sup>. However, because AGES includes only individuals 21 and under, there are only four SARP groups represented in AGES. In addition, the proportion of individuals in the most severe SARP groups is lower in our study than in the original SARP study. Both the kmeans clusters and SARP groups are related to PCs 1 and 2 in AGES (Figure 2.2 A and B, respectively). The scatterplot of PCs 1 and 2 displays a continuous spectrum of points without clear breaks in the points that could be easily distinguished as clusters. Both the kmeans clusters and the SARP groups divide this spectrum of points along lines in the plane of PCs 1 and 2. This indicates that the cluster groups form a categorical variable from the continuous PCs.

**Figure 2.2. A) Kmeans clusters and B) SARP groups separate PCs 1 and 2 by slicing a continuous set of points into groups.** For both plots, the y-axis is the PC2 coordinate of each individual and the x-axis is the PC1 coordinate of each individual. The colors indicate membership in the three kmeans clusters or the four SARP groups present in AGES in panels A and B, respectively.



**Figure 2.3. Clustergrams indicate that three kmeans clusters are the most stable in AGES. A) Average of PC1 in each cluster vs. number of clusters (k). B) Average of PC1 in each cluster vs. repeated kmeans clustering runs with k=3 to check the stability of points.** Each vertical bar represents one cluster and the total number of bars at each point on the x-axis is k. The width of each bar indicates the proportion of points in each cluster. The horizontal lines show the changing cluster membership of each point as k changes. The color of each horizontal line indicates each point's PC1 value according to the scale on the right. Even when  $k > 3$ , most of the points from the  $k=3$  clusters stay together.



### 2.3.3. Principal Components Predict All Outcomes Better than Clusters

We measured the area under the ROC curves (AUC) for Random Forest, PCs 1 and 2, kmeans clusters, SARP groups, and the reduced PCs 1 and 2 for each outcome (Table 2.4). We measured these AUCs both for the original and imputed data. In addition, we plotted the ROC curves for hospitalization, ER visits, and oral steroid use in the last 12 months from the original data (Figure 2.4). Random Forest, our estimate of the maximum predictive ability of the input variables, predicts all outcomes better than any other predictor. Across all three outcomes, the mean AUC for Random Forest is 71%. The mean AUC of 68.7% for PCs 1 and 2 is very close to Random Forest. In contrast, the mean AUCs of 61.3% for kmeans clusters and 55.3% for SARP groups are considerably lower than Random Forest or PCs 1 and 2. The mean AUC of 63.7% for the reduced PCs was between the AUCs for PCs 1 and 2 and kmeans clusters.

**Table 2.4. AUCs for all outcomes and predictors with non-missing or imputed data.**

|                   | <b>Hospitalizations*</b> | <b>ER Visits*</b> | <b>Oral Steroid Use*</b> |
|-------------------|--------------------------|-------------------|--------------------------|
| Random Forest     | 0.67 (0.70)              | 0.72 (0.72)       | 0.74 (0.74)              |
| PCs 1 & 2         | 0.65 (0.69)              | 0.70 (0.68)       | 0.71 (0.69)              |
| Kmeans, k=3       | 0.56 (0.67)              | 0.63 (0.64)       | 0.65 (0.64)              |
| SARP groups       | 0.58 (0.57)              | 0.53 (0.52)       | 0.55 (0.53)              |
| Reduced PCs 1 & 2 | 0.62 (0.69)              | 0.66 (0.68)       | 0.64 (0.69)              |

\*AUCs from imputed data are shown in parentheses.

### 2.3.4. Analysis of Imputed Data Supports Original Results

To investigate the effect of dropping individuals with missing data from our primary analysis, we repeated our original analysis using multiple chained imputation to fill in missing data. We ran 5 imputations with 20 iterations each. After 20 iterations, we assessed the convergence and bias of the imputation chains by plotting the mean and standard deviation of each variable versus iteration number for each imputation chain

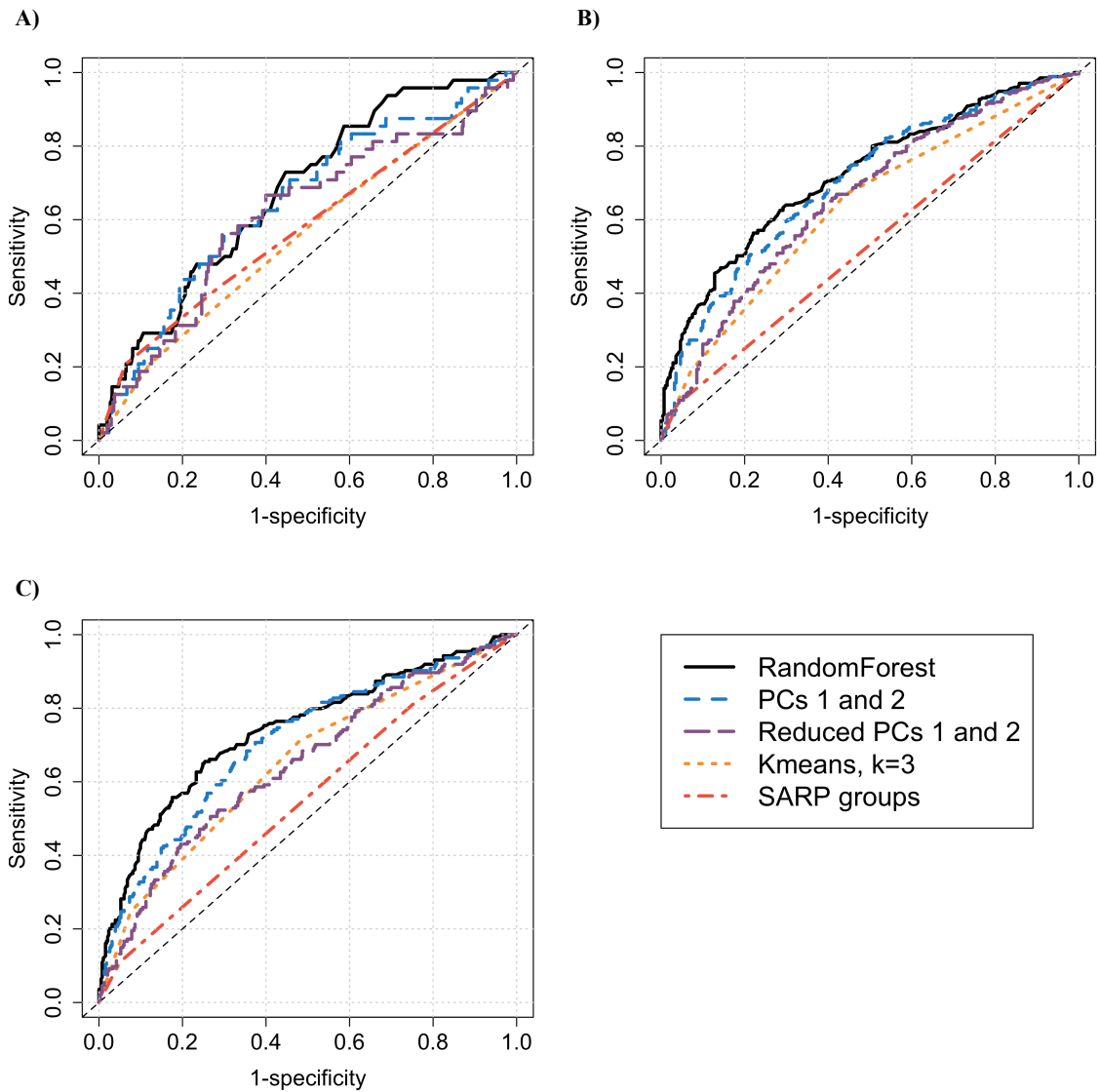
(e.g. Figure 2.5). The similarity and overlap of the lines for each of the 5 imputations indicated that there was no bias in any of the 5 imputations. However, since the mean and standard deviation did not converge, most of the variables showed no evidence of convergence after 20 iterations. Despite the fact that most of the variables did not converge, the distribution of imputed values for these variables was similar to the distribution of the original values (data not shown). Therefore, we used the data from the first imputation chain to repeat our original analysis.

Although most of the results from the imputed data were similar to the original analysis, the most important variables loading PCs 1 and 2 were somewhat different (Figure 2.6). Many of the most important variables for PC 1 in the imputed data had a high percent missingness (Table 2.3). Despite the difference in variable importance, 3 kmeans clusters were the best fit for the imputed data. These clusters were related to PCs 1 and 2 in the same way as in the original data.

Most of the AUCs from the imputed data follow the same trends as in the original data. However, in the imputed data the AUCs for hospitalization are higher than the original AUCs for hospitalization. This may be due to the fact that hospitalizations are the most rare exacerbation and are thus the most difficult to predict in the smaller original data set. Thus, the mean AUCs for all of the predictors are higher. The mean AUC for Random Forest in the imputed data is 72%. Following the trend in the original data, the mean AUC in the imputed data for PCs 1 and 2 is 68.7%. Similarly, the mean AUCs in the imputed data for kmeans clusters and SARP groups are 65% and 54%, respectively. The mean AUC in the imputed data for the reduced PCs differed from the original data. The mean AUC in the imputed data for the reduced PCs 1 and 2 is 66.7%. Although the

means are slightly higher because the prediction of hospitalization is slightly better in the imputed data than in the original data, the trends are the same for all but the reduced PCs.

**Figure 2.4. Principal components predict A) hospitalization, B) ER visits, and C) oral steroid use better than kmeans cluster or SARP groups. ROC curves plot sensitivity vs. 1-specificity for each outcome.**

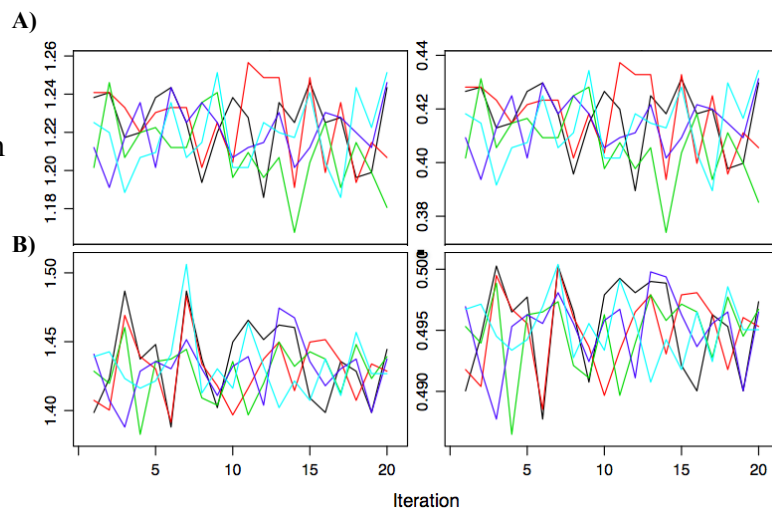


**Figure 2.5. Most variables mean and standard deviation are not biased but do not converge with imputation.** Two variables are given as examples: **A) trouble sleeping because of wheezing/coughing in the last two weeks** and **B) smoke making wheezing/coughing worse**. The change in the mean (left) and standard deviation (right) over 20 iterations are shown for each of 5 imputation chains (colors). Most variables looked similar to A

where the mixing of the strands indicated no bias in any imputation chain.

However, there was no convergence of the chains.

B is an example of a variable where the chains converged.

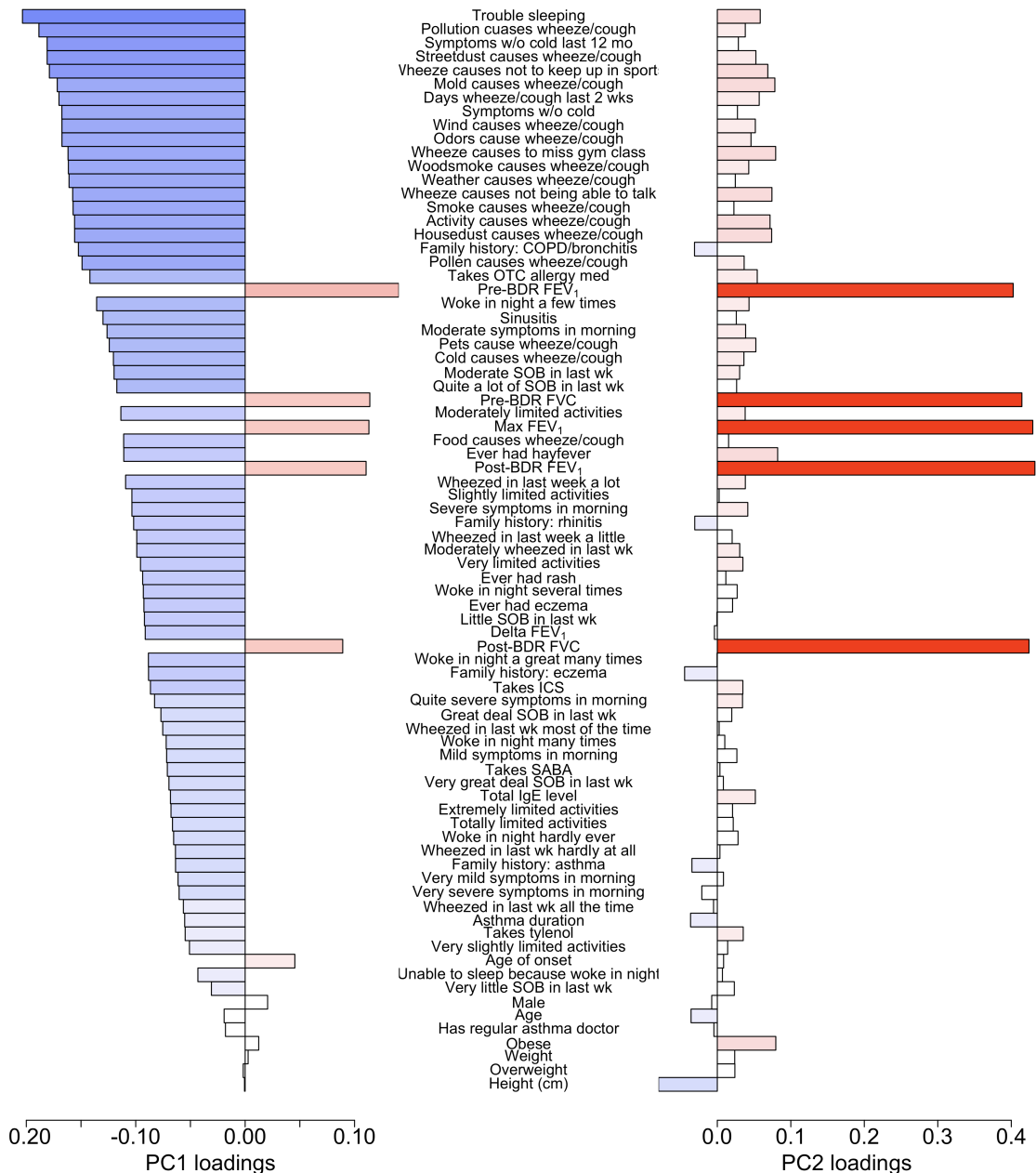


## 2.4. Discussion

In this study, we compared the predictive ability of PCs, kmeans clusters and SARP cluster groups in Latino and African American patients with asthma from AGES. We found that PCs predict exacerbations better than kmeans clusters, which predict exacerbations better than SARP groups. In our study, the AUC was 7.4% and 13.4% higher on average for PCs compared to kmeans clusters and SARP groups, respectively (Table 2.5). Furthermore, the AUC was only 2.3% lower on average for PCs compared to Random Forest, which we used as an estimate of the maximum predictive ability of our

**Figure 2.6. Loadings of PCs 1 & 2 after imputation differ from original PCs 1 & 2.**

The x-axis shows PC1 loading on the left and PC2 loading on the right. Bars are colored from darker blue to darker red based on increasing positive loading. The y-axis (down the middle) shows the variables names that contributed to the PCs. Abbreviations: BDR= bronchodilator response; SOB = shortness of breath; OTC = over-the-counter; SABA = short-acting beta agonist; ICS = inhaled corticosteroid.





input variables. In addition to demonstrating that PCs predict exacerbations better than cluster groups, our results suggest the reason for this observation. We found that both kmeans clusters and SARP groups split the 2-dimensional continuum created by PCs 1 and 2 into categories (Figure 2.3). Since the data appear to be continuous, we would expect a categorical representation of these data to have less power than a continuous one<sup>14</sup>. Thus, it is not surprising that clusters that split the continuous PC variables into categories have lower predictive ability. Our results show that PCs 1 and 2 are a better representation of the data than either kmeans clusters or SARP groups. Given the relationship between PCs 1 and 2 and the cluster groups from either method, PCs are likely to predict both exacerbations and other clinical outcomes like response to therapy better than cluster groups.

A few researchers have suggested that asthma phenotypes like clusters or PCs should be clinically useful by either predicting an outcome like exacerbations or being related to asthma pathology<sup>3,8</sup>. To be useful in predicting exacerbations, asthma phenotypes will need to perform better than the current best clinical predictors of exacerbations. Previous exacerbations and asthma control are the best clinical predictors of exacerbations identified so far, although neither is perfect<sup>21</sup>. Three recent studies have found that previous exacerbations are associated with increased risk of future exacerbations<sup>22-24</sup>. In addition, one study constructed a clinical score for predicting exacerbations based on 17 questions<sup>25</sup>. This study found AUCs of 0.75 in their original sample and 0.69 in a second sample, which are in the same range as the AUCs from PCs in the current study. Although one clustering study found statistically significant differences in exacerbations based on cluster membership<sup>3</sup>, the present study is, to our

knowledge, the first to directly compare prediction of a clinical outcome from asthma phenotypes based on clustering methods and a continuous measurement. To determine the ultimate utility of asthma phenotypes in predicting exacerbations, a direct comparison of clusters, PCs and other predictors of exacerbations in the same study will be necessary. In addition, the ability of asthma phenotypes to predict other clinical outcomes and their relation to asthma pathology will need to be investigated to determine their clinical utility.

As researchers investigate the clinical utility of asthma phenotypes, they will also refine the phenotypes. Given the continuous nature of our data, a continuous measure of the asthma spectrum is a promising candidate for the best asthma phenotype. However, measuring 40 variables in the clinic might be prohibitive. On the other hand, our results suggest that too much reduction in the number of variables results in lower prediction. This is true when comparing both the reduced PCs to the original PCs and the SARP groups to the kmeans clusters. The reduced PCs and SARP groups were formed based on only nine and three variables, respectively. It is possible that an intermediate number of variables would provide accurate prediction that was reproducible across studies. In addition, prediction may be improved by including more information on other predictors of clinical outcomes, like previous exacerbations or asthma control. The appropriate number of variables and the optimal set of variables are not clear and should be assessed across several clinical populations and in comparison to other predictors of exacerbations.

The results of our imputed data analysis indicate that it may be possible to maintain a high level of prediction with a reduced number of input variables for PCs. However, it is not clear how accurate our imputed analysis was. Although our

imputations appeared to be unbiased, they never converged (Figure 2.5). Furthermore, the top variables in our imputed PCs were slightly different than those in our original analysis (Figure 2.6). Many of the top variables in the imputed PCs were those with a high percent missingness in the original data. The imputation may have caused the variance to be larger in these variables than it actually is. This would cause them to be important variables in the imputed PCs when they might not otherwise be important. Nonetheless, the level of prediction from the imputed data in general is similar to that in the original analysis. Since the AUCs are not lower in the imputed analysis than in the original analysis, it is unlikely that the loss of data in the original analysis caused an overestimation of the predictive ability of the PCs and clusters. However, evaluation of both PCs and clusters should be performed in a dataset with fewer observations with missing data.

In conclusion, we found that a continuous measure of the asthma spectrum predicts exacerbations better than two clustering methods that define clinical subgroups of patients with asthma. Our findings suggest that a continuous measure of the asthma spectrum is a better description of the data than cluster groups and will likely predict many outcomes better than cluster groups. However, using the number of variables we used to create the continuous PC measure may be impractical in the clinic. A more sparse but continuous measure of the asthma spectrum that balances the number of variables with the desired predictive ability should be developed across several studies and in comparison with the current standard for predicting exacerbations and response to therapy.

## 2.5. References

1. Wenzel, S.E. (2006). Asthma: defining of the persistent adult phenotypes. *Lancet* 368, 804–813.
2. Wardlaw, A.J., Silverman, M., Siva, R., Pavord, I.D., and Green, R. (2005). Multi-dimensional phenotyping: towards a new taxonomy for airway disease. *Clin Exp Allergy* 35, 1254–1262.
3. Haldar, P., Pavord, I.D., Shaw, D.E., Berry, M.A., Thomas, M., Brightling, C.E., Wardlaw, A.J., and Green, R.H. (2008). Cluster Analysis and Clinical Asthma Phenotypes. *American Journal of Respiratory and Critical Care Medicine* 178, 218–224.
4. Moore, W.C., Meyers, D.A., Wenzel, S.E., Teague, W.G., Li, H., Li, X., D'agostino, R., Castro, M., Curran-Everett, D., Fitzpatrick, A.M., et al. (2010). Identification of Asthma Phenotypes Using Cluster Analysis in the Severe Asthma Research Program. *American Journal of Respiratory and Critical Care Medicine* 181, 315–323.
5. Fitzpatrick, A.M., Teague, W.G., Meyers, D.A., Peters, S.P., Li, X., Li, H., Wenzel, S.E., Aujla, S., Castro, M., Bacharier, L.B., et al. (2011). Heterogeneity of severe asthma in childhood: Confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. *J Allergy Clin Immunol* 127, 382–389.e13.
6. Weatherall, M., Travers, J., Shirtcliffe, P.M., Marsh, S.E., Williams, M.V., Nowitz, M.R., Aldington, S., and Beasley, R. (2009). Distinct clinical phenotypes of airways disease defined by cluster analysis. *European Respiratory Journal* 34, 812–818.
7. Gupta, S., Siddiqui, S., Haldar, P., Entwisle, J.J., Mawby, D., Wardlaw, A.J., Bradding, P., Pavord, I.D., Green, R.H., and Brightling, C.E. (2010). Quantitative analysis of high-resolution computed tomography scans in severe asthma subphenotypes. *Thorax* 65, 775–781.
8. Siroux, V., and Garcia-Aymerich, J. (2011). The investigation of asthma phenotypes. *Current Opinion in Allergy and Clinical Immunology* 11, 393–399.
9. Zhao, Y., and Karypis, G. (2003). Clustering in Life Sciences. In *Functional Genomics: Methods and Protocols*, M. Brownstein, and A. Khodursky, eds. (New Jersey: Humana Press), pp. 183–218.
10. Ding, C. (2004). Principal component analysis and effective k-means clustering. *Proceedings of the 2004 SIAM International Conference on Data Mining*.
11. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38, 904–909.
12. Novembre, J., and Stephens, M. (2008). Interpreting principal component analyses of

spatial population genetic variation. *Nat Genet* 40, 646–649.

13. Boyko, A., Lohmueller, K., and Novembre, J. (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome* ....

14. Royston, P., Altman, D.G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 25, 127–141.

15. Hankinson, J.L., Odencrantz, J.R., and Fedan, K.B. (1999). Spirometric reference values from a sample of the general U.S. population. *American Journal of Respiratory and Critical Care Medicine* 159, 179–187.

16. R Development Core Team (2010). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).

17. Dimitriadou, E. cclust: Convex Clustering Methods and Clustering Indexes.

18. Schonlau, M. (2004). Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. *Comput Stat* 19, 95–111.

19. Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. *R News* 2, 18–22.

20. van Buuren, S., and Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45, 1–67.

21. Forno, E., and Celedón, J.C. (2012). Predicting asthma exacerbations in children. *Current Opinion in Pulmonary Medicine* 18, 63–69.

22. Wu, A.C., Tantisira, K., Li, L., Schuemann, B., Weiss, S.T., and Fuhlbrigge, A.L. (2011). Predictors of Symptoms are Different from Predictors of Severe Exacerbations from Asthma in Children. *Chest* –.

23. Haselkorn, T., Zeiger, R.S., Chipps, B.E., Mink, D.R., Szeffler, S.J., Simons, F.E.R., Massanari, M., and Fish, J.E. (2009). Recent asthma exacerbations predict future exacerbations in children with severe or difficult-to-treat asthma. *J Allergy Clin Immunol* 124, 921–927.

24. Covar, R.A., Szeffler, S.J., Zeiger, R.S., Sorkness, C.A., Moss, M., Mauger, D.T., Boehmer, S.J., Strunk, R.C., Martinez, F.D., and Taussig, L.M. (2008). Factors associated with asthma exacerbations during a long-term clinical trial of controller medications in children. *J Allergy Clin Immunol* 122, 741–747.e744.

25. Forno, E., Fuhlbrigge, A., Soto-Quirós, M.E., Avila, L., Raby, B.A., Brehm, J., Sylvia, J.M., Weiss, S.T., and Celedón, J.C. (2010). Risk factors and predictive clinical scores for asthma exacerbations in childhood. *Chest* 138, 1156–1165.

## CHAPTER 3

# GENOME-WIDE ASSOCIATION STUDY AND ADMIXTURE MAPPING OF BRONCHODILATOR RESPONSE IMPLICATES RARE VARIANTS

### 3.1. Introduction

Short-acting  $\beta_2$ -adrenergic receptor ( $\beta_2$ AR) agonists (SABAs) are the primary rescue medication for individuals having an asthma attack<sup>1,2</sup>. SABAs cause rapid bronchodilation, or smooth muscle relaxation in the airways, by stimulating  $\beta_2$ AR. Response to SABAs is measured by bronchodilator response (BDR), the percent change in forced expiratory volume in one second ( $FEV_1$ ) after administration of a SABA. There is wide inter-individual variability in BDR and not every patient responds<sup>2,3</sup>. The reason for this variability is unknown, but causes likely include both genetic and environmental factors<sup>2</sup>.

Genetic studies of BDR have reported five candidate genes<sup>4,5</sup>. The most commonly studied gene is *ADRB2*, which encodes  $\beta_2$ AR. Evidence for the association of *ADRB2* and BDR is inconsistent. This may be in part because studies have generally failed to attempt replication across the same single nucleotide polymorphisms (SNPs), endpoints, time, comparison, and genetic models<sup>6</sup>. Four other candidate genes have been much less commonly studied: *CRHR2*, *ADCY9*, *ARG1*, and *THRB*. These genes are good candidates because they are functionally similar to  $\beta_2$ AR (*CRHR2*), downstream of  $\beta_2$ AR in its signaling pathway (*ADCY9*), implicated in asthma and part of a pathway that inhibits smooth muscle relaxation (*ARG1*), and differentially expressed after exposure to a  $\beta_2$ -agonist and part of a pathway that stimulates smooth muscle relaxation (*THRB*)<sup>5,7</sup>.

Although these genes are good candidates, lack of replication in more than one paper makes it unclear how important they are for BDR. Despite many candidate gene studies of BDR, no genome-wide association study (GWAS) of BDR has been published to date<sup>8</sup>.

An alternative to GWAS is to use admixture mapping to identify genomic regions containing disease-associated variants that differ in frequency across human populations. Differences in disease prevalence across human populations are a good indication that disease-associated variants may differ in frequency across these populations. As with genetic studies of BDR, there is a paucity of information available on population differences in BDR. In one study, Puerto Ricans had lower BDR than either African Americans or Mexicans, despite having more severe asthma<sup>9,10</sup>. However, this study was limited by the fact that it recruited subjects from specialized asthma clinics. In another study, African Americans had lower BDR than white patients<sup>11</sup>. These patients were recruited from a specialty clinic and newspaper ads. The lack of a population-based study of BDR makes it difficult to assess true differences across populations. There have also been inconsistencies between populations in candidate gene studies of BDR. For example, one paper reported that a variant in *ADRB2* was associated with increased BDR in Puerto Ricans but not in Mexicans<sup>12</sup>. It is possible that these inconsistencies are due to a lack of understanding of which variants are causal and differing linkage disequilibrium patterns across populations. Nevertheless, these findings together with evidence for population differences in BDR provide limited evidence that variants specific to certain ancestries may play a role in determining an individual's BDR. However, no admixture mapping study of BDR has been published to date.

We hypothesized that both common genetic variation and genetic variation more frequent on certain ancestral haplotypes contribute to differences in BDR. To test these hypotheses, we performed a genome-wide association study of BDR in 1,782 Latino patients with asthma from across the US and Puerto Rico. Specifically, we tested for both allelic associations and ancestry associations with BDR.

### 3.2. Methods

#### 3.2.1. Discovery Population: GALA II

**Table 3.1. Participating study centers and institutions in the GALA II study.**

| <b>Study Center</b>    | <b>Institution</b>   | <b>Number of Cases</b> |
|------------------------|--|------------------------|
| Chicago                | Northwestern University  | 95                     |
|                        | Children's Memorial Hospital   | 206                    |
| Houston                | Baylor College of Medicine Ben Taub Children's Center, Texas Children's Hospital | 194                    |
|                        | Jacobi Medical Center  | 285                    |
| Puerto Rico            | Centro de Neumologia Pediatrica (San Juan)                                       | 408                    |
|                        | VA Medical Center (San Juan)   | 296                    |
| San Francisco Bay Area | Kaiser Permanente  | 5                      |
|                        | Richmond Medical Center  | 4                      |
|                        | Oakland Medical Center   | 33                     |
|                        | Vallejo Medical Center   | 3                      |
|                        | Bay Area Pediatrics (Oakland)  | 14                     |
|                        | Children's Hospital (Oakland)  | 99                     |
|                        | La Clinica de la Raza (Oakland)  | 130                    |
|                        | San Francisco General Hospital   | 10                     |
|                        | Alta Vista (Oakland)   |                        |

Subjects in the discovery population were from the Genes-Environments & Admixture in Latino Americans (GALA II) study. Recruitment for the GALA II study began in 2006 and is an ongoing, clinic-based study of children ages 8-21 with and without asthma. Subjects are recruited from urban study centers across the mainland U.S. and Puerto Rico (Table 3.1). A total of 4,045 participants, 1,976 of whom were asthma cases, were recruited through June 2011, when genotyping began.



All participants who met criteria for enrollment (Table 3.2) completed in-person questionnaires related to their medical, asthma, allergic, social, environmental and demographic histories. In addition, all participants provided blood for genetic analysis and underwent spirometry and skin allergen testing. Each participant or parent was also required to identify all four grandparents as Latino. Local institutional review boards approved the studies and all subjects or legal guardians provided written informed consent.

**Table 3.2. Eligibility criteria for participation of asthma cases in GALA II and GALA I.**

| <b>Criterion</b>   | <b>GALA II</b> | <b>GALA I</b> |
|--|----------------|---------------|
| Age between 8 and 21 years old (GALA II) or 8 and 40 years old (GALA I)              | Yes            | Yes           |
| Child and all four grandparents self-identified as Latino/Hispanic origin            | Yes            | No            |
| Child and all four grandparents self-identified as Mexican or Puerto Rican in origin | No             | Yes           |
| History of physician-diagnosed asthma  | Yes            | Yes           |
| Symptoms of coughing, wheezing or shortness of breath in the past 2 years            | Yes            | Yes           |
| No respiratory infections for $\geq 6$ weeks (clinical stability)                    | Yes            | Yes           |
| No asthma exacerbations for $\geq 6$ weeks (clinical stability)                      | Yes            | Yes           |
| Less than 10 pack year smoking history and no smoking in the last year               | Yes            | Yes           |
| If pregnant, < 3rd trimester   | Yes            | Yes           |
| No history of other lung diseases or other chronic illnesses                         | Yes            | No            |

The primary outcome for the current study was BDR after two doses of albuterol. Each subject's baseline FEV<sub>1</sub> was measured prior to administering four puffs of albuterol. After 15 minutes, a post-bronchodilator FEV<sub>1</sub> was measured followed by an additional two (if <17 years of age) or four (if >16 years of age) puffs of albuterol and a second post-bronchodilator FEV<sub>1</sub> measurement. BDR was calculated as the percent change in FEV<sub>1</sub> between the second post-bronchodilator measurement and baseline. Since the prevalence of asthma in Puerto Rico is very high and asthma is frequently discussed<sup>13</sup>, patient self-report of a physician diagnosis of asthma may be inaccurate.

Therefore, patients from the Centro de Neumologia Pediatrica in Puerto Rico were required to have a BDR of at least 8% for inclusion in GALA II. These samples were included in our study and enrich the study for the more extreme upper end of the distribution of BDR.

The only clinical covariate we adjusted for was the subject's self-reported ethnicity. Ethnicity was divided into categories of Puerto Rican, Mexican, Mixed Latino, and Other Latino. Puerto Rican and Mexican ethnicities were available as selections on the questionnaire. Mixed Latinos were defined as any individual who identified with more than one Latino ethnic group. Other Latinos were defined as those who chose only one Latino ethnic group from among Spanish/Hispanic/Latino, Cuban, Dominican, El Salvadorian, Guatemalan, Nicaraguan, Honduran, Colombian, Brazilian, and Argentinian. GALA II is a heterogenous population made up of many Latino ethnicities and a wide range of ancestry and BDR. To balance the need for power and our ability to detect population-specific variants, we performed all analyses in all of GALA II as well as separately in the largest two subsets of GALA II: the Puerto Ricans and the Mexicans.

### **3.2.2. Replication Population: GALA I**

Subjects in the replication population were from the previously described Genetics of Asthma in Latino Americans (GALA I) Study<sup>9</sup>. GALA I is a study of children (probands) and their biological parents recruited from schools, clinics, and hospitals across four sites: San Francisco Bay Area, New York City, Puerto Rico, and Mexico City. All probands who met criteria for enrollment (Table 3.2) completed in-person questionnaires related to their medical, asthma, allergic, and demographic

histories. In addition, all participants provided blood for genetic analysis and underwent spirometry. The Institutional Review Board at the University of California San Francisco approved the research.

BDR in GALA I was measured after two (if < 17 years of age) or four (if > 16 years of age) puffs of albuterol. Each subject's baseline FEV<sub>1</sub> was measured, they were given albuterol, and then, after 15 minutes, post-bronchodilator FEV<sub>1</sub> was measured. BDR was calculated as the percent change in FEV<sub>1</sub> between the post-bronchodilator measurement and baseline.

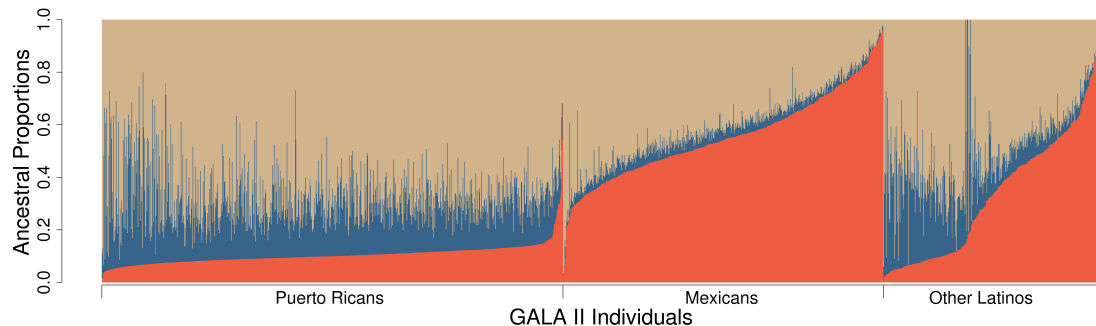
As for GALA II, we adjusted for or stratified on each subject's self-reported ethnicity. Ethnicity was either Mexican or Puerto Rican in the proband and all four biological grandparents.

### **3.2.3. Genotyping and Genetic Ancestry Estimation in GALA II**

Genotyping of GALA II subjects was performed using the Affymetrix Axiom LAT array (Affymetrix, Santa Clara, CA) that contained 817,810 SNPs prior to quality control (QC). We removed SNPs with >5% missing values, failing platform specific SNP quality criteria, or deviating from Hardy-Weinberg equilibrium ( $p < 10^{-6}$ ) within their respective populations. The total number of SNPs passing QC was 568,037. Subjects were filtered based on 95% call rates and gender discrepancies, IBD and standard Affymetrix Axiom metrics. The total number of subjects with asthma passing QC for genotyping was 1,879. We also removed individuals with missing BDR or who were outliers for BDR (BDR > 80 or < -50). Following QC, a total of 1,782 subjects with asthma were included in this study.

We used data from three populations to represent the ancestral haplotypes of Latinos for estimating genetic ancestry: HapMap European (CEU), HapMap African (YRI) and 95 Native American samples. First, the CEU and YRI genotypes were combined across Phase II, Illumina Omni, and Affymetrix Axiom platforms for maximum coverage. In addition, we genotyped 95 Native American samples kindly provided by Cheryl Winkler, Andres Moreno, Karla Sandoval and Carlos Bustamante on the Axiom LAT array. Global admixture was estimated using ADMIXTURE<sup>14</sup>, unsupervised and assuming 3 ancestral populations (Figure 3.1). Local ancestry was estimated using LAMP-LD under a 3-population model, assuming 20 generations of admixture<sup>15</sup> and after phasing the ancestral haplotypes using BEAGLE<sup>16</sup>.

**Figure 3.1. Admixture proportions for GALA II cases.** Each bar represents one individual. For each individual, the proportions of Native American (red), African (blue), and European (tan) ancestry are displayed.



### 3.2.4. Genotyping, Genetic Ancestry Estimation and Imputation in GALA I

Genotyping and estimates of genetic ancestry for GALA I have been described previously<sup>17</sup>. Briefly, GALA I subjects were genotyped using the Affymetrix 6.0 GeneChip that contained more than 900,000 SNPs before quality control. Quality control filters included call rates, Hardy-Weinberg equilibrium, unambiguous mapping to the

human reference genome, consistency between genetic and reported sex, principal components analysis, high or low autosomal heterozygosity, and unexpected pairwise relatedness or genetic identity. After quality control filtering on markers and subjects, genotypes were available for 729,685 markers in 529 children with asthma (253 Mexican and 276 Puerto Rican subjects). Global genetic ancestry was estimated using the program ADMIXTURE<sup>14</sup> assuming 3 ancestral populations. Local ancestry was estimated using the program LAMP<sup>18</sup> under a 3-population model, assuming 20 generations of admixture.

For replication of SNPs that were not among the 729,685 markers that passed quality control in GALA I, we either directly genotyped or imputed them in GALA I. We used pre-designed TaqMan SNP Genotyping Assays (Applied Biosystems, Carlsbad, CA) to genotype two SNPs, rs1281748 and rs1281743. For SNPs that failed TaqMan assay design we first phased the data using the program SHAPE-IT<sup>20</sup> and accounting for relationships within trios. Then, we imputed the SNPs using the program IMPUTE2<sup>19</sup> separately in the GALA I Mexicans and Puerto Ricans. Reference haplotypes for imputation were from Phase I version 3 of the 1000 Genomes Project<sup>21</sup>. We imputed 100kb regions around each SNP with a 20kb buffer on each side. Finally, we filtered SNPs that had an info score  $> 0.3$  to indicate high quality imputation. For all analyses of imputed data, we used the gene dosage output from IMPUTE2 to account for the uncertainty in imputation.

### **3.2.5. Analysis of Allelic Associations**

For each SNP in GALA II, we used a linear regression to test whether the number of minor alleles present was associated with BDR after adjusting for ethnicity, local

African ancestry, local Native American ancestry, global African ancestry, and global Native American ancestry. In addition, we performed the same analysis without adjusting for ethnicity separately in the GALA II Puerto Ricans and Mexicans. The p-values for the SNPs were examined using Manhattan and QQ plots. Confidence bands on QQ plots were determined using a  $\text{beta}(j, n-j+1)$  distribution for the  $j^{\text{th}}$  order statistic when  $n$  SNPs are tested<sup>22</sup>. Although Bonferonni corrections for genome-wide significance are often used in GWAS, they are too conservative if tests are not independent, resulting in Type II errors<sup>23</sup>. To avoid Type II errors, we established a genome-wide significance threshold using random permutations. Random permutations provide an empirical distribution of genome-wide minimum p-values under the null hypothesis of no association between any of the SNPs and BDR. We randomly permuted BDR in all of GALA II 100 times, keeping ethnicity paired with BDR. For each of these 100 permutations of BDR, we tested the association of the permuted BDR with every SNP as described above and tracked the minimum p-value across all SNPs. From the distribution of minimum p-values, we found the p-value that corresponded to a GWAS-wide  $\alpha = 0.05$  to be  $1.6 \times 10^{-7}$ . We used this p-value as our threshold for genome-wide significance in allelic tests of association.

For each SNP that met our threshold for genome-wide significance, we examined the dose-response relationship of the number of alleles and BDR by calculating the mean BDR for individuals who had 0, 1, or 2 minor alleles. In addition, we examined this relationship after removing individuals with a BDR > 60 who were at the upper end of the BDR distribution. SNPs that were genome-wide significant, that had more than three heterozygotes with BDR data, and whose dose-response relationship was consistent even

after potential outliers were removed were carried forward for replication. We attempted replication of allelic associations by imputing genotypes in GALA I as described above. We used linear regression to test for the association between the minor allele dosage and BDR adjusting for local African ancestry, local Native American ancestry, global African ancestry, and global Native American ancestry. We performed these tests separately in the GALA I Mexicans and Puerto Ricans.

### **3.2.6. Analysis of Ancestry Association**

In addition to testing for allelic associations, we tested for the association of local ancestry with BDR separately for African, Native American, and European ancestry using linear regression and adjusting for global ancestry. We performed these tests in all of GALA2 adjusting for ethnicity and separately in the Mexicans and Puerto Ricans.

A Bonferonni correction for genome-wide significance is not appropriate for admixture mapping since ancestry blocks are relatively large, making tests at adjacent SNPs non-independent. Therefore, we used random permutation tests to establish a genome-wide significance cutoff for each ancestry. For these permutations, we used the same 100 permuted BDR and ethnicity values as for the allelic associations and repeated the linear regressions for each ancestry as described above. This resulted in three genome-wide significance cutoffs corresponding to  $\alpha = 0.05$ :  $p < 1.6 \times 10^{-4}$  for African ancestry and  $p < 1.0 \times 10^{-4}$  for European and Native American ancestry. We defined an admixture mapping peak as the region on a chromosome bounded by loci with p-values less than the appropriate permutation cutoff. For each admixture mapping peak, we attempted to replicate the peak in GALA I data by testing for an association of BDR and

local ancestry as described above. Furthermore, we investigated each peak to see whether there were significant allelic associations using linear regression as described above under the peak. We used a Bonferonni correction for the number of SNPs under the peak to determine significance. We attempted replication of significant allelic associations under the admixture mapping peaks in GALA I.

All statistical analyses described in sections 3.2.5 and 3.2.6 were conducted using R (v2.14.1)<sup>24</sup>.

### **3.3. Results**

#### **3.3.1. Rare Variants are Associated with BDR in GALA II**

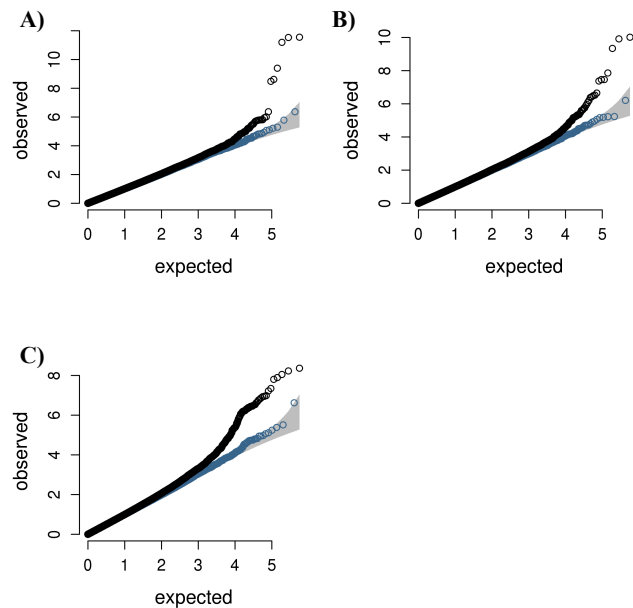
We tested for an association of BDR with each of 568,037 SNPs across the genome in 1,782 children with asthma from GALA II. We performed these tests in all of GALA II and separately in the subsets of 823 Puerto Rican children and 572 Mexican children. For all three subsets of GALA II, there was some inflation in p-values over what was expected by chance around  $p=10^{-3}$  (Figure 3.2). If SNPs with a minor allele frequency (MAF) <5% were removed, the p-values were around those expected by chance, indicating that the signal was driven by less common variants (Figure 3.2). The MAF of these less common variants ranged from singletons at 0.03% with only one heterozygous individual with BDR data to 5%. The singletons are problematic since they may be false positives because of either genotyping error or BDR measurement error. However, the other SNPs driving the signal are promising rare variants.



**Figure 3.2. QQ plots for allelic associations with bronchodilator response (BDR) show that signal is driven by rare variants in A) all of GALA II, B) GALA II Puerto Ricans and C) GALA II Mexicans.**

Y-axis: Observed  $-\log_{10}$  of the p-value for the number of minor alleles in the linear regression of BDR on the number of minor alleles, ethnicity, African local and global ancestry, and Native American local and global

ancestry. X-axis: Expected  $-\log_{10}$  p-values based on a uniform distribution. The black QQ plot shows all SNPs that passed QC filters (see methods). The shaded region is the 95% concentration band. The superimposed blue QQ plot shows only SNPs with a minor allele frequency  $> 5\%$ , indicating that the signal is driven by rare variants.



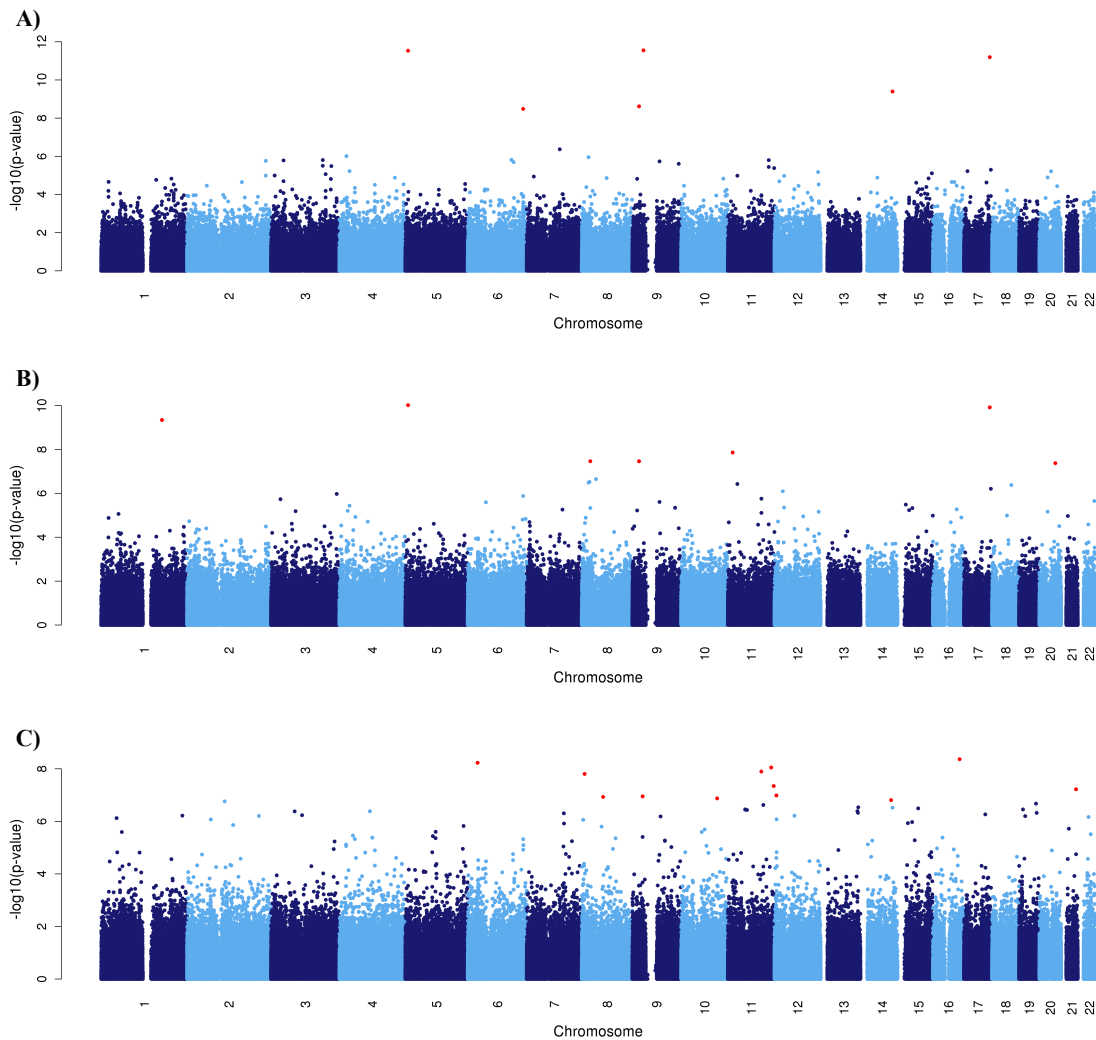
In all of GALA II, six SNPs had p-values lower than the permutation-based genome-wide significance cutoff of  $p < 1.7 \times 10^{-7}$  (Figure 3.3A). All of these six SNPs had a MAF  $< 5\%$  and had no individuals with BDR data who were minor homozygotes (Table 3.3). Four of these SNPs were singletons that we did not follow-up because we could not validate them without validating both the genotype and the BDR measurement. There were two genome-wide significant SNPs, rs8191725 and rs77441273, with more than one heterozygous individual with BDR data (genotype clusters in Appendix A). Both of these SNPs had a dose-response relationship that supported the hypothesis that the minor allele confers greater BDR (Table 3.3). The mean BDR was 9.91% and 10.0% in the major homozygotes and 20.4% and 33.6% in the heterozygotes for rs8191725 and

rs77441273, respectively. These relationships remained consistent even after removing heterozygotes with a BDR > 60, which indicates that the dose-response relationships are not entirely driven by individuals with extremely high BDR.

In the Puerto Rican and Mexican subsets of GALA II, 7 and 12 SNPs had p-values lower than the permutation-based genome wide significance cutoff, respectively (Figure 3.3 B and C). Three of the 7 SNPs that were genome-wide significant in the Puerto Ricans were also significant in all of GALA II. As in all of GALA II, all of the significant SNPs in the Puerto Ricans and Mexicans had a MAF < 5%. Again, most of these SNPs had very few heterozygotes with BDR data and we did not investigate them further because we could not validate them without validating both the genotype and the BDR measurement. Only one SNP in each of these subsets of GALA II had more than three minor homozygotes (genotype clusters in Appendix A). These SNPs, rs77977790 and rs71513949 for the Puerto Ricans and Mexicans, respectively, had dose-response relationships that supported the hypothesis that the minor allele confers greater BDR (Table 3.3). The mean BDR was 11.5% in the major homozygotes and 20.6% in the heterozygotes for rs77977790. For rs71513949, the mean BDRs were 7.4%, 14.8% and 17% in the major homozygotes, heterozygotes, and one minor homozygote, respectively. Again, these relationships remained consistent even after removing heterozygotes with a BDR > 60.

Since all of the significant SNPs in GALA II and its subsets were rare, identifying them through genotyping in the smaller GALA I population was unlikely. Indeed, custom TaqMan SNP Genotyping Assays for rs8191725, rs77441273, rs77977790 and rs71513949 were unable to cluster heterozygotes separately from major homozygotes.

**Figure 3.3. Genome-wide allelic associations with bronchodilator response in A) all GALA II, B) GALA II Puerto Ricans and C) GALA II Mexicans.** Y-axis:  $-\log_{10}$  of the p-value for the number of minor alleles in the linear regression of BDR on the number of minor alleles, ethnicity, African local and global ancestry, and Native American local and global ancestry. X-axis: Chromosome and position. Colors alternate for clarity. SNPs meeting a genome-wide significance cutoff of  $p < 1.6 \times 10^{-7}$  are colored in red.



Therefore, we attempted replication of these four SNPs *in silico* by imputing them in GALA I. We analyzed only the three SNPs that had an information score  $> 0.3$ . One of these, rs77441273, was significant in the GALA I Mexicans ( $p = 6.49 \times 10^{-4}$ ,  $\beta = 93.5$ ).

However, this imputed SNP was very rare and fairly poorly imputed in the GALA I Mexicans, having only a few possible heterozygote individuals and an info score = 0.366.

**Table 3.3. Genome-wide significant hits from allelic associations with BDR in GALA II.**

| Population  | SNP ID*            | MAF    | p-value | Beta  | Mean BDR (# individuals) |                |                |
|-------------|--------------------|--------|---------|-------|--------------------------|----------------|----------------|
|             |                    |        |         |       | 0 <sup>+</sup>           | 1 <sup>+</sup> | 2 <sup>+</sup> |
| All GALA II | <b>rs8191725</b>   | 0.815% | 3.3E-09 | 10.2  | 9.9 (1750)               | 20.4 (29)      |                |
|             | <b>rs77441273</b>  | 0.168% | 4.0E-10 | 23.6  | 10 (1776)                | 33.6 (6)       |                |
|             | <u>rs41313772</u>  | 0.028% | 2.8E-12 | 64.3  | 10 (1778)                | 73.9 (1)       |                |
|             | <u>rs4510053</u>   | 0.028% | 6.4E-12 | 63.2  | 10 (1778)                | 75.8 (1)       |                |
|             | <u>rs16879355</u>  | 0.028% | 2.9E-12 | 64.3  | 10.1 (1780)              | 75.8 (1)       |                |
|             | <u>rs115856718</u> | 0.028% | 2.4E-09 | 54.9  | 10.1 (1781)              | 67.7 (1)       |                |
| GALA II PR  | <b>rs77977790</b>  | 2.798% | 4.6E-10 | 9.5   | 11.5 (776)               | 20.6 (46)      |                |
|             | <u>rs77149876</u>  | 0.182% | 1.4E-08 | 32.5  | 11.9 (819)               | 44.1 (3)       |                |
|             | <u>rs115501901</u> | 0.182% | 4.2E-08 | 31.5  | 11.9 (819)               | 43.1 (3)       |                |
|             | <u>rs4510053</u>   | 0.061% | 1.2E-10 | 63.7  | 11.9 (819)               | 75.8 (1)       |                |
|             | <u>rs16879355</u>  | 0.061% | 9.6E-11 | 64.3  | 11.9 (822)               | 75.8 (1)       |                |
|             | <u>rs74667495</u>  | 0.061% | 3.5E-08 | 55.0  | 11.9 (822)               | 67.7 (1)       |                |
|             | <u>rs115856718</u> | 0.061% | 3.5E-08 | 54.9  | 11.9 (822)               | 67.7 (1)       |                |
| GALA II MX  | <b>rs71513949</b>  | 3.671% | 1.1E-07 | 6.9   | 7.4 (531)                | 14.8 (40)      | 17 (1)         |
|             | <u>rs116551936</u> | 0.264% | 5.9E-09 | 27.0  | 7.8 (566)                | 36 (3)         |                |
|             | <u>rs79889346</u>  | 0.175% | 6.0E-08 | -30.8 | 8 (568)                  | -21.8 (2)      |                |
|             | <u>rs74973995</u>  | 0.175% | 1.3E-08 | 32.6  | 7.8 (569)                | 39.9 (2)       |                |
|             | <u>rs115719051</u> | 0.175% | 1.0E-07 | 30.9  | 7.8 (570)                | 38.4 (2)       |                |
|             | <u>rs115428154</u> | 0.175% | 1.2E-07 | 30.6  | 7.8 (570)                | 38.4 (2)       |                |
|             | <u>rs77420108</u>  | 0.089% | 8.9E-09 | 46.9  | 7.9 (563)                | 53.1 (1)       |                |
|             | <u>rs114488285</u> | 0.088% | 4.3E-09 | 47.7  | 7.8 (568)                | 53.1 (1)       |                |
|             | <u>rs114475415</u> | 0.088% | 1.3E-07 | -42.4 | 8 (569)                  | -34.3 (1)      |                |
|             | <u>rs78708267</u>  | 0.087% | 1.6E-08 | 45.6  | 7.8 (571)                | 53.1 (1)       |                |
|             | <u>rs77253533</u>  | 0.087% | 4.5E-08 | 45.8  | 7.8 (571)                | 53.1 (1)       |                |
|             | <u>rs117004957</u> | 0.087% | 1.6E-07 | -42.2 | 8 (571)                  | -34.3 (1)      |                |

\*We attempted to replicate SNPs in **bold**

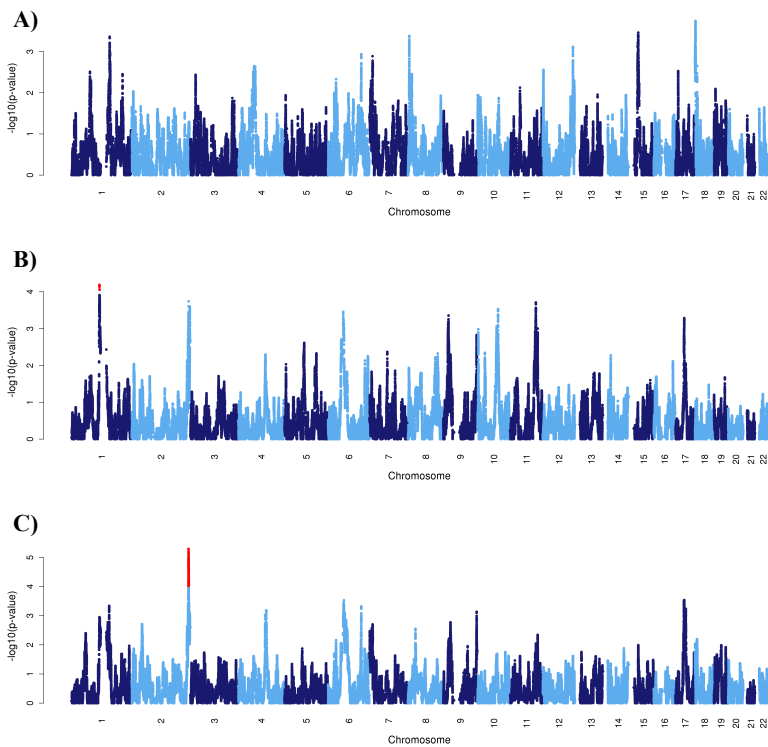
+Indicates the number of minor alleles

### **3.3.2. Admixture Mapping Supports the Association of Rare Variants with BDR in GALA II**

At each of the 568,037 loci, we tested for the association of each of three local ancestry components with BDR in 1,782 children with asthma from GALA II (Figure 3.4). We performed these tests in all of GALA II and separately in the subsets of 823 Puerto Rican children (Figure 3.5) and 572 Mexican children (Figure 3.6). In total, we identified 5 peaks that were significant using permutation-based cutoffs for each ancestry. Two of the significant peaks were strongest in the Puerto Ricans but also significant in all of GALA II. These two peaks also had signals for both Native American and European ancestry in the Puerto Ricans. On Chromosome 1, an increase in Native American ancestry was associated with a decrease in BDR in both the Puerto Ricans and all of GALA II. At the same locus, an increase in European ancestry was associated with an increase in BDR in Puerto Ricans. A similar pattern existed for the peak on Chromosome 2. On Chromosome 2, an increase in Native American ancestry was associated with an increase in BDR in both the Puerto Ricans and all of GALA II. At the same locus, an increase in European ancestry was associated with a decrease in BDR in the Puerto Ricans. The signal from these two peaks indicates that genetic variation at these loci is associated with BDR and differs according to the level of Native American and European ancestry at these loci, especially in the Puerto Ricans.

The other three peaks we identified were all driven by African ancestry. The peaks on Chromosome 1 and 8 were in the Puerto Ricans. For these peaks, an increase in African ancestry was associated with a decrease and increase in BDR, respectively. The Chromosome 1 peak did not overlap with the previously mentioned Native

**Figure 3.4. Genome-wide ancestry associations with bronchodilator response in all GALA II for A) African ancestry, B) Native American ancestry, and C) European ancestry.** Y-axis:



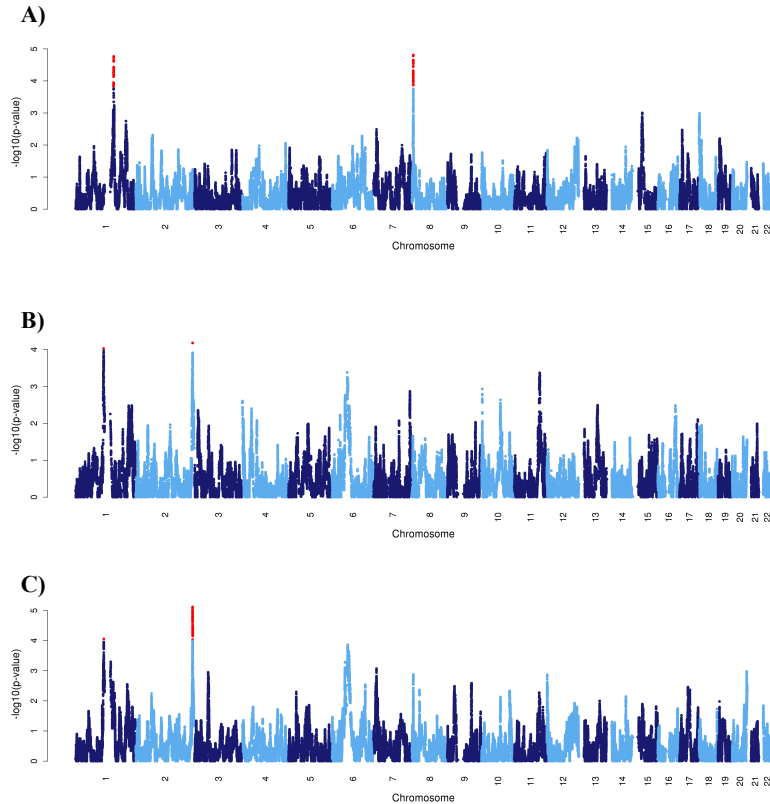
–log<sub>10</sub> of the p-value for the specified local ancestry in the linear regression of BDR on the specified local ancestry, corresponding global ancestry, and ethnicity. X-axis: Chromosome and position. Colors alternate for clarity. Significant loci are highlighted in red.

American/European ancestry Chromosome 1 peak. The third African ancestry peak was in the Mexicans on Chromosome 14. In this peak, an increase in African ancestry was associated with an increase in BDR. These peaks were all unique to the subset of GALA II in which they were identified.

We further investigated the admixture mapping peaks we found in two ways. First, we attempted to replicate all 5 admixture mapping peaks using existing GALA I data. None of the peaks replicated in GALA I ( $p > 0.05$ ). Second, we looked for significant allelic associations in GALA II under the peaks. For each peak, we used a Bonferonni correction for the number of SNPs tested under that peak. We tested for associations in all three subsets of GALA II. Two SNPs under the peak on Chr 1 from Native American/European ancestry, rs1281748 and rs1281743, were significantly

**Figure 3.5. Genome-wide ancestry associations with bronchodilator response in GALA II Puerto Ricans for A) African ancestry, B) Native American ancestry, and C) European ancestry.**

Y-axis:  $-\log_{10}$  of the p-value for the specified local ancestry in the linear regression of BDR on the specified local ancestry, corresponding global ancestry, and ethnicity. X-axis: Chromosome and position. Colors alternate for clarity. Significant loci are highlighted in red.

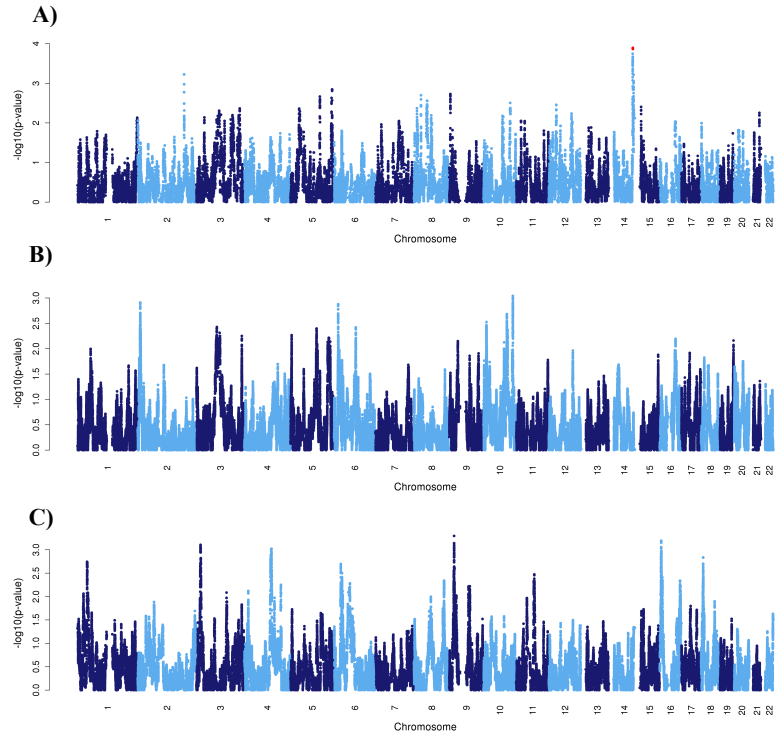


associated with BDR in the Mexicans ( $p = 8.8 \times 10^{-5}$  for both). These two SNPs were in LD and had MAFs of 0.26%. Being heterozygous for the minor allele at these SNPs was associated with an increase of 18.8% in BDR in the Mexicans. There were three heterozygote Mexican individuals for this SNP, all with a BDR between 25% and 28%. All three of these individuals were also heterozygous for African ancestry. We attempted to replicate these two SNPs by genotyping them in GALA I; neither SNP was significant. However, both SNPs were very rare: there were only 1 or 2 heterozygous individuals with BDR data in the GALA I Mexicans and 8 or 9 heterozygous individuals with BDR data in the GALA I Puerto Ricans for rs1281748 and rs1281743, respectively. Furthermore, the trend for both SNPs in the GALA I Puerto Ricans was the same as in

the GALA II Mexicans. In the GALA I Puerto Ricans, rs1281748 and rs1281743 had coefficients of 5.9 and 7.7, respectively ( $p=0.28$  and  $p=0.19$ , respectively).

**Figure 3.6. Genome-wide ancestry associations with bronchodilator response in GALA II Mexicans for A) African ancestry, B) Native American ancestry, and C) European ancestry.**

Y-axis:  $-\log_{10}$  of the p-value for the specified local ancestry in the linear regression of BDR on the specified local ancestry, corresponding global ancestry, and ethnicity. X-axis: Chromosome and position. Colors alternate for clarity. Significant loci are highlighted in red.



**Table 3.4. Significant Admixture Mapping Peaks**

| Population                 | Ancestry* | Coefficient <sup>+</sup> | Chr | Start Position | End Position | Length (kb) |
|----------------------------|-----------|--------------------------|-----|----------------|--------------|-------------|
| Puerto Rican & All GALA II | NAM / EUR | -2.58 / 2.05             | 1   | 116204807      | 117505312    | 1300.5      |
| Puerto Rican & All GALA II | NAM / EUR | 2.96 / -2.56             | 2   | 235202022      | 236278203    | 1076.2      |
| Puerto Rican               | AFR       | -2.83                    | 1   | 157995576      | 158687163    | 691.6       |
| Puerto Rican               | AFR       | 2.69                     | 8   | 5585682        | 6024650      | 439.0       |
| Mexicans                   | AFR       | 3.88                     | 14  | 98812269       | 98932579     | 120.3       |

\*NAM = Native American; EUR = European; AFR = African ancestries

+Coefficient is for locus with lowest p-value in peak



**Table 3.5. Summary of Top SNPs Identified in GWAS and Admixture Mapping**

| SNP ID*            | Chr | Position  | Gene                  | SNP Function | Functional Notes   |
|--------------------|-----|-----------|-----------------------|--------------|--|
| <b>rs8191725</b>   | 6   | 160429357 | IGF2R                 | Intronic     | Activation of TGFβ2 among other functions  |
| <b>rs77441273</b>  | 14  | 92959857  | SLC24A4               | Missense     | Organic ion transporter; associated with skin/hair/eye pigmentation  |
| <b>rs79777790</b>  | 1   | 176695479 | PAPPA2                | Intronic     | Cleaves IGF-binding protein 5 and is thought to be local regulator of IGF bioavailability; also interacts with SMAD4     |
| <b>rs71513949</b>  | 9   | 29938098  | --                    |              |  |
| <b>rs1281748</b>   | 1   | 116607254 | SLC22A15              | Intronic     | Organic ion transporter  |
| <b>rs1281743</b>   | 1   | 116610388 | SLC22A15              | 3'UTR        | Organic ion transporter  |
| <b>rs41313772</b>  | 9   | 32405589  | ACO1                  | Missense     | Regulates cellular iron homeostasis; also an aconitase   |
| <b>rs4510053</b>   | 17  | 74557681  | Uncharacterized locus |              |  |
| <b>rs16879355</b>  | 5   | 7893013   | MTRR                  | Missense     | Electron transferase that reactivates methionine synthase  |
| <b>rs115856718</b> | 9   | 19509957  | SLC24A2               | 3'UTR        | Sodium-calcium exchanger important for light adaptation in the retinal rod and cone                                      |
| <b>rs77149876</b>  | 11  | 14085131  | SPON1                 | Intronic     | Cell adhesion protein for vascular smooth muscle cell; highest expression in lung  |
| <b>rs115501901</b> | 20  | 46282708  | NCOA3                 | 3'UTR        | Nuclear receptor coactivator that interacts with nuclear receptors to enhance their transcriptional activation functions |
| <b>rs74667495</b>  | 8   | 24804373  | --                    |              |  |
| <b>rs116551936</b> | 6   | 28619532  | --                    |              |  |
| <b>rs79889346</b>  | 21  | 43306070  | C2CD2                 | 3'UTR        | Contains one calcium-dependent domain  |
| <b>rs74973995</b>  | 11  | 96962052  | --                    |              |  |
| <b>rs115719051</b> | 12  | 5606764   | --                    |              |  |
| <b>rs115428154</b> | 8   | 61773728  | CHD7                  | Intronic     | Probable transcription regulator associated with CHARGE syndrome   |
| <b>rs77420108</b>  | 11  | 125453419 | EI24                  | Missense     | Negative growth regulator through p53-mediated apoptosis   |
| <b>rs114488285</b> | 16  | 77472082  | --                    |              |  |
| <b>rs114475415</b> | 10  | 104525335 | C10orf26              | Intronic     |  |
| <b>rs78708267</b>  | 8   | 8256534   | --                    |              |  |
| <b>rs77253533</b>  | 11  | 132822741 | OPCML                 | Intronic     | Probably involved in cell contact  |
| <b>rs117004957</b> | 14  | 88761653  | KCNK10                | Intronic     | Rapidly activating outward rectifying potassium channel  |

\*SNPs that we attempted to replicate are indicated in **bold** text.

### 3.4. Discussion

We hypothesized that both common genetic variation and genetic variation more frequent on certain ancestral haplotypes contribute to differences in BDR. We tested this hypothesis by performing tests for allelic and ancestry associations with BDR in 1,782 Latino patients with asthma from across the US and Puerto Rico. We found evidence that rare variants are involved in BDR from tests for both allelic and ancestry associations. We found 22 rare variant alleles that were significantly associated with BDR in either all of GALA II, the Puerto Ricans alone, or the Mexicans alone. Four of these associations are promising candidates because they did not appear to be driven by a single individual with an extreme BDR. We also found 5 significant admixture mapping peaks. Since rare variants are often specific to one ancestral background<sup>25</sup>, these admixture mapping peaks may be driven by rare variants associated with BDR. In support of this idea, we found two rare SNPs under one of the admixture mapping peaks whose alleles were associated with BDR.

Sixteen of the 24 rare variants identified by either GWAS or admixture mapping fell in or near 15 unique genes and one was in an uncharacterized locus (Table 3.5). Many of these genes are plausible BDR genes including four ion transporters, two genes involved in transcriptional regulation, and several other genes involved in some sort of regulation or homeostasis. Two genes, which were identified from the four most promising allelic associations, are especially noteworthy. IGF2R has many functions including activation of the asthma gene TGFB2<sup>26</sup>. In addition, PAPP2 cleaves IGF-binding protein 5 and may be a local regulator of IGF bioavailability in some cases<sup>27</sup>. IGF

is involved in airway inflammation and remodeling<sup>28</sup>. Taken together, our results suggest that rare variants contribute to inter-individual variability in BDR.

Interest in the association of rare variants and complex diseases has piqued in recent years<sup>29,30</sup>. In fact, it is even possible that some signals identified from common variants in GWAS are driven by rare variants<sup>31</sup>. Although direct sequencing is the best approach to identify rare variants, the LAT array was designed to have better coverage of rare variants than typical genotyping arrays<sup>32</sup>. Admixture mapping can also be used to identify signals from rare variants, but will generally need to be followed up with sequencing to identify the causal ones. Even though studies of rare variants and complex diseases have become more common, replication of these rare variants is still very difficult for several reasons. First, rare variants are more likely to be population-specific than common variants<sup>33</sup>. Therefore, attempting replication across populations of different ancestries is unlikely to be fruitful. Second, studies have less power to detect rare variants than common variants. Thus, investigators need very large populations of the same ancestry for discovery and replication. These resources do not always exist, particularly for minority populations. Sequencing the regions around promising rare variants to identify other rare variants with similar effects and test all of these jointly may also improve power. However, sequencing large populations is still cost-prohibitive for many studies. Finally, rare variants may affect only a specific subset of subjects with a heterogeneous phenotype. In this case, this specific subset would need to be well represented in the replication population. However, researchers will not always know that they identified a rare variant because it acts only in the specific subset of subjects they studied. Since phenotypes are often measured in slightly different ways across studies or

subjects are recruited from different sources, ensuring that the phenotypes are consistent in the discovery and replication populations is difficult. As more data are collected and we develop a better understanding of the causes of variation in BDR, we will be able to replicate and understand the function of rare variants such as the ones identified in this study.

In this study, it was difficult to replicate the rare variants and ancestry signals that were associated with BDR for many of the reasons discussed above. The study of BDR in Latino populations is limited by a lack of data. GALA II and GALA I are the only large studies of Latinos with asthma that have data on BDR to our knowledge. In fact, only a few of the large genetic studies of asthma of any ethnicity have BDR data. Thus, GALA I was our best available replication population. GALA I has only 700 subjects with asthma and BDR data. This is simply not large enough to replicate most of the rare variants we identified in GALA II. In addition, there are differences in recruiting sites and the distribution of BDR between GALA I and GALA II. In GALA I, Mexicans were recruited from both the San Francisco Bay Area and Mexico City. GALA II Mexicans were recruited only from the continental US. The GALA I Mexicans had higher BDR than the Puerto Ricans<sup>9</sup>. In GALA II, BDR was higher in the GALA II Puerto Ricans than in the Mexicans because of the recruiting protocol. In addition, the protocols for administration of albuterol and measurement of BDR were slightly different between GALA I and GALA II (see Methods). These differences between GALA I and GALA II mean that variants that are only associated with very low or very high BDR will not replicate well between these two populations. Thus, these differences may explain the lack of replication of our admixture mapping peaks in GALA I.

We were also unable to replicate many of the rare variant allelic associations identified in GALA II. The fact that they are rare alone suggests that several of these variants may be false positives. This is particularly true for the variants whose associations were driven by a single individual with an extreme BDR. However, some of the rare variants with more than three heterozygotes seem very promising. For example, rs8191725 and rs77977790 both had associations that were consistent after removing BDR outliers, are in genes that have known ties to asthma (*IGF2R* and *PAPPA2*, respectively), and had plausible genotype clusters in GALA II (Appendix A). Unfortunately, we were not able to design genotype assays for these SNPs and they did not replicate using imputed genotypes. The differences between GALA I and GALA II may explain some of the lack of replication. However, the small sample size in GALA I also clearly limited our ability to replicate rare variants. The two SNPs in LD identified under one of the admixture mapping peaks, rs1281748 and rs1281743, are promising candidates with trends in GALA I similar to those in GALA II. However, there were no more than 11 heterozygous individuals for either of these SNPs in GALA I. This fact alone makes replication of these rare variants extremely unlikely. Replication using imputed genotypes is even more unlikely, since imputation of rare variants is more difficult than imputation of common variants. Thus, the apparent lack of replication of these promising SNPs from GALA II does not indicate that they are not important for BDR.

In conclusion, we identified four promising rare variants associated with BDR in GALA II through GWAS and admixture mapping. We also identified several other regions in which rare variants may be associated with BDR. Our findings suggest that

rare variants play an important role in BDR in Latino populations. Since rare variants are difficult to replicate, the variants and signals identified in this study will require follow-up through sequencing and functional studies in large populations. In addition, their relevance will need to be assessed across multiple racial and ethnic populations.

### 3.5. References

1. Nelson, H.S. (1995). Beta-adrenergic bronchodilators. *N Engl J Med* 333, 499–506.
2. Palmer, L.J., Silverman, E.S., Weiss, S.T., and Drazen, J.M. (2002). Pharmacogenetics of asthma. *American Journal of Respiratory and Critical Care Medicine* 165, 861–866.
3. Drazen, J.M., Silverman, E.K., and Lee, T.H. (2000). Heterogeneity of therapeutic responses in asthma. *Br. Med. Bull.* 56, 1054–1070.
4. Tantisira, K., and Weiss, S. (2008). The pharmacogenetics of asthma treatment. *Current Allergy and Asthma Reports* 9, 10–17.
5. Duan, Q.L., and Tantisira, K.G. (2009). Pharmacogenetics of Asthma Therapy. *Curr Pharm Des* 15, 3742–3753.
6. Contopoulos-Ioannidis, D.G., Alexiou, G.A., Gouvias, T.C., and Ioannidis, J.P.A. (2006). An empirical evaluation of multifarious outcomes in pharmacogenetics: beta-2 adrenoceptor gene polymorphisms in asthma treatment. *Pharmacogenet Genomics* 16, 705–711.
7. Duan, Q.L., Du, R., Lasky-Su, J., Klanderma, B.J., Partch, A.B., Peters, S.P., Irvin, C.G., Hanrahan, J.P., Lima, J.J., Blake, K.V., et al. (2012). A polymorphism in the thyroid hormone receptor gene is associated with bronchodilator response in asthmatics. *The Pharmacogenomics Journal* 1–7.
8. Hindorff, L., MacArthur, J., Wise, A., Junkins, H., Hall, P., Klemm, A., and Manolio, T. A Catalog of Published Genome-Wide Association Studies. *Genome.Gov*.
9. Burchard, E.G., Avila, P.C., Nazario, S., Casal, J., Torres, A., Rodriguez-Santana, J.R., Toscano, M., Sylvia, J.S., Alioto, M., Salazar, M., et al. (2004). Lower bronchodilator responsiveness in Puerto Rican than in Mexican subjects with asthma. *American Journal of Respiratory and Critical Care Medicine* 169, 386–392.
10. Naqvi, M., Thyne, S., Choudhry, S., Tsai, H.-J., Navarro, D., Castro, R.A., Nazario, S., Rodriguez-Santana, J.R., Casal, J., Torres, A., et al. (2007). Ethnic-Specific Differences in Bronchodilator Responsiveness Among African Americans, Puerto Ricans, and Mexicans with Asthma. *J Asthma* 44, 639–648.

11. Blake, K., Madabushi, R., Derendorf, H., and Lima, J. (2008). Population Pharmacodynamic Model of Bronchodilator Response to Inhaled Albuterol in Children and Adults With Asthma. *Chest* 134, 981–989.
12. Choudhry, S. (2004). Pharmacogenetic Differences in Response to Albuterol between Puerto Ricans and Mexicans with Asthma. *American Journal of Respiratory and Critical Care Medicine* 171, 563–570.
13. Lara, M., Akinbami, L., Flores, G., and Morgenstern, H. (2006). Heterogeneity of childhood asthma among Hispanic children: Puerto Rican children bear a disproportionate burden. *Pediatrics* 117, 43–53.
14. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19, 1655–1664.
15. Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*.
16. Browning, S. (2006). Multilocus Association Mapping Using Variable-Length Markov Chains. *The American Journal of Human Genetics* 78, 903–913.
17. Torgerson, D.G., Gignoux, C.R., Galanter, J.M., Drake, K.A., Roth, L.A., Eng, C., Huntsman, S., Torres, R., Avila, P.C., Chapela, R., et al. (2012). Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. *Journal of Allergy and Clinical Immunology* 1–19.
18. Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating local ancestry in admixed populations. *Am J Hum Genet* 82, 290–303.
19. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* 5, e1000529.
20. Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181.
21. Consortium, T.I.G.P., author, C., committee, S., Medicine, P.G.B.C.O., BGI-Shenzhen, Broad Institute of MIT and Harvard, Illumina, Technologies, L., Max Planck Institute for Molecular Genetics, Science, R.A., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073.
22. Casella, G., and Berger, R.L. (2002). *Statistical inference* (Duxbury Pr).
23. Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7, 781–791.

24. R Development Core Team. (2010). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).
25. Gravel, S., and Henn, B. (2011). Demographic history and rare allele sharing among human populations.
26. Dennis, P.A., and Rifkin, D.B. (1991). Cellular activation of latent transforming growth factor beta requires binding to the cation-independent mannose 6-phosphate/insulin-like growth factor type II receptor. *Pnas* 88, 580–584.
27. Yan, X., Baxter, R.C., and Firth, S.M. (2010). Involvement of pregnancy-associated plasma protein-A2 in insulin-like growth factor (IGF) binding protein-5 proteolysis during pregnancy: a potential mechanism for increasing IGF bioavailability. *J. Clin. Endocrinol. Metab.* 95, 1412–1420.
28. Yamashita, N., Tashimo, H., Ishida, H., Matsuo, Y., Arai, H., Nagase, H., Adachi, T., and Ohta, K. (2005). Role of insulin-like growth factor-I in allergen-induced airway inflammation and remodeling. *Cell. Immunol.* 235, 85–91.
29. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753.
30. Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature* 456, 18–21.
31. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D.B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biol.* 8, e1000294.
32. Hoffmann, T., Zhan, Y., Kvale, M., and Hesselson, S. (2011). Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* 98, 422-430.
33. Stephens, J.C. (2001). Haplotype Variation and Linkage Disequilibrium in 313 Human Genes. *Science* 293, 489–493.



## CHAPTER 4

### HAPLOTYPE INFERENCE ERROR VARIES ACROSS POPULATIONS

#### 4.1. Introduction

Haplotype inference methods have some intrinsic error that influences effect estimates in studies that rely on these methods<sup>1</sup>. These methods are used in case-control association studies to improve power and as a necessary precursor to population genetics studies or genotype imputation<sup>2,3</sup>. In the past, two approaches for statistical haplotype inference were commonly used: maximum likelihood in an expectation-maximization algorithm or the parsimony method<sup>4,5</sup>. In 2001, Stephens *et al.* introduced a haplotype inference method implemented in the program PHASE that incorporated knowledge from population genetics and coalescent theory and greatly reduced error compared to traditional methods<sup>6</sup>. Since the introduction of PHASE, several groups have published methods that make use of population genetics models and are computationally feasible for genome-wide data<sup>e.g. 7-9</sup>.

Although these modern haplotype inference methods use different statistical and computational techniques, they are all based on population genetics models. These models make assumptions about random mating and linkage disequilibrium (LD) patterns whose validity almost certainly varies across populations. For example, linkage disequilibrium is lower in older populations like those of African ancestry than in younger populations like those of European ancestry<sup>10</sup>. In addition, human populations are known to mate assortatively based on many factors, including socioeconomic status, religion and attitudes<sup>11</sup>. Recently, two studies have demonstrated that people mate

assortatively by ancestry<sup>12,13</sup>. If LD patterns and assortative mating vary between populations then the validity of the assumptions underlying haplotype inference methods also varies. Thus, the error from haplotype inference methods is also likely to vary across populations. In turn, bias in effect estimates will also vary across populations making effect estimates from studies that rely on haplotype inference incomparable. Haplotype inference error has been compared between African and European-ancestry populations<sup>2</sup>. However, to our knowledge, no studies have compared haplotype inference error across more than these two populations.

We hypothesized that haplotype inference error varies across human populations. To test this hypothesis and to determine the extent to which this error varies, we repeatedly sampled candidate-gene sized regions in trios from different HapMap populations and compared haplotype inference error in these populations. Then, we characterized this error by determining the extent to which measurable factors such as linkage disequilibrium and minor allele frequency contributed to this error.

## **4.2. Methods**

### **4.2.1. Data**

To test the hypothesis that haplotype inference error varies across populations, we used genome-wide single nucleotide polymorphism (SNP) data from four populations of trios from the HapMap Project Phase III<sup>14</sup>. These four populations included Utah residents of Northern and Western European ancestry (30 CEU trios), Yoruba in Ibadan, Nigeria (30 YRI trios), Maasai in Kinyawa, Kenya (28 MKK trios) and of Mexicans in the Los Angeles area (23 MEX trios). All trios were complete, consisting of a mother,

father and child. We sampled repeated candidate-gene sized regions from the 54,495 SNPs on Chromosome 1 that were genotyped in all four populations. We randomly sampled 100 20kb regions across Chromosome 1. For each of these regions in each population, we inferred haplotypes using four phasing methods and compared them to a gold standard as described below.

#### **4.2.2. Haplotype Inference and Gold Standard**

We inferred haplotypes using four publicly available phasing methods to ensure that our results were not an artifact of a specific method. We used PHASE, fastPHASE, BEAGLE, and SHAPE-IT<sup>6-9</sup>. Since all four methods rely on using unrelated individuals from the same population, we inferred haplotypes separately in the parents and children from each population. All methods were run with default settings.

We used information from each complete trio to obtain a gold standard haplotype for comparison using logic rules. For each SNP in each region, we determined which allele came from each parent. This is possible as long as at least one member of the trio is not heterozygous. For example, if a mother and child both have an AG genotype and the father has a GG genotype, we can determine that the haplotype the child obtained from the mother contains the A allele and the haplotype the child obtained from the father contains the G allele. Similarly, we know that the mother's haplotype that she gave to the child contains the A allele and the haplotype she did not pass on contains the G allele. Our gold standard makes the assumption that there is no recombination in the generation between the parents and the child in the specified region. In addition, we cannot identify errors in phasing of SNPs that are heterozygous in all three members of a trio.

### **4.2.3. Calculation of Error Proportion**

We compared our gold standard to the inferred haplotype from each phasing method to determine the number of individuals with incorrectly inferred haplotypes. Then, we calculated an error proportion within each region for each population and haplotype inference method. This error proportion was calculated as the number of individuals with incorrectly inferred haplotypes divided by the number of individuals with ambiguous haplotypes. Ambiguous haplotypes are haplotypes for which at least two of an individual's SNPs are heterozygous. Since the denominator of error proportion was ambiguous haplotypes, only haplotypes where each haplotype inference method could have made an error were counted towards the error proportion.

### **4.2.4. Calculation of Factors Related to Inference Error**

To determine the extent to which several measurable factors contribute to phasing error, we measured each of these factors (Table 4.1). These factors were measured either for each region across all populations, for each region within each population, or for each individual. We tested whether each of these factors were associated with inference error as described below.

### **4.2.5. Statistical Analysis**

We used t-tests to make pairwise comparisons between the mean error proportions in the populations for each phasing method. Our significance level was  $\alpha=0.0083$  based on a Bonferonni correction for 6 tests within each method.

**Table 4.1. Factors that may be associated with haplotype inference error**

| Factor                           | Measurement Level      | Definition   | Hypothesized Direction of Effect       |
|----------------------------------|------------------------|--|--|
| Number of SNPs                   | Across all populations | Number of SNPs in selected 20kb region   | Increasing with increasing error       |
| Distance between SNPs            | Across all populations | Mean number of base pairs between every pair of adjacent SNPs in the haplotype                                       | Decreasing with increasing error       |
| Genic/Intergenic                 | Across all populations | Genic if any SNP in the haplotype is contained within 2 kb upstream or downstream of a gene region according to NCBI | Increasing error in intergenic regions |
| Linkage disequilibrium           | Within each population | Mean pair-wise $r^2$ for all pairs of SNPs in haplotype and $r^2$ between the two SNPs at each end of the haplotype  | Decreasing with increasing error       |
| SNP minor allele frequency       | Within each population | Minimum MAF of all SNPs in haplotype   | Decreasing with increasing error       |
| Hardy-Weinberg Equilibrium (HWE) | Within each population | Maximum HWE chi-squared statistic of all SNPs in haplotype   | Increasing with increasing error       |
| Number of heterozygous SNPs      | For each individual    | Number of SNPs in the haplotype for which an individual is heterozygous (ambiguous)                                  | Increasing with increasing error       |

We used logistic regressions to determine whether each measured factor described above was associated with haplotype inference error. In addition, we used logistic regressions to test whether each region, individual and population was associated with haplotype inference error. Specifically, we regressed a binary error variable (*i.e.* yes/no error for each individual in each region) on each factor, one at a time. For minimum minor allele frequency, LD measured by  $r^2$  and mean between-SNP distance, we adjusted the scale of the factor so that the resulting odds ratios would be more meaningful to interpret. We multiplied linkage disequilibrium and minor allele frequency by 100 so that the odds ratios would be the change in the odds for an increase of 0.01 in these factors. We divided mean between-SNP distance by 100 so that the odds ratio would be the change in the odds for an increase of 100 base pairs in mean between-SNP distance. We used  $\alpha=1.2 \times 10^{-4}$  to determine significance level for these factors based on a Bonferroni correction for 413 tests.

To test whether the association of significant regions with haplotype inference error was driven by the factors we measured, we included principal components (PCs) created from the measured factors as covariates in the logistic regressions. We used all of the measured factors (Table 4.1) to create PCs. Since several of the measured factors were related, adjusting for all of the PCs might have caused regions that would be non-significant after adjustment for only a few PCs to become significant. Therefore, we determined the number of PCs to adjust for by examining the trends in p-values for each region after adjusting for varying numbers of PCs.

### **4.3. Results**

#### **4.3.1. Haplotype Inference Error Varies by Population**

The mean haplotype inference error proportion was highest in the African populations and lowest in the European population using all four phasing methods (Figure 4.1). Mean error proportions were consistently less than 1% different for PHASE and SHAPE-IT. The error proportions for PHASE and SHAPE-IT were the lowest overall. For SHAPE-IT, the mean error proportions across all regions were 5.1%, 7.5%, 8.3% and 9.2% for the European, Mexican, Yoruba and Maasai populations, respectively. Since BEAGLE is not designed for use in small sample sizes, it had the highest error proportions overall. The mean error proportions for BEAGLE were 13.4%, 15.3%, 20.6%, and 19.4% for the European, Mexican, Yoruba and Maasai populations, respectively. The mean error proportions for fastPHASE followed the same trend as for SHAPE-IT and PHASE and were intermediate between these and the BEAGLE error proportions.

**Figure 4.1. Haplotype Inference Error Varies by Population.** The distribution of error

proportion per region is

displayed by population and

phasing method. The y-axis

shows the error proportion that

is calculated by region as the

number of individuals with

incorrectly inferred haplotypes

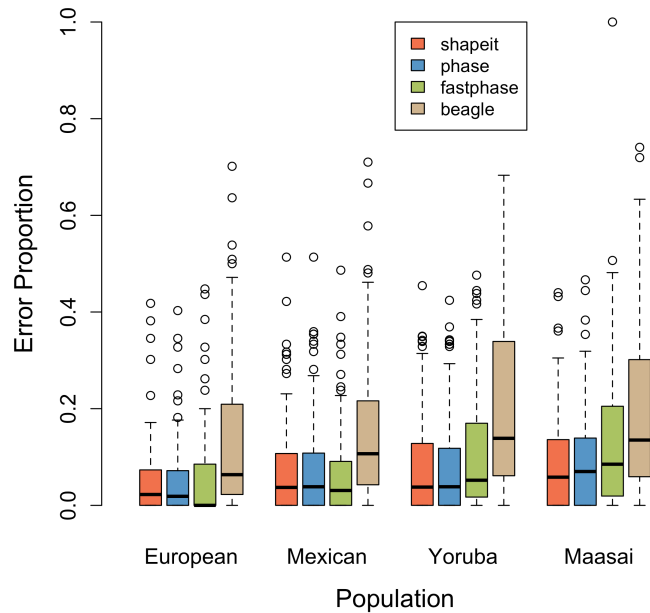
divided by the number of

individuals with ambiguous

haplotypes. The x-axis

displays the HapMap Phase 3

populations.



We used t-tests to test for significant differences in the mean error proportion between populations for each method (Table 4.2). The difference between the mean error proportions in the European population and the African population with the highest error proportion was significant for all four methods. For fastPHASE, the difference between the European population and both African populations were significant. In addition, the difference between the Mexicans and Maasai was significant for fastPHASE.

**Table 4.2. Significant Differences in Haplotype Inference Error by Population**

| Method    | Population | European         | Mexican         | Yoruba |
|-----------|------------|------------------|-----------------|--------|
| Shape-IT  | Mexican    | 0.0798           |                 |        |
|           | Yoruba     | 0.02104          | 0.5661          |        |
|           | Maasai     | <b>0.00267</b>   | 0.2347          | 0.5566 |
| Phase     | Mexican    | 0.07256          |                 |        |
|           | Yoruba     | 0.01681          | 0.5394          |        |
|           | Maasai     | <b>0.002206</b>  | 0.2224          | 0.5591 |
| fastPhase | Mexican    | 0.4345           |                 |        |
|           | Yoruba     | <b>0.001341</b>  | 0.0097          |        |
|           | Maasai     | <b>0.0001361</b> | <b>0.001118</b> | 0.4118 |
| Beagle    | Mexican    | 0.3906           |                 |        |
|           | Yoruba     | <b>0.002616</b>  | 0.02428         |        |
|           | Maasai     | 0.01163          | 0.07963         | 0.63   |

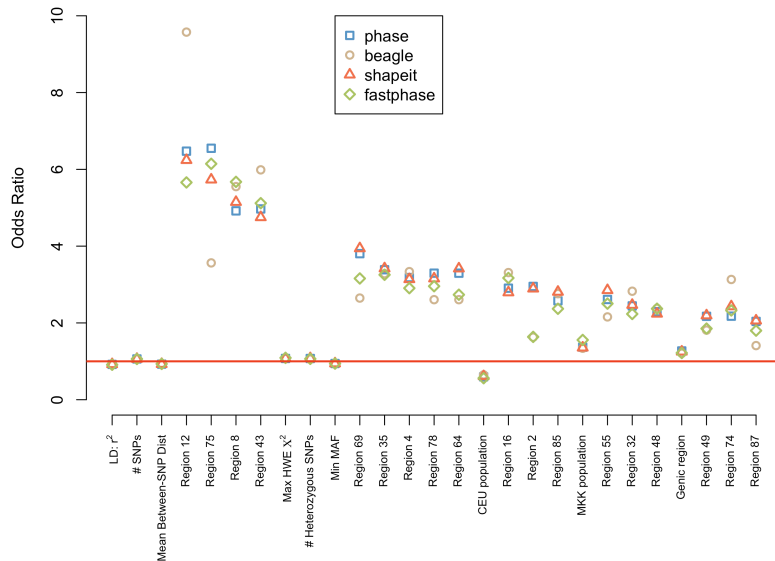
**Bold-face** p-values are lower than the Bonferonni alpha level of 0.0083 correcting for 6 tests within each method.

**Figure 4.2. Seven measured factors, two populations, and 18 regions are associated**

**with error.** The y-axis displays odds ratios for each variable from four logistic regressions of error on the variable. The four logistic regressions are of error from each of the phasing methods. The x-axis displays the 27 variables that were significantly associated with error from Phase after a Bonferonni correction for 413 tests (7 measured factors, 100 regions, 4 populations, and 302 individuals). The red line is at OR=1.

Variables with ORs above and below this line are associated with increased and decreased odds of error,

respectively. For linkage disequilibrium and minimum minor allele frequency the ORs are the change in odds for an increase of 0.2 in these variables.





### **4.3.2. Measurable Factors and Certain Regions are Associated with Haplotype Inference Error**

We tested the association of seven measured factors, each region, each population and each individual with error to further characterize the causes of error. All seven measured factors, 18 regions, and the European and Maasai populations were significantly associated with error from PHASE after correcting for 413 tests (Figure 4.2). Although we selected variables based on significant associations with PHASE error, the associations and directions of effect were consistent across all phasing methods. The direction of effect for the seven measured factors was consistent with what we hypothesized for all factors except genic/intergenic region (Table 4.1). For example, for every 0.01 increase in mean pairwise linkage disequilibrium  $r^2$  the odds of having an error from Phase was 0.92 times lower (Figure 4.2). The Maasai population had a 1.37 times the odds of having an error from Phase compared to the other populations. In contrast, the European population had 0.6 times the odds of having an error from Phase compared to the other populations. The 18 significantly associated regions all had odds ratios (ORs)  $> 1$ . These regions all had higher odds of error than the rest of the regions combined.

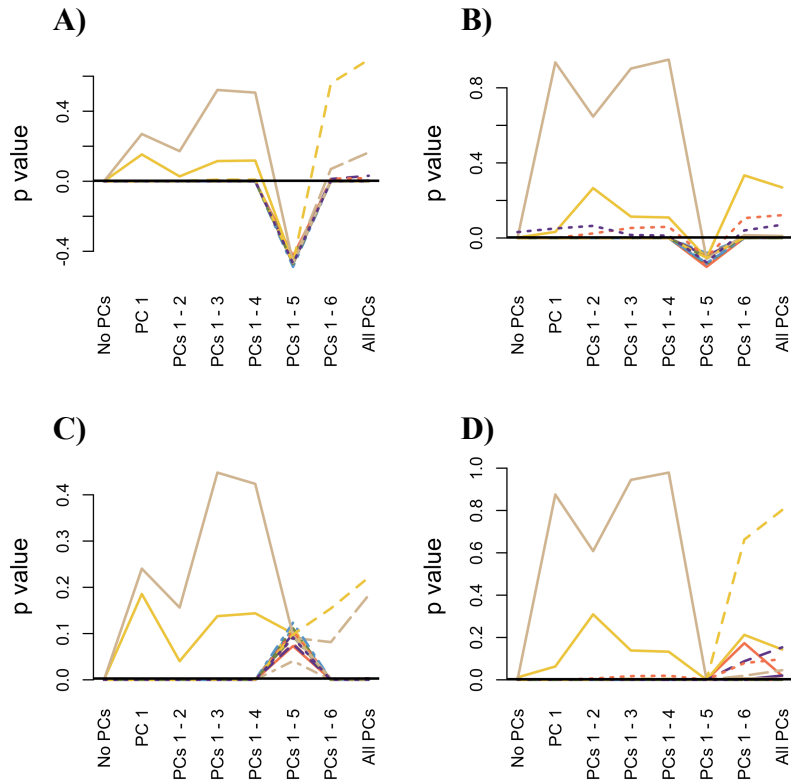
### **4.3.3. Most Previously Significant Regions are Associated with Error Even After Adjustment for Other Factors**

The association of regions with haplotype inference error may be driven by the measured factors that were also associated with error. Therefore, we built principal components from these factors and controlled for them in the 18 previously significant

**Figure 4.3. Trends in p-values for significant regions after adjusting for increasing numbers of PCs for A) Phase, B) Beagle, C) Shape-IT and D) fastPhase.**

Each plot shows the p-value for each region on the y-axis against the number of PCs that were adjusted for on the x-axis. A line connects the points for each of the 18 originally significant regions.

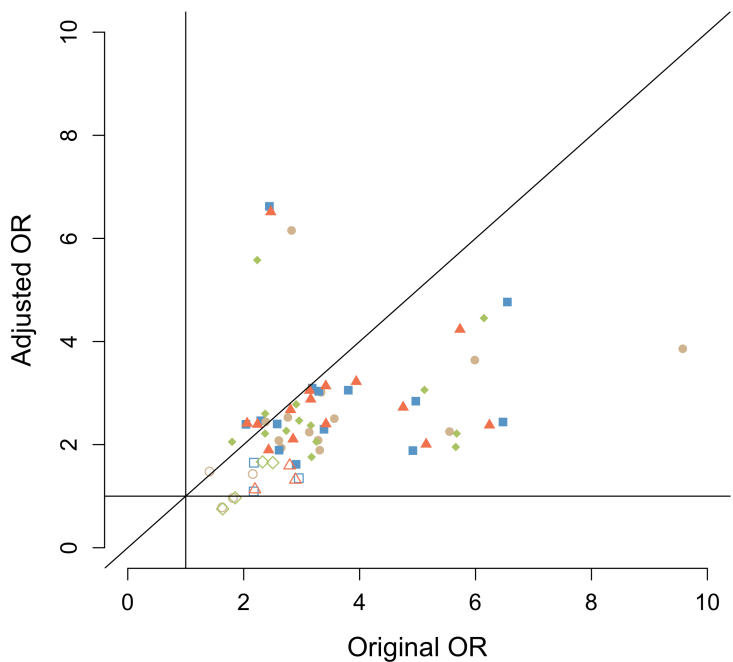
The thick horizontal black line is the alpha level after a Bonferonni correction for 18 tests. Adjusting for six or more PCs causes regions that became non-significant after adjusting for only one PC to become significant again.



logistic regressions of error on region. We determined the appropriate number of PCs to adjust for by examining the trends in p-values for each region after adjusting for varying numbers of PCs (Figure 4.3). We adjusted for PCs 1 through 5 because two regions that became non-significant after adjusting for only PC 1 became significant again when we adjusted for PCs 1 through 6 or all seven PCs. These regions may have become

significant again because there was collinearity between the PCs because of underlying collinearity in some of the input variables.

**Figure 4.4. Most Regions Remain Associated with Error After Adjusting for Other Measurable Factors.** X-axis: Original unadjusted OR from regression of error on region. Y-axis: Adjusted OR from regression of error on region, adjusting for PCs 1 and 2. Colors and shape indicate phasing method (Shape-IT=red/triangle, Phase=blue/square, fastPhase=green/diamond, Beagle=tan/circle). Filled shapes remain significant after



adjusting for PCs 1 and 2.

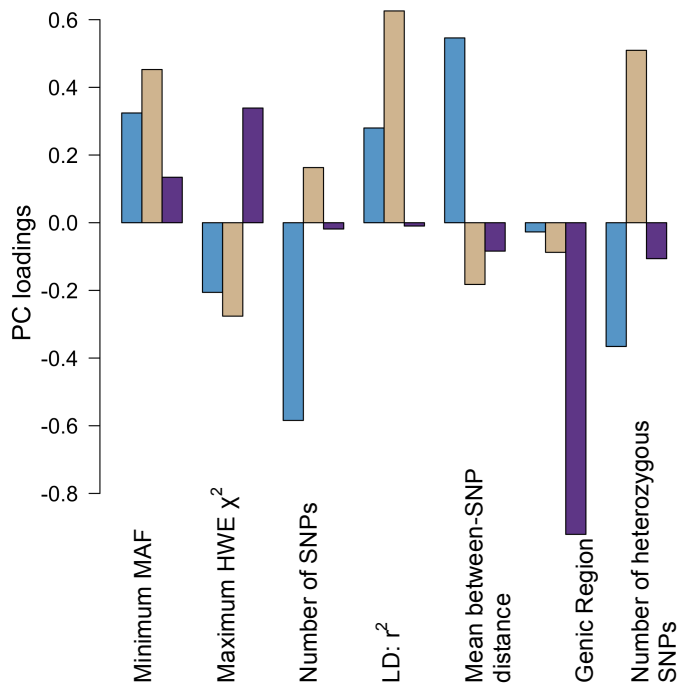
Horizontal and vertical lines are for ORs=1. The diagonal line is the identity line between the original and adjusted ORs. Points above and below this line had ORs that increased and decreased, respectively, after adjusting for PCs 1 and 2.

After controlling for PCs 1 through 5, 15 of the 18 previously significant regions were still significantly associated with error from PHASE or SHAPE-IT (Figure 4.4). Similarly, 14 of the previously significant regions were still significantly associated with

error BEAGLE or fastPHASE. 15 of the ORs decreased after adjusting for PCs 1 through 5 for all the methods. The mean change in OR was a decrease ranging from 0.76 for SHAPE-IT to 0.96 for BEAGLE. PCs 1, 2 and 3 were significantly associated with error from Phase for all regions ( $p < 0.005$ , data not shown). The largest driving factors of PCs 1, 2 and 3 were number of SNPs / mean between-SNP distance, LD / MAF / number of heterozygous SNPs, and genic region, respectively (Figure 4.5).

**Fig 4.5. PCs 1, 2 and 3 are driven by number of SNPs / between-SNP distance, LD / MAF / number of heterozygous SNPs and genic region, respectively.**

The y-axis shows the loading for each variable on the x-axis. Colors indicate PC (blue=PC1, tan=PC2, purple=PC3).



#### 4.4. Discussion

Overall, our results show that haplotype inference error varies across four HapMap Phase III populations. Furthermore, although this error is associated with seven of the factors we measured, there are regions whose high error cannot be explained by

these factors. We found that haplotype inference error was highest in two African populations, intermediate in a Mexican population, and lowest in a European population using four different haplotype inference methods. These results are sensible given that haplotype inference methods rely on LD, which is generally lower in African populations than either European or Mexican populations<sup>10</sup>. Unsurprisingly, we found that decreasing LD was associated with higher haplotype inference error.

Other factors we found to be associated with error are also sensible, including minor allele frequency, the number of SNPs, between-SNP distance, the number of heterozygous SNPs per individual, non-random mating, and genic vs. intergenic regions. Decreasing minor allele frequency was associated with higher haplotype inference error. This is sensible since when there are rare alleles there will be rare haplotypes that are more difficult for the phasing methods to infer. Since the size of the regions was fixed at 20kb, the number of SNPs and between-SNP distance are inversely related. An increasing number of SNPs and an increasing between-SNP distance were associated with increasing and decreasing error, respectively. Even by chance, the error proportion should increase with an increasing number of SNPs because it is harder for phasing methods to infer every single SNP correctly. We observed the same trend for the number of heterozygous SNPs per individual, since this represents the number of SNPs that need to be inferred for this individual. Non-random mating, which is assumed by the haplotype inference methods we used, can cause an increase in HWE  $\chi^2$ <sup>15</sup>. Sensibly, increasing HWE  $\chi^2$  was associated with increasing error. Finally, we found increasing error in genic regions despite our hypothesis that error would be higher in intergenic regions due to

lower levels of LD<sup>16</sup>. This higher error in genic regions may be due to other factors acting in these regions, either measured or not.

Overall, we found that haplotype inference error varies across populations and is associated with certain factors in a way that is consistent with previous knowledge. Even though we identified seven factors that were associated with haplotype inference error, these factors did not explain the high inference error in 14 regions. Our results suggest that there are other factors that we have not measured that contribute to high haplotype inference error in certain regions of the genome.

The differences we found in haplotype inference error across populations have important implications for genetic association studies of haplotypes and imputed SNPs. Many modern genotype imputation methods depend on inferring haplotypes before imputing SNPs<sup>3</sup>. Therefore, error in haplotype inference will cause misclassification in both haplotypes and imputed genotypes. This misclassification will cause bias in the effect estimates from these haplotypes or imputed genotypes<sup>17</sup>. Since the level of haplotype inference error varies across populations, the resulting levels of misclassification and bias in effect estimates will also vary across populations. Therefore, our results imply that differences in effect estimates across populations from haplotypes or imputed SNPs may be caused by haplotype inference error rather than true differences in effect estimates. Investigators should use caution when interpreting differences in these effect estimates.

Most published studies of haplotype inference error are methods comparisons that use simulated data or data from European populations. Although we used four different haplotype inference methods, we did so to make sure our results were not an artifact of

one phasing method. To our knowledge, this study is the first to examine differences in haplotype inference error across more than two populations. The fact that we used trios creates a biased sampling scheme because we can't evaluate error at SNPs that are heterozygous in all three individuals and we have more information to determine the child's haplotype than the parents'. However, using real populations similar to those used in association studies was critical to determining whether differences in error exist across populations and what the magnitude of these differences is.

In conclusion, the results of this study show that haplotype inference error varies across four HapMap Phase III populations. Furthermore, this error is associated with but not fully explained by factors we expected to cause error. Our study suggests that variation in haplotype inference error across populations may cause differences in effect estimates from haplotype association studies and imputed SNPs across populations. Furthermore, investigators cannot rely on the factors we expected to be associated with haplotype inference error to predict the regions that will have high error.

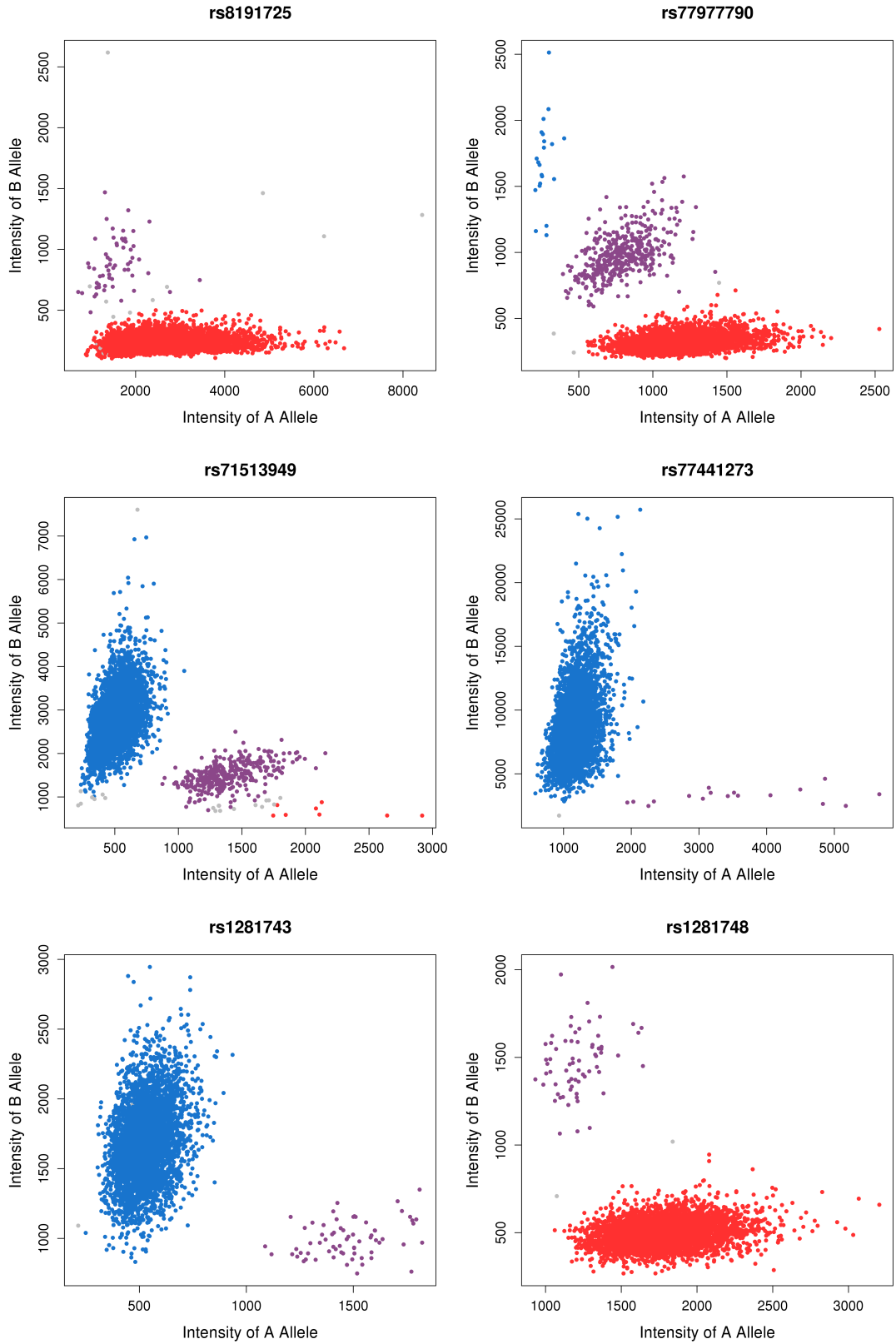
#### **4.5. References**

1. Li, W., and Gregersen, P.K. (2003). Reconstructing haplotypes in pedigrees: Importance of parental information. *Am. J. Med. Genet.* *124A*, 107–109.
2. Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., et al. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* *78*, 437–450.
3. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* *5*, e1000529.
4. Excoffier, L., and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* *12*, 921–927.

5. Clark, A. (1990). Inference of Haplotypes From Pcr-Amplified Samples of Diploid Populations. *Mol Biol Evol* 7, 111–122.
6. Stephens, M., Smith, N., and Donnelly, P. (2001). A New Statistical Method for Haplotype Reconstruction from Population Data. *The American Journal of Human Genetics* 68, 978–989.
7. Delaneau, O., Coulonges, C., and Zagury, J.-F. (2008). Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* 9, 540.
8. Browning, S., and Browning, B. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics* 81, 1084–1097.
9. Scheet, P., and Stephens, M. (2006). A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics* 78, 629–644.
10. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., Vanliere, J.M., Fung, H.-C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451, 998–1003.
11. Evans, D.M., Gillespie, N.A., and Martin, N.G. (2002). Biometrical genetics. *Biol Psychol* 61, 33–51.
12. Risch, N., Choudhry, S., Via, M., Basu, A., Sebro, R., Eng, C., Beckman, K., Thyne, S., Chapela, R., Rodriguez-Santana, J.R., et al. (2009). Ancestry-related assortative mating in Latino populations. *Genome Biol* 10, R132.
13. Sebro, R., Hoffman, T.J., Lange, C., Rogus, J.J., and Risch, N.J. (2010). Testing for non-random mating: evidence for ancestry-related assortative mating in the Framingham heart study. *Genet. Epidemiol.* 34, 674–679.
14. Consortium, T.I.H.3. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52.
15. Hartl, D.L. (2000). *A primer of population genetics* (Sinauer Associates Inc).
16. Eberle, M.A., Rieder, M.J., Kruglyak, L., and Nickerson, D.A. (2006). Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. *PLoS Genetics* 2, e142.
17. Kopec, J.A., and Esdaile, J.M. (1990). Bias in case-control studies. A review. *Journal of Epidemiology & Community Health* 44, 179–186.



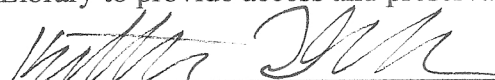
## APPENDIX A. GENOTYPE CLUSTERS FOR TOP HITS IN GALA II



## **Publishing Agreement**

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

  
\_\_\_\_\_  
Author Signature

6/11/12  
Date