IMPUTATION-BASED GENETIC ASSOCIATION ANALYSIS OF COMPLEX TRAITS IN ADMIXED POPULATIONS

Qing Duan

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Bioinformatics and Computational Biology in the School of Medicine.

Chapel Hill
2016

Approved by:

Terry S. Furey

Ethan M. Lange

Yun Li

Karen L. Mohlke

Kari E. North

ABSTRACT

Qing Duan: Imputation-Based Genetic Association Analysis of Complex Traits in Admixed Populations
(Under the direction of Yun Li)

Genetic association studies in admixed populations have drawn increasing attention from the genetic community, as performing association analysis in diverse populations allows us to gain deeper understanding of the genetic architecture of human diseases and traits. However, population stratification due to admixture poses special challenges. To address the challenges, I conducted the following studies from the perspectives of enhancing genotype imputation quality and providing proper treatment of local ancestry in the association analysis.

First, I provided a new resource of marker imputability information with commonly used reference panels to guide the choice of reference and genotyping platforms. To be specific, I systematically evaluated marker imputation quality using sequencing-based reference panels from the 1000 Genomes Project and released the information through a user-friendly and publicly available data portal. This is the first resource providing variant imputability information specific to each continental group and to each genotyping platform.

Second, I established a paradigm for better imputation in African Americans using study-specific sequencing based reference panels. I built an internal reference panel consisting of variants derived from the NHLBI Exome Sequencing Project for African American subjects, which significantly increased effective sample size comparing with that from the 1000 Genomes Project. No loss of imputation quality was observed using a panel built from phenotypic

extremes. In addition, I recommended using haplotypes from Exome Sequencing Project alone or concatenation of the two panels over quality score-based post-imputation selection or IMPUTE2's two-panel combination.

Finally, I proposed a robust and powerful two-step testing procedure for association analysis in admixed populations. Through extensive numeric simulations, I demonstrated that our testing procedure robustly captures and pinpoints associations due to allele effect, ancestry effect or the existence of effect heterogeneity between the two ancestral populations. In particular, our testing procedure is more powerful in identifying the presence of effect heterogeneity than traditional cross-product interaction model. I further illustrated its usefulness by applying the two-step testing procedure to test for the association between genetic variants and hemoglobin trait in African American participates from CARe.

Taken together, the above studies guide genotype imputation practice and substantially improve the power of imputation-based genetic association studies in admixed populations, leading to more accurate discovery of disease-associated variants and ultimately better therapeutic strategies in admixed populations.

To my father, mother and Hexuan

ACKNOWLEDGEMENTS

My research projects would not have been completed without the support of many people. It is a great pleasure to convey my gratitude to those who made this thesis possible.

First of all, I would like to express my profound gratitude to my adviser, Dr. Yun Li, for her invaluable mentorship. Her generous guidance, support and encouragement deeply inspired me and made this dissertation possible.

My sincere gratitude goes to Dr. Karen Mohlke, Dr. Leslie Lange, Dr. Kari North, Dr. Ethan Lange and Dr. Terry Furey for their invaluable advice and support in many ways.

I am truly grateful to my group colleagues for creating a stimulating, inspiring and friendly environment. Many thanks go in particular to Dr. Eric Yi Liu, Dr. Zheng Xu and Dr. Ying Wu who has been especially helpful to me in bringing these projects to completion.

My deepest gratitude goes to my beloved parents and husband. Their love, caring, understanding and support accompanied me and encouraged me throughout my study.

Finally, I would like to extend my best regards to all of those who helped and supported me in various ways throughout my PhD study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1000G	1000 Genomes Project
AIM	Ancestry informative makers
ARIC	Atherosclerosis Risk in Communities
CARDIA	Coronary Artery Risk Development in Young Adults
CARe	Candidate-gene Association Resource
CLHNS	Cebu Longitudinal Health and Nutrition Survey
ESP	Exome Sequencing Project
GWAS	Genome Wide Association Studies
HCHS/SOL	Hispanic Community Health Study / Study of Latinos
HMM	Hidden Markov Model
IBD	Identity-by-Descent
JHS	Jackson Heart Study
LD	Linkage Disequilibrium
LDL	Low-Density Lipoprotein
MAF	Minor Allele Frequency
MESA	Multi-Ethnic Study of Atherosclerosis
MI	Myocardial Infarction
QC	Quality Control
SE	Standard Error of the Mean
SNP	Single Nucleotide Polymorphism
WHI	Women's Health Initiative
WHISP	WHI Sequencing Project

CHAPTER 1: MOTIVATION AND SIGNIFICANCE

In this document, I will discuss computational resources and statistical approaches for facilitating genetic association studies with complex human traits in admixed populations. This section provides an overview of the motivation and significance of this study.

Genome-wide association studies (GWAS) have been successful in improving our understanding of the genetic basis of numerous heritable diseases and quantitative traits (Visscher *et al.*, 2012). They have been useful in identifying genes associated with complex traits in various relevant biological pathways (Visscher *et al.*, 2012). Although GWAS have initially been performed with individuals of European ancestry, the field has expanded to non-European populations (Rosenberg *et al.*, 2010). Performing genetic association analysis in diverse populations allows us to gain deeper understanding of the genetic architecture of human diseases and traits, through assessing the generalizability of risk variants (G. K. Chen *et al.*, 2010; Ioannidis *et al.*, 2004), narrowing down the location of the functional variants over the risk region (International HapMap, 2005) or identifying novel disease loci which are absent or in low frequency in European population (Rosenberg *et al.*, 2010).

In the US, genetic association studies in admixed populations have been receiving increasing attention, whereas it is challenged by the complex local ancestry structure resulted from admixture process, where gene flow occurs between two or more distinct populations (ancestral populations). Consequently, admixed chromosomes can be viewed as mosaic segments (local ancestry) originating from each of the ancestral populations (Shriner, 2013). Due to the

presence of local ancestry, many methods developed in homogeneous populations needs to be modified before they could be applied to admixed populations. To address the challenges, I perform three studies from the aspects of enhancing genotype imputation quality with a focus on rare variants and applying proper treatment of local ancestry information in association analysis.

Chapter 2 provides a literature review of genotype imputation and association analysis in admixed populations. To be specific, for genotype imputation, I will review several commonly used genotype imputation methods, how the methods are tuned to accommodate the unique LD structure in admixed populations and the practical guidelines for performing genotype imputation in genetic association studies with a focus on improving rare variants imputation quality; for association analysis in admixed populations, I will review methods used for local ancestry inference and statistical methods that are adopted in performing association studies in admixed populations.

Chapter 3 discusses a new resource of marker imputability information based on the commonly used reference panels from the 1000 Genomes Project. Marker imputability information is highly desirable to guide study design, to prioritize imputable markers and to serve as a post-imputation quality control. However, there is no direct access to this information without performing genotype imputation. I fill in this gap by providing marker imputation accuracy information of four major continental groups through a user-friendly publicly available data portal. This is the first study to provide genome-wide high resolution profiling of variants imputability. This imputability information will be very useful to association studies in diverse populations including admixed populations.

Chapter 4 shows as a proof-of-principle that imputation quality in low frequency variants of African American samples can be substantially enhanced by using sequencing-based study-

specific reference panels. Also I provide guidelines regarding the optimal manner to maximize information from internal and external reference. Increasing number of medical studies includes deep sequencing of a fraction of the participants as part of the study design. Imputation using these internally built reference panel can potentially achieve better performance, particularly at low frequency variants, than using a publicly available panel, e.g., from 1000G, due to better match of ancestry, larger effective sample size and disease status (Fridley *et al.*, 2010). However, limited studies have been performed to investigate the usefulness of study-specific reference panels in imputation, as compared with that from the public domain and none has been performed in African American sample. The study in Chapter 4 serves as a proof-of-principle to show the feasibility and gain of using the internal reference. Additionally, this is the first study that addresses the potential influence on imputation performance by using a reference panel with sample ascertainment bias, which is common in re-sequencing-based studies. Moreover, this is the first study that explores the optimal usage of the combined information from internal and external reference.

Chapter 5 proposes a testing procedure to model and test allele and ancestry effect jointly taking into account of effect heterogeneity in association analysis with African American sample. Historically, genetic linkage studies and admixture mapping have been treated as distinct methods and are performed separately in samples from admixed populations. It has been shown that the two methods contain independent information (Seldin *et al.*, 2011). Thus, jointly modeling allele and local ancestry effect would potentially increase statistical power (Pasaniuc *et al.*, 2011). In Chapter 5, I propose a robust and powerful two-step testing procedure for association analysis in admixed populations to robustly capture and to identify associations due to allele effect, ancestry effect and, particularly, the existence of effect heterogeneity between the

two ancestral populations. It is noteworthy that, by taking advantage of the inferred joint distribution of allele and ancestry, this method is more powerful than the traditional interaction model when effect heterogeneity presents.

CHAPTER 2: LITERATURE REVIEW

2.1 Genome-wide association studies in non-European populations

Many genetic variants associated with human diseases and traits have been successfully identified and replicated in Europeans (Locke *et al.*, 2015; Loth *et al.*, 2014). Nonetheless, European population only consists of a proportion of human genetic variation, which may have different properties, such as allele frequencies, from those in other populations (Coram *et al.*, 2013; Dhandapany *et al.*, 2009; Myles *et al.*, 2008). It is insufficient to study a single population to achieve the goal of fully uncovering the genetic architecture of complex human diseases and traits. Consequently, genome-wide association studies have been expanding into diverse populations such as Asian, African American and Hispanic populations (Franceschini *et al.*, 2013; Kooner *et al.*, 2011; Norris *et al.*, 2009).

In non-European populations, the characteristics of genetic variation, maybe different from that in Europeans, provide a unique opportunity for gene mapping and refining our understanding of genetic variations. First, a genetic variant may be fixed for different alleles among diverse populations. For example, DARC null allele for white blood cells is close to fixation in sub-Saharan Africa, whereas the other wild-type allele is fixed in non-African populations (Lautenberger *et al*., 2000; Tournamille *et al*., 1995). Similar examples include SLC24A5 for skin pigmentation (Lamason *et al.*, 2005) and APOL1 for kidney disease (Genovese *et al.*, 2010). The identification of such genetic variants would have been difficult if not impossible without association studies in African American populations. Second, even when

the same risk allele presents across populations, the risk allele frequency may vary. Risk alleles in certain populations with higher allele frequency may be easier to be identified in association analysis (McCarthy & Hirschhorn, 2008; Teo *et al.*, 2009). Third, the strength of linkage disequilibrium (LD) and the distance over which it extends may differ in different populations. It is observed that the extent of LD in a Nigerian population is markedly far less compared with that in a north-European population (D. E. Reich *et al.*, 2001). The short-range LD in some populations may offer opportunity for fine mapping (Gabriel *et al.*, 2002). Fourth, heterogeneity in allelic effects may present in distantly related populations. Causal variants may differ in the marginal allelic effect across populations as a result of differential environmental exposure (Morris, 2011). Thus some risk variants are more easily detected by using samples from certain relevant populations (Rosenberg *et al*., 2010). Finally, diverse populations facilitate the study of rare variant associations. Because of the substantially unexplained proportion of heritability as well as the drop in the whole-genome sequencing cost, many association studies have been focusing on rare variants. As the results of recent mutational events, rare variants are more likely to be geographically restricted (The 1000 Genomes Project *et al.*, 2012). Therefore, different populations may have distinct pool of rare variants, which emphasizes the necessity of studying rare variant associations in diverse populations.

Despite of the advantage of genetic analysis in non-European populations, challenges have been the lack of well-designed genotyping platforms and the availability of appropriate reference panels for genotype imputation (Jallow *et al.*, 2009). Initial tagSNP selection and genotyping chip design are based on HapMap CEU panel from the International HapMap Project focusing on European populations (International HapMap *et al.*, 2007; Sudmant *et al.*, 2015), which may result in ascertainment bias. In addition, the tagSNPs used to be selected based on LD

in CEU to maximize the number of markers "covered" by a tagSNP (in adequate LD with the tagSNP measured by r^2) (Carlson *et al.*, 2004). The coverage of genomic region can be reduced when the same tagSNP is applied in non-European populations due to the LD structure differences. Nevertheless, more recent high-throughput genotyping chips are denser and less biased towards European ancestral populations. Additionally, recent chips have been designed with worldwide populations in mind. For example, Affymetrix Axiom Genome-Wide Population-Optimized Human Array is customized for Caucasian, Asian and West African populations. Recently, Illumina collaborating with PAGE, CAAPA, and T2D-Genes Consortia has developed a multi-ethnic genotyping array, which adopts a novel tagSNP selection algorithm to optimize imputation accuracy across diverse populations (Gignoux, 2015). Moreover, the 1000 Genomes Project, a large-scale international collaboration, has substantially accelerated sequencing-based genetic association studies. The 1000 Genome Project aims to produce an extensive public catalog of human genetic variation by conducting whole genome sequencing of over 2500 individuals from 26 continental groups (The 1000 Genomes Project *et al.*, 2010; The 1000 Genomes Project *et al.*, 2012; The 1000 Genomes Project *et al.*, 2015). Sequencing uncovers genetic variants in diverse populations with little ascertainment bias issue. The comprehensive reference panels provided by the 1000 Genome Project increase the chance of matched LD patterns between the study sample and reference haplotypes, thereby, permitting better imputation in diverse populations from Africa, Asia, America and Europe.

In the United States, marked difference in disease prevalence has been reported between European Americans and admixed populations such as African Americans, including CRP (D. Reich *et al.*, 2007), prostate cancer (Freedman *et al.*, 2006), hypertension (Zhu *et al.*, 2005), type II diabetes (Elbein *et al.*, 2009) and obesity (Cheng *et al.*, 2009). As genetic factors may in part

account for the differences, these observations motivate the search for genetic loci, which contribute to the disease disparity in admixed populations.

Different from ancestrally homogeneous populations, such as Europeans, Africans or Asians, admixed populations have ancestry from more than one populations due to admixture process (Shriner, 2013). The chromosomes of admixed populations can be viewed as mosaic segments with different parental origins, leading to variable "global ancestry", the individual ancestry proportion, as well as different "local ancestry", the ancestral origin at a particular locus, across the genome (Bryc *et al.*, 2010; Silva-Zolezzi *et al.*, 2009; S. Wang *et al.*, 2008).

Therefore, the LD patterns in admixed populations are complicated with a fine scale of ancestral LD and a coarse scale of admixture LD. It imposes challenges for genotype imputation, as there are no large and closely matched reference panels from the 1000 Genomes Projects (The 1000 Genomes Project *et al.*, 2015). To meet the challenges, imputation methods have been modified to accommodate admixed populations and practical guidelines have been proposed, which will be reviewed in section 2.2.

Due to the genetic composition of multiple ancestries, population stratification is an intrinsic issue in admixed populations. The underlying population structure may give rise to spurious associations or false negative signals (Rosenberg & Nordborg, 2006). Consequently, computational tools and statistical methods have been developed aiming to resolve this issue in admixed populations, which will be reviewed in section 2.3.

2.2 Genotype imputation in admixed populations

Genotype imputation is an approach to predict genotypes for markers that are not experimentally genotyped in a study sample. With genotype imputation, one would obtain the genotype of a dense set of markers cost-effectively by imputing reference haplotypes into a study

sample that is moderately genotyped at a subset of the reference markers (Y. Li *et al.*, 2010; Marchini & Howie, 2010). Genotype imputation has become a standard practice in GWAS, which markedly enhances statistical power (Spencer *et al.*, 2009), facilitates fine mapping (Scott *et al.*, 2007) and enables meta-analysis (de Bakker *et al.*, 2008). Genotype imputation assumes that "unrelated" individuals could share identical by descent chromosome segment (IBD), i.e., identical short stretches of nucleotide sequences inherited from distant common ancestors. Conceptually, genotype imputation works by identifying the IBD segment for a study sample from the pool of reference haplotypes, based on the genotyped markers overlapping between reference and study sample. Then the genotype of the untyped markers in the study sample can be obtained by copying the corresponding marker genotype from the shared reference haplotype.

Currently, in genetic community, MaCH-Admix (E. Y. Liu *et al.*, 2013), MaCH/minimac (B. Howie *et al.*, 2012; Y. Li *et al.*, 2010) and IMPUTE2 (B. N. Howie *et al.*, 2009) are the widely used imputation software. These methods are highly accurate with adequate computational efficiency and have facilitated genotype imputation in many large-scale genome-wide genetic studies.

MaCH (Y. Li *et al.*, 2010) is based on Hidden Markov model (HMM), which has been developed as a sampling scheme for modeling LD and identifying recombination hotspot by treating a sampled haplotype as an "imperfect mosaic" of a set of reference haplotypes (Daly *et al.*, 2001; N. Li & Stephens, 2003). Each mosaic segment can be viewed as a hidden state in HMM. The goal is to infer the posterior probability, $P(S|G,H)$, of the sequence of hidden states (S) under each observed genotype conditioning on the target individual's genotype vector (G) and the pool of reference haplotypes (H). The posterior probabilities can be calculated through multiple Markov iterations where $P(S|G,H) \propto P(G,S|H)$. The model can be written as

$$P(G,S|H) = P(S_1|H)\prod_{j=2}^{L} P(S_j|S_{j-1},H)\prod_{j=1}^{L} P(G_j|S_j,H)$$

Where $P(S_1|H)$ is the prior probability of the initial mosaic state which is treated to be equal in all configurations. The term $P(S_j|S_{j-1},H)$ is the transition probability which models how the mosaic state changes along the haplotype. The switching of state between marker j and j-1 depends on the historical recombination events which is modeled as a function of the crossover parameter θ_j:

$$P\big(S_j\big|S_{j-1},H\big) = \begin{cases} \theta_j^2/H^2 & if\ x_j \neq x_{j-1}\ and\ y_j \neq y_{j-1} \\ \big(1-\theta_j\big)\theta_j/H + \theta_j^2/H^2 & if\ x_j \neq x_{j-1}\ or\ y_j \neq y_{j-1} \\ (1-\theta)^2 + 2\big(1-\theta_j\big)\theta_j/H + \theta_j^2/H^2 & if\ x_j = x_{j-1}\ and\ y_j = y_{j-1} \end{cases}$$

The term $P(G_j|S_j,H)$ is the emission probability which allows the observed genotype at marker j differs from the genotype of the underlying state which reflects the effect from mutation, gene conversion or genotyping error. It is modeled as a function of the error parameter ϵ_j (As shown below). Both θ_j and ϵ_j are inferred from data and updated in each iteration.

$P(G_j\|S_j,H)$		G		
		0	1	2
S	0	$\big(1-\epsilon_j\big)^2$	$2\epsilon_j(1-\epsilon_j)$	ϵ_j^2
	1	$\epsilon_j(1-\epsilon_j)$	$\big(1-\epsilon_j\big)^2+\epsilon_j^2$	$\epsilon_j(1-\epsilon_j)$
	2	ϵ_j^2	$2\epsilon_j(1-\epsilon_j)$	$\big(1-\epsilon_j\big)^2$

IMPUTE1 (Marchini *et al.*, 2007) is another imputation method also based on HMM, which differs from MaCH in terms of its implementation. For example, the transition probability in IMPUTE1 is modeled as a function of $\rho_j = 4N_e r_j$ where r_j is the recombination rate, pre-

calculated from the HapMap or the 1000 Genome project reference panel and N_e is the pre-set effective population size. In addition, the mutation rate used in calculating emission probability is assumed to be constant which is from population genetics theory. The pre-calibrated parameters can help reduce computation cost if using standard reference panels from the HapMap project and the 1000 Genome Project. However, with the increase of medical sequencing projects, more and more study-specific reference panels are available, where the model parameters may differ from the pre-calibrated ones. In these scenarios, MaCH, using the data-dependent parameters, may be advantageous.

IMPUTE2 is an improved version of IMPUTE1. In IMPUTE2, imputation accuracy is improved through two separate phasing and imputation steps. Because computational burden increases quadratically with the increase of reference size, IMPUTE2, similar to MaCH's implementation, selects a subset of haplotypes so that computation cost increase linearly with a fixed number of selected haplotypes. The subset of reference haplotypes is selected based on their "closeness" to the haplotype of the target individual according to their Hamming distance. This is implemented in both phasing (B. N. Howie *et al.*, 2009) and imputation (B. Howie *et al.*, 2011) step. In addition, a data configuration was implemented to use multiple reference panels simultaneously.

As an updated version of MaCH, MaCH-Admix (E. Y. Liu *et al.*, 2013) has most of the advanced improvement made in IMPUTE2. Importantly, it is tailored towards imputation in admixed population. As mentioned earlier, admixture LD imposes challenges to genotype imputation. Thus it is critical to incorporate the underlying ancestry information for each marker to ensure good imputation quality. MaCH-Admix achieves this by selecting a set of effective reference panel corresponding to the local ancestry composition – the piecewise IBS-matching

strategy. Rather than identifying the effective reference haplotypes through calculating whole-chromosome Hamming distance, the piecewise IBS-matching method divides the chromosome or target region into chunks and then the selection of effective reference haplotypes is conducted within each chunk for each individual, so that it allows the set of effective reference haplotypes differs across genomic regions and across individuals. It is shown that the piecewise IBS method is highly robust and stable and is particularly advantageous for uncommon variants imputation. With the implementation of piecewise IBS method in MaCH-Admix, one can impute study sample from admixed populations with high accuracy by using a cosmopolitan reference panel consisting of haplotypes from diverse continental groups. Another approach implemented in MaCH-Admix for handling admixed population imputation is ancestry-weighted approach. This method uses weighted combination panel, which is produced by duplicating reference haplotypes based on certain weights. Usually the weights are given according to the ancestry composition. For example, in African Americans, 2:8 CEU: YIR weighting scheme is preferred because on average the African American populations are composed of 20% European ancestry and 80% African ancestry. Besides taking the weights based on reported ancestry proportions, MaCH-Admix can estimate the ancestry proportions for target individuals internally, allowing more precise and flexible weighting scheme that can be the same for all or subgroup of individuals or specific for each target individuals.

Besides the computational approaches used in genotype imputation, many other factors can affect imputation accuracy. For example, minor allele frequency – rarer variants are harder to impute, which agrees with the previous observation that rare variants are more difficult to tag as compared with common variants; choice of genotyping chips – chips with denser set of markers, on average, have higher imputation accuracy, which is usually the highest when imputing into

Europeans, followed by Asians and Africans; choice of reference panel – reference panel that better represents the genetic diversity of the study sample results in better imputation quality; reference size – a larger pool of reference haplotypes increases the chance that a good match of haplotype between reference panel and target individual can be found.

Improving genotype imputation accuracy is critical, as better imputation accuracy leads to enhanced statistical power in the downstream genetic association analysis.

2.3 Association studies in admixed populations

Admixed populations are a growing proportion of the US population and suffer from disproportionately higher rates of cardiovascular diseases (Mensah *et al.*, 2005) and certain type of cancer, such as prostate cancer (Bunker *et al.*, 2002). Extending genetic association studies into admixed populations will provide a more complete understanding of the genetic bases of complex traits in human by identifying the shared and distinct genetic components compared with other population groups, which will in turn reduce medical disparity and benefit all people from the development of precision medicine.

2.3.1 Local ancestry inference

Unlike populations with a single ancestry origin, individuals from admixed populations vary in their proportion of parental ancestry. For example, the average proportion of African ancestry is about 80% in African Americans (Stefflova *et al.*, 2011). The exact percentage in individual African American may vary between 0 and 1. This variation is due to the difference in ancestry proportion at each specific locus, which is known as local ancestry. Both global and local ancestry are unobserved, and yet they may be inferred from genotype data. Once local ancestry estimates are obtained, global ancestry can be calculated.

To address this issue, many local ancestry inference methods have been developed. One

of the earliest methods is STRUCTURE (Falush *et al.*, 2003; Pritchard *et al.*, 2000), which uses a Bayesian framework and applies Markov Chain Monte Carlo (MCMC) to sample from the posterior distribution. This early method has complex model and suffers from high computational cost, which is less practical for the analysis involving dense genome-wide genotyping data.

Later methods divide into two main classes – HMM-based and window-based approaches. The HMM-based methods take advantage of dense genotyping data, leveraging information from neighboring markers and explicitly modeling background LD. In these methods, hidden state corresponds to ancestral haplotype segments and the transition between states corresponds to recombination that happened within and between ancestral populations. The earlier method in this class, SABER (Tang *et al.*, 2006), uses an extension of HMM, Markov–hidden Markov model, to model background LD, whereas this LD model cannot fully capture the complex nature of the LD pattern in the genome thus can lead to biased estimates (Price *et al.*, 2009). More accurate LD models were developed and implemented in HAPMIX (Price *et al.*, 2009) and HAPAA (Sundquist *et al.*, 2008), where reference haplotypes are used to account for background LD. The choice of reference haplotypes is critical for inference accuracy, which serves as the proxies for the true haplotypes of the ancestral populations. HapMap CEU and YRI panel have been shown to be good reference haplotypes for inferring local ancestry in African American populations (Price *et al.*, 2009). HAPMIX can only handle ancestry inference in populations with two-way admixture. The limitation of HMM-based methods is its high computational cost due to the large number of parameters involves in modeling LD.

The other class adopts window-based method, which divides the whole genome into overlapping or non-overlapping windows and performs inference within each window. These

methods are shown to be highly accuracy and much faster than HMM-based methods, even though they do not explicitly model LD and uses only unlinked markers. These methods usually find their application in populations with multi-way admixture, typically three-way admixture. The major methods include LAMP (Sankararaman *et al.*, 2008), its extension WINPOP (Pasaniuc *et al.*, 2009) and recently LAMP-LD (Baran *et al.*, 2012). LAMP (Sankararaman *et al.*, 2008) assumes no recombination within each window, and performs a clustering algorithm to estimate the local ancestry for each SNP. In the end, a most likely ancestry of the SNP is taken by a majority vote over all windows that cover the SNP. LAMP is highly accurate when applied to populations with distant ancestral populations, whereas performs poorly in admixed populations whose ancestral populations are closely related. As an extension of LAMP, WINPOP (Pasaniuc *et al.*, 2009), enhances the inference accuracy in closely related ancestral populations by improving the modeling of recombination events by allowing less than one admixture event per window and allowing adaptive window size depending on the local ancestral structure. To leverage haplotype structure for local ancestry inference without sacrificing computation speed, LAMP-LD (Baran *et al.*, 2012) uses combination of window-based method and HMM where HMM-based inference is applied within each window. It reduces the estimation bias in haplotype-based methods resulting limited reference size. Pasaniuc et al. (Pasaniuc *et al.*, 2013) performs the first empirical assessment of local ancestry inference accuracy in Latinos, by measuring the rate of local ancestry assignments that cause Mendelian inconsistencies in local ancestry (MILANC) in trios. They demonstrate the superior performance of WINPOP and LAMP-LD as compared with two other methods.

Although methods within the two main classes are the commonly used ones in application requiring local ancestry information, many other methods have been developed to

handle local ancestry inference, such as PCAdmix (Brisbin *et al.*, 2012), a PC-based method; SeqMix (Hu *et al.*, 2013), which tackles local ancestry inference in exome-sequencing samples and RFMix (Maples *et al.*, 2013), a random forest based method.

2.3.3 Association tests in admixed populations

Subpopulation structure in a study sample may confound phenotype-SNP associations leading to spurious results, primarily at markers with differential allele frequencies among subpopulations (Price *et al.*, 2010). This issue of population stratification is well addressed in recent years. As a common practice, genomic control (Devlin & Roeder, 1999) is used to quantify the extent of inflation cased by population stratification or other confounders (such as cryptic relatedness) and principal components analysis (Price *et al.*, 2006) is used to correct for stratification if necessary.

In admixed populations, the population stratification issue is intrinsic. Although controlling for global ancestry can reduce false positives, it cannot necessarily eliminate the confounding effect from local ancestry (Qin *et al.*, 2010).

As an alternative, taking advantage of the correlation between phenotype and local ancestry, admixture mapping is developed to test for associated loci in admixed populations. The idea is that the genetic factors leading to disease disparity between the parental populations would have differential allele frequencies. By testing for the association between the ancestry of the chromosome segment and phenotype in admixed sample, the region harboring the causal variant may be identified (Winkler *et al.*, 2010). Admixture mapping can be traced to Chakraborty and Weiss (Chakraborty & Weiss, 1988), who theoretically demonstrate that the admixture LD in admixed sample can be used to detect the linkage relationship between two loci. Admixture mapping is particularly effective when the difference of disease risk in the parental

populations is large, as long as it is accounted for by genetic factors rather than entirely by environmental factors (Smith *et al.*, 2004). One key advantage of genome-wide admixture mapping over genome-wide genetic linkage study is that it requires the use of much less markers (Smith *et al.*, 2004), which reduces multiple testing burden. Moreover this is particularly advantageous when genotyping is expensive. Due to recent admixture, the extend of admixture LD could be several megabases. Therefore, a few thousand ancestry informative markers (AIM; markers with large differential allele frequencies between parental populations) would be sufficient. Since the availability of the first dense map of AIM between Europeans and Africans (Smith *et al.*, 2004), admixture mapping has been successfully applied in gene discovery in African Americans for diseases that differ in prevalence in parental populations, such as hypertension (Zhu *et al.*, 2005), prostate cancer (Freedman *et al.*, 2006), type II diabetes (Ng, 2015) and Proliferative Diabetic Retinopathy (Tandon *et al.*, 2015). On the other hand, however, the long stretch of admixture LD limits the resolution of admixture mapping, which is higher than that of family based linkage analysis but is lower than that of ancestry LD based genetic association analysis (Shriner, 2013).

As genotyping cost dropped drastically, more and more dense genotype data from cohorts of admixed populations are available, such as that from the Women's Health Initiative (WHI), the Atherosclerosis Risk in Communities Study (ARIC), the Jackson Heart Study (JHS), the Coronary Artery Risk Development in Young Adults Study (CARDIA), the Multi-Ethnic Study of Atherosclerosis (MESA) and the Hispanic Community Health Study / Study of Latinos (HCHS/SOL). The dense genotype data provide an opportunity to apply GWAS in admixed populations. Given the challenges from population stratification, statistical methods have been developed to control false positives due to local population structure. Qin et al. showed that PCs

calculated in local genomic regions (local PCs) strongly correlated with local ancestry. It is more effective in eliminating spurious findings by incorporating local PCs in genetic association studies (Qin *et al.*, 2010). Rather than controlling for local PCs, Wang et al. directly controls for local ancestry estimate and shows its effectiveness in controlling type I error (X. Wang *et al.*, 2011). Furthermore, genotype and ancestry have been shown to contain independent information, thus genetic association test and admixture mapping may complement each other (Tang *et al.*, 2010). Taking advantage of both tests, Reiner et al. (Reiner *et al.*, 2012) conducted SNP and admixture mapping separately using the same set of markers, where they used admixture mapping results to explain the peaks observed in GWAS scan and *vice versa* (Reiner *et al.*, 2012). Along the same line, Tang et al. developed a test to jointly test ancestry and SNP effect in a family study design (Tang *et al.*, 2010). Later, a joint test (MIXSCORE) under case/control design is proposed (Pasaniuc *et al.*, 2011) as well as a Bayesian method (BMIX) (Shriner *et al.*, 2011). Due to the difference in the strength or direction of shared LD between the tested SNP and causal variant in the parental populations, association effect may have different size or directions. Liu et al. proposed to capture effect heterogeneity among the ancestral populations by including interaction term between local ancestry and genotype (J. Liu *et al.*, 2013).

CHAPTER 3: A COMPREHENSIVE SNP AND INDEL IMPUTABILITY DATABASE[1]

3.1 Introduction

Genotype imputation has proven to be a powerful tool in genome-wide association studies (GWAS) by facilitating fine mapping and the merging of datasets from different genotyping platforms (Y. Li *et al.*, 2009; Marchini & Howie, 2010). It is a way to predict genotypes computationally based on linkage disequilibrium patterns instead of obtaining genotypes by laboratory-based procedure (Browning & Yu, 2009; B. Howie *et al.*, 2011; Y. Li *et al.*, 2010). As it has been shown to directly affect downstream analysis, imputation accuracy needs to be taken into consideration when designing and performing GWAS (Zheng *et al.*, 2011). For instance, at the study design stage, a question of interest would be which commercially available genotyping platform can provide the optimal imputation quality genome-wide or in certain genomic region(s) of interest. Such a question can be answered by assessing the imputation accuracy of relevant variants. However, there has been no resource available to provide variant imputability information without actually performing imputation.

A commonly used evaluation method is to mask a subset of markers, impute their dosages and compare those dosages with the true (masked) genotypes for those markers (Y. Li *et al.*, 2010). This method, however, can only be used after genotypes have already been obtained

[1] This chapter previously appeared as an article in Bioinformatics. The original citation is as follows: Duan, Q., *et al.*, A comprehensive SNP and indel imputability database, *Bioinformatics*, 2013, 29(4):528-531.

and therefore cannot help guide study design decisions. In addition, the evaluation procedure can be computationally costly because of the requirement of conducting imputation, particularly with the emergence of reference panels built through re-sequencing efforts (Sampson *et al.*, 2012) To facilitate genetic studies in the era of genomic re-sequencing, we have built a database containing imputation accuracy information for SNPs and indels identified from the 1000 Genomes Project (The 1000 Genomes Project *et al.*, 2010), a sequencing-based reference resource, which has demonstrated its potential for enhancing the power of genetic association studies in the post-GWAS era (Day-Williams *et al.*, 2011; Holm *et al.*, 2011; Huang *et al.*, 2012). The assessment of marker imputability was carried out through a leave-one-out imputation procedure: a single individual serves as the imputation target, and imputation is performed using haplotypes from all the other individuals as reference. Imputation accuracy was quantified within each of the four major continental groups surveyed by the 1000 Genomes Project. We anticipate this database containing imputation accuracy information searchable by continental group and by GWAS genotyping platform will be a useful resource for geneticists in this sequencing era.

3.2 Data setup and retrieval

3.2.1 Database

The database contains imputation quality information (as measured by dosage r^2, the squared Pearson correlation coefficient between the imputed dosage—ranging continuously from 0 to 2—and the observed/masked genotypes—taking values 0, 1 or 2 copies of a given allele) for every non-singleton SNP and indel discovered by and passing default quality filters in the 1000 Genomes Project (The 1000 Genomes Project *et al.*, 2010). The dosage r^2 of each variant reflects its potential imputation accuracy when conducting imputation using haplotypes from the 1000 Genomes Project as reference. Imputability information is available for multiple genotyping

platforms, and separately for each of the four major continental groups [Europeans (EUR), Africans (AFR), Asians (ASN) and Americans (AMR)]. Details regarding sub-population constituents of the continental groups can be found at ftp://ftp-trace. ncbi.nih.gov/1000genomes/ftp/release/20110521/.

3.2.2 Methods

The dosage r^2 of each variant was obtained using a leave-one-out imputation procedure with MaCH-Admix (E. Y. Liu *et al.*, 2013) [high Pearson correlation (0.85–0.94) with those obtained using minimac (B. Howie *et al.*, 2012) and IMPUTE2 (B. Howie *et al.*, 2011), and lower correlation (0.71–0.85) with those from BEAGLE (Browning & Yu, 2009), data not shown] on samples from the latest release of the 1000 Genomes Project (version 3 March 2012 release, 2184 haplotypes). We mimicked typical GWAS imputation practice by masking genotypes at markers absent from the selected genotyping platform and treating them as untyped. These untyped markers were imputed in one individual at a time using the haplotypes of all the remaining individuals as reference (2182 haplotypes). The imputation accuracy of each marker, measured by dosage r^2, was calculated separately in each of four continental groups currently available in the 1000 Genomes Project. The genotyping platforms we have evaluated include Affymetrix 5.0, Affymetrix 6.0, Affymetrix Axiom, Illumina Human1M, Illumina Omni 5 M and Illumina Omni ZhongHua. The results of the assessment are searchable through a publicly available database.

3.2.3 Usage

Our database can take as input either a list of marker names or the start and end position of a genomic region on a specified chromosome. Users can choose to view information corresponding to one or more specific genotyping platforms. Given the marker or region input

and the choice of genotyping platform, our database returns imputability information for variants of interest ordered by their genomic location according to NCBI Build 37. Users have the option to display or to download the imputability information for each continental group or the maximum dosage r^2 across the four continental groups (max-r^2). Moreover, users can filter results by max-r^2. Markers with no rsID follow chromosome:physical-coordinate nomenclature (**Figure 3.1A**). In addition, for an SNP–indel pair with the same genomic location, the SNP is always listed before the indel (**Figure 3.1B**).

3.2.4 Examples

The first example shows the utility of our database at the study design stage. Specifically, suppose an investigator wants to decide between two genotyping platforms, Affymetrix 6.0 and Affymetrix Axiom, based on imputation accuracy within a 1-kb region on chromosome 9p21 (22,095,555 to 22,096,555 bp) harboring the SNP rs10757274 known to be associated with risk of coronary heart disease and multiple related phenotypes (Cunnington *et al.*, 2010; McPherson *et al.*, 2007). Our database interface, the example query, as well as the results of the query are shown in **Figure 3.2**. Given the regional input (start and end position 22,095,555 and 22,096,555 on chromosome 9), our data- base returns a list of markers within the region (only the top three are shown). For each marker, the database shows its marker name, genomic location and dosage r^2 for the two selected genotyping platforms across four continental groups. To ease comparison, users can choose to display max-r^2 instead of r^2 values for each population separately and/or filter by setting non-zero max-r^2 threshold. Based on what is shown in **Figure 3.2**, we would recommend the Axiom over the 6.0 panel, unless the samples under study are Americans (e.g. Hispanic or African Americans) and the SNP of primary interest is rs139492236. Note that this is a toy example mainly meant to introduce the interface of our database where we show only the

top three SNPs. For more realistic settings where the region of interest typically includes many more markers, we recommend prioritization of markers in the region (e.g. according to functional annotation and/or evidence from existing association or functional studies, if available), followed by the examination and comparison of the max-r^2 distribution through 'Download Results' or 'Genome-wide Graphical Comparison'. Such comparison of imputation accuracy across platforms will facilitate decision making regarding the choice of genotyping assays.

Once the investigator has decided on the genotyping platform, a typical question is whether specific markers or markers in specific regions of interest can be imputed well (e.g. novel variants or associated regions identified in other cohorts). When computational resources are limited or when an investigator is interested in a considerable number of markers/regions, imputability information can help prioritize markers/regions that have the potential to be well-imputed as well as avoid wasting resources on markers/regions that have little potential for high-quality imputation. As shown in **Figure 3.2**, our database contains four dosage r^2 values (one for each continental group) for each marker, given a genotyping platform. As false-negatives (markers that can be well-imputed but with bad predicted imputation accuracy such that one would not perform actual imputation) are typically more costly than false-positives (the consequence would be wasted computational resources on markers/regions that are truly not imputable), we recommend using the maximum dosage r^2 across the four continental groups (max-r^2) to guide decisions, particularly for samples involving admixed individuals. **Figure 3.3** shows the receiver-operating characteristic curve for data from the Cebu Longitudinal Health and Nutrition Survey (CLHNS) when max-r^2 is used for thresholding. In this cohort of Filipinos (Adair *et al.*, 2011; Marvelle *et al.*, 2007), we have 81 individuals who have both Affymetrix 5.0

(Lange *et al.*, 2010) and Metabochip (Croteau-Chonka *et al.*, 2012) genotypes. We imputed the Metabochip SNPs from the Affymetrix 5.0 data, using haplotypes from the 1000 Genomes Project as reference. We computed the imputation accuracy in this sample (CLHNS-specific dosage r^2) by comparing the imputed dosages with the genotypes obtained through genotyping using Metabochip. The *y*-axis shows the proportion of poorly imputed SNPs (CLHNS-specific dosage $r^2 < 0.2$) removed and the *x*-axis shows the proportion of well-imputed SNPs (CLHNS-specific dosage $r^2 > 0.8$) sacrificed for SNPs in different minor allele frequency (MAF) categories (defined within CLHNS). Using a max-r^2 threshold of 0.7, which removes ~15 million of the ~31 million markers in the latest release from the 1000 Genomes Project, we found that the database filters out 77%, 58%, 51% and 42% of the poorly imputed SNPs (again, SNPs with CLHNS-specific dosage $r^2 < 0.2$) at the cost of 0.3%, 0.8%, 1.5% and 4.6% well-imputed markers (SNPs with CLHNS-specific dosage $r^2 > 0.8$) in the MAF categories of >5%, 3–5%, 1–3% and 0.5–1%, respectively. Using a different threshold of 0.5 (0.9), which removes ~12 (~20) million of the ~31 million markers, we can filter out 54%, 32%, 29% and 26% (92%, 80%, 75% and 66%) of the poorly imputed SNPs at the cost of 0.1%, 0.3%, 0.2% and 2.1% (4.8%, 6.1%, 7.5% and 17.2%) well-imputed SNPs. We also confirmed in samples of Caucasians and samples of African Americans (data not shown) that a max-r^2 in the range of 0.5–0.8 serves as a reasonable thresh- old in terms of a trade-off between sensitivity and specificity. The actual threshold an investigator selects can be tailored according to MAF and available computational resources (including both CPU times and disk space). We and others have previously observed lower imputation quality for rarer variants (International HapMap *et al.*, 2010; L. Li *et al.*, 2011; Liu *et al.*, 2012). Our database now shows that imputation quality of rarer variants is also more challenging for prediction estimation: the total area under the receiver-operating characteristic

curve is 0.97, 0.91, 0.88 and 0.79, respectively, for markers with MAF>5%, 3–5%, 1–3% and 0.5–1%.

3.3 Conclusion

In summary, we have built a publicly available database for marker imputability to aid genetic association studies in the re-sequencing era (Fridley *et al.*, 2010; Y. Li *et al.*, 2011; Sampson *et al.*, 2012). Reference panels built from re-sequencing studies bring us the benefits of improved imputation accuracy and the potential to impute low-frequency variants. These benefits come, however, at the cost of heavy computational burden for imputation if we impute every marker discovered by sequencing, which is 430 million in the latest release from the 1000 Genomes Project. It is therefore desirable to have direct access to marker imputability information without actually conducting genotype imputation. Our marker imputability database provides direct access to imputation accuracy information for SNPs and indels identified from the 1000 Genomes Project across four major continental groups using multiple genotyping platforms. We anticipate that this database will serve as a useful resource for researchers in this re-sequencing era in terms of design and analysis of genetic association studies. In addition, although the database is developed mainly for guidance before actual imputation, it can be used for post-imputation quality assurance by comparing estimated r^2 values in the imputed study sample with those in our database in an SNP-specific manner. Using a cohort of Filipinos, we estimate that we can, with up to 48.6% reduced computation efforts (by imputing only the top 51.4% markers according to imputation quality estimated from individuals in the 1000 Genomes Project), filter out 42–77% of poorly imputed markers at the cost of 0.3–4.6% well-imputed markers. Finally, two caveats should be kept in mind by database users. First, we record results from the MaCH-Admix software. Although more than moderate level of correlation is observed

with results from other imputation software, caution needs to be taken when generalizing to other imputation methods, particularly those that are not based on the Li and Stephens model (N. Li & Stephens, 2003). Second, loss of some typed markers due to quality control in real studies could lead to reduced imputation quality of specific markers, which cannot be modeled generically and are thus not reflected by our database. We will update the database when new data releases of the 1000 Genomes Project or new genotyping platforms become available.

Figure 3.1. An example SNP/indel imputability record from the database. A. SNPs and indels with no rsID are named by chromosome number followed by genomic location in base-pairs (e.g., chr20:4910201). B. When a SNP and an indel have the same genomic location, the SNP is listed first and the indel second (e.g., at position 4895999, SNP rs77916149 is listed first followed by indel chr20:4895999).

A

rs143826645	20	4909947	0.878049	0.680036	N/A	N/A
rs111452017	20	4909963	0.000966	N/A	N/A	0.054442
rs114496810	20	4910064	0.844389	0.970425	N/A	0.788591
chr20:4910201	20	4910201	0.990739	0.994424	0.995163	0.997714
rs142616219	20	4910372	N/A	N/A	0.001081	N/A

B

rs1715362	20	4895803	0.99943	1	0.99883	1
rs77916149	20	4895999	0.954525	0.980537	N/A	N/A
chr20:4895999	20	4895999	0.369739	0.022765	0.085031	0.00304
rs149699301	20	4896003	0.867443	0.913105	N/A	N/A

Figure 3.2. The SNP and indel imputability database interface.

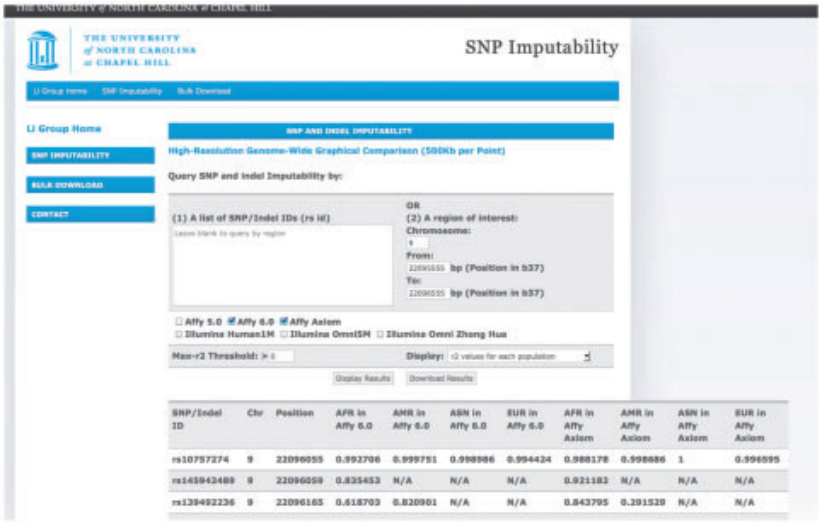
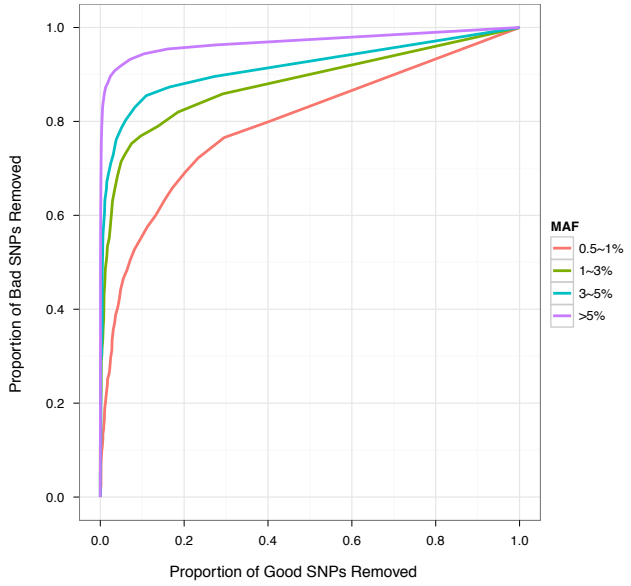


Figure 3.3. Receiver operating characteristic (ROC) curve in the Cebu Longitudinal Health and Nutrition Survey.



CHAPTER 4: IMPUTATION OF CODING VARIANTS IN AFRICAN AMERICANS[2]

4.1 Introduction

Increasingly large reference panels available in the public domain [e.g. those from the 1000 Genomes Project (The 1000 Genomes Project *et al.*, 2010; The 1000 Genomes Project *et al.*, 2012) and UK10K project (Futema *et al.*, 2012) together with improved statistical methods (B. Howie *et al.*, 2012; E. Y. Liu *et al.*, 2013) have enhanced imputation quality, especially for rare variants with minor allele frequency (MAF) < 5%. These improvements have resulted in both discovery and refined mapping of association with complex traits (Auer *et al.*, 2012; Holm *et al.*, 2011; Huang *et al.*, 2012). However, few studies have examined the use of large study-specific reference panels, particularly the use of exome sequencing-derived panels in admixed populations. Here, we present a new resource for imputation in African Americans, built from 1692 African Americans sequenced by the Exome Sequencing Project (ESP) (Tennessen *et al.*, 2012). We assessed the use of the ESP data as an imputation reference panel and compared the results with those obtained using the 1000 Genomes Project Phase1 data (1000G; version 3, March 2012 release) (The 1000 Genomes Project *et al.*, 2012). Additionally, we evaluated the potential consequences of using a reference panel built from samples selected on the basis of phenotypic extremes or disease status instead of a population-based random sample. Lastly, we

---

[2] This chapter previously appeared as an article in *Bioinformatics*. The original citation is as follows: Duan, Q., *et al*., Imputation of coding variants in African Americans: better performance using data from the exome sequencing project, *Bioinformatics*, 2013, 29(21):2744-2749.

compared multiple approaches to combine the ESP and 1000G panels for the imputation of rare coding variants.

4.2 Methods

4.2.1 Exome Sequencing Project

The complete ESP dataset (Fu *et al.*, 2013) consists of whole exome data for 6823 individuals. Samples were sequenced at the University of Washington (SeattleGO) and the Broad Institute (BroadGO). Among the 6823 individuals, 1692 participants were African Americans with genome-wide association data available for analysis. The 1692 African Americans ESP samples include 845 from the Women's Health Initiative (WHI) study (The Women's Health Initiative Study, 1998) as part of the WHI Sequencing Project (WHISP), and a total of 847 including Atherosclerosis Risk in Communities (ARIC) (Muntaner *et al.*, 1998) (N=282), Jackson Heart Study (JHS) (Taylor *et al.*, 2005) (N=366), Multi-Ethnic Study of Atherosclerosis (MESA) (Bild *et al.*, 2002) (N=146) and Coronary Artery Risk Development in Young Adults (CARDIA) (Friedman *et al.*, 1988) (N=53) as part of HeartGO. Most WHISP and HeartGO participants were selected on the basis of primary phenotypes for ESP, which included extremes of body mass index, blood pressure, low-density lipoprotein (LDL), cholesterol, early onset myocardial infarction (MI) cases and controls, ischemic stroke with either early onset or positive family history. Approximately 15% of samples were selected because of having non-missing data for a selected set of core phenotypes, but were not ascertained based on trait values.

4.2.2 Exome Sequencing

Exome sequencing was performed at the University of Washington (SeattleGO) and the Broad Institute (BroadGO). Initial quality control (QC) on all samples involved sample

quantification (PicoGreen), confirmation of high-molecular weight DNA, fingerprint genotyping and sex determination. Samples were failed if total mass, concentration, integrity of DNA or quality of preliminary genotyping data was too low or sex typing was discordant. Following QC, 2mg of extracted genomic DNA was subjected to shotgun library preparation and exome capture as previously described (Tennessen *et al.*, 2012).

4.2.2.1 Genotype Calling

For read mapping and variant analysis, samples were aligned to a human reference (hg19) using Burrows–Wheeler Aligner (H. Li & Durbin, 2009). Variant detection and genotyping were performed on both exomes and flanking 50bp of intronic sequence.

Typical mean coverage of the target was 60–80x. Variant data for each sample were formatted (variant call format) as 'raw' calls for all samples. Filters considered the total read depth, the number of individuals with coverage at the site, the fraction of variant reads in each heterozygote, the ratio of forward and reverse strand reads carrying reference and variant alleles and the average position of variant alleles along a read. Variant calling was performed across all 6515 samples at the University of Michigan (UMich). Only single nucleotide polymorphisms (SNPs) that passed the UMich support vector machine quality filter were retained for analysis. Details were previously described (Fu *et al.*, 2013).

4.2.2.2 Reference Panel Construction

A reference panel of 2163 individuals (including the 1692 African Americans used in this study and 471 European Americans) was constructed. All of the 2163 individuals have both Genome-wide association study (GWAS; Affymetrix 6.0) genotypes and whole exome sequencing data. When combining the two sources of data, a total of 375 024 bi-allelic autosomal SNPs with minor allele count ≥ 4 (in the 2163 reference panel subjects) did not

overlap with the 702 205 GWAS SNPs. There were 10130 SNPs that overlapped between ESP and the 702 205 GWAS markers. SNPs with concordance 595% were removed (65 SNPs). For overlapping SNPs that passed this concordance filter, GWAS genotype was retained for consistency with the target individuals. A total of 1 077 164 autosomal SNPs were included in the reference panel. These 1 077 164 markers were phased across all 2163 samples using BEAGLE v3.3.1 (Browning & Yu, 2009).

4.2.2.3 ESP 'Extreme' and 'Normal' Panel Construction

The 1692 ESP African Americans were selected based on the following phenotypic traits: (i) LDL (N = 254: 131 with high LDL and 123 with low LDL), (ii) blood pressure (N = 247: 132 with high blood pressure and 115 with low blood pressure), (iii) body mass index (BMI, N = 609: 429 with high BMI and 180 with normal to low BMI), (iv) early onset MI (EOMI, N = 324: 39 EOMI cases and 285 EOMI controls), (v) stroke (N = 40, all cases) and (vi) random samples (N = 218). We constructed one ESP 'Extreme' panel and one ESP 'Normal' panel each with 853 individuals. The ESP 'Extreme' panel included (i) 254 individuals with high/low LDL (131 with high LDL and 123 with low LDL), (ii) 247 individuals with high/low blood pressure (132 with high blood pressure and 115 with low blood pressure), (iii) 40 stroke cases, (iv) 39 EOMI cases and (v) 273 individuals with high BMI. The ESP 'Normal' panel consists of 80% individuals with 'non-extreme' phenotypes and 20% with extreme phenotypes so as to better represent a population sample. Individuals with 'non-extreme' phenotypes (N = 683) are from random sample, EOMI controls and low BMI group. Individuals with extreme phenotypes (N = 170) are from high (N = 85) and low LDL (N = 85) group.

4.2.2.4 The 1000 Genomes Project (1000G)

The 1000 Genomes Phase1 data were downloaded from http://www.sph.umich.edu/csg/yli/ mach/download/1000G.2012-03-14.html. Details regarding the generation of the data can be found in the Phase 1 article (The 1000 Genomes Project *et al.*, 2012).

4.2.3 Target African Americans

4.2.3.1 GWAS Data

All of the 1661 target African Americans in this study were genotyped using the Affymetrix 6.0 genotyping platform as part of the WHI SNP Health Association Resource study. Before phasing and imputation, we removed Affymetrix 6.0 SNPs with genotype call rates <90%, or Hardy–Weinberg exact test (Wigginton *et al.*, 2005) $P<10^6$ or MAF < 1%. QC details were described previously (Auer *et al.*, 2012; Reiner *et al.*, 2011).

4.2.3.2 Metabochip data

All of the 1661 target African Americans in this study were also genotyped using the Metabochip (Voight *et al.*, 2012) in an attempt to generalize genetic effects across racial groups by the WHI Population Architecture using Genomics and Epidemiology (PAGE) study. Standard QC was performed, including removal of markers with genotype call rate < 95% or Hardy–Weinberg $P<10^6$, as well as exclusion of individuals who showed excess heterozygosity, were part of an apparent first-degree relative pair, or were ancestry outliers as determined by Eigensoft (Price *et al.*, 2006). Details can be found in the PAGE Metabochip article (Buyske *et al.*, 2012).

Genotypes at the Metabochip SNPs were not used for imputation but rather used for assessment of imputation quality. In total 5035 markers, which were on Metabochip, in 1000G and in ESP, but not on Affymetrix 6.0, were used for imputation quality assessment.

4.2.3.3 Overlap with ESP African Americans

African Americans present in ESP were not included as target. In other words, individuals in the reference ESP and the target were mutually exclusive. In addition, we removed any target with PLINK (Purcell *et al.*, 2007) estimated identity-by-descent (IBD) ≥ 0.2 with any reference individual such that our final target set did not contain any apparent first-degree relative with the reference ESP.

4.2.3.4 Imputation using IMPUTE2

In the main text, unless otherwise specified, we present results using minimac for imputation. **Figure 4.7** and **Table 4.8** showed that our recommendation of ESP alone or concatenation of ESP with 1000G (ESP_U_1000G) over 1000G still held when IMPUTE2 was used for imputation. We note that in the main text, our recommendation against IMPUTE2's two panel mode (option 3: ESPþ1000G) was confounded by software/method choice: ESP alone or ESP_U_1000G using minimac performed better than IMPUTE2's ESP + 1000G, but when using IMPUTE2 for all, ESP alone or ESP_U_1000G performed similarly as ESP + 1000G.

4.3 Results

4.3.1 Comparison of imputation quality

We first performed imputation, using either ESP or 1000G as reference, into 1661 African Americans in the WHI study (the 'target' sample) who were genotyped by both the Affymetrix 6.0 (Auer *et al.*, 2012) and the Illumina Metabochip array (Buyske *et al.*, 2012; Liu *et al.*, 2012). We used MaCH (Y. Li *et al.*, 2010), a hidden Markov model that leverages linkage disequilibrium information among samples of unrelated individuals, to pre-phase the 1661 WHI African Americans at the Affymetrix 6.0 markers. The ESP reference panel was built from 1692 African Americans with genotypes from both the Affymetrix 6.0 platform and whole exome sequencing. These genotypes were merged and phased using BEAGLE (Browning & Yu, 2009).

Imputation into the 1661 target WHI African Americans was performed with minimac (B. Howie *et al.*, 2012) (similar results were obtained with IMPUTE2; see Methods) using their Affymetrix 6.0 genotypes only; genotypes from the Metabochip genotyping were saved for evaluation. Following the literature (Browning & Yu, 2009; Y. Li *et al.*, 2010), we used dosage r^2 [squared Pearson correlation between imputed dosages (ranging continuously from 0 to 2) and experimental genotypes (coded as 0, 1 or 2)], which directly determines effective sample size for subsequent association analysis (Pritchard & Przeworski, 2001), to gauge imputation quality. We also use Rsq, the estimated dosage r^2 generated by minimac, as the post-imputation QC metric. We observed 8.3–11.4% increases in average dosage r^2 for variants with MAF<1% using the ESP reference panel compared with the 1000G reference panel (paired Wilcoxon $P<1.3\times10^{-4}$ - 4.1×10^{-16}). Such increases were observed without applying any post-imputation QC, that is, when every imputed variant was retained. Similarly increased dosage r^2 was observed across a broad range of post-imputation QC stringency (removing 0–90% of variants; **Figure 4.1** and **Table 4.1**). As imputation is routinely performed in 10 000–100 000 individuals (Auer *et al.*, 2012; Cho *et al.*, 2012; Dastani *et al.*, 2012; Holm *et al.*, 2011; Teslovich *et al.*, 2010), such an increase would correspond to increasing the sample size for association testing by 1000–10 000 samples.

Because the ESP panel is larger and consists entirely of African Americans, we conducted more comparisons by assessing the performance of 10 random subsets from ESP of the same size as 1000G (both for the full 1000G panel [Number of haplotypes (H) = 1092 × 2; reference panels termed ESP.1092 and 1000G.1092] and the most relevant panel [AFR + EUR, H = 625 × 2; reference panels termed ESP.625 and 1000G.625]). The difference in effective sample size derived from the ESP and 1000G reference panels, although smaller, remains

(**Figure 4.2** and **Table 4.2**). For example, when comparing ESP.1092 with 1000G.1092 and retaining all imputed variants in the analysis (no post-imputation QC), we observed an average dosage r^2 increase of 11.3, 4.6 and 6.1% for variants with MAF < 0.2%, 0.2–0.5% and 0.5–1%, respectively. The corresponding dosage r^2 increases for a comparison of ESP.625 with 1000G.625 were 13.9%, 1.0 and 3.1%, respectively. The superior performance of ESP over 1000G was likely driven by two primary factors. First, genotypes for rare variants from ESP were derived from high coverage sequencing, whereas those from 1000G were in part from low coverage sequencing (1000G data we used here are the integrated panel constructed from low coverage whole genome sequencing, deep exome sequencing and SNP array genotyping). Second, ESP African Americans (~50% also from WHI, detailed in Materials and Methods) were better matched to the 'target' WHI African Americans for ancestry than were the samples in the 1000G panel, which were pooled from several populations of European, African and African American ancestry.

As expected, better quality imputation using the ESP panel produces a larger number of well-imputed rare coding variants than using the 1000G panel (Rsq>0.6 for MAF<0.5%; detailed in **Table 4.3**). For example, the number of well-imputed variants was 2.28, 2.83, and 1.54 times greater than that from 1000G for MAF<0.2, 0.2–0.5 and 0.5–1%, respectively (**Table 4.3**). The boost in imputation quality as well as in the number of well-imputed markers is expected to enhance power for testing association with phenotypic traits. For example, out of the eight novel blood trait associated variants reported in Auer *et al*. (Auer *et al.*, 2012), two are not in 1000G but ESP only (**Table 4.4**).

4.3.2 Impact of imputation reference panel

Many subjects sequenced in ESP were selected on the basis of phenotypic extremes or disease status (detailed in Materials and Methods), an approach that has been shown to increase power for association testing of the specific phenotype (Barnett *et al.*, 2013; Guey *et al.*, 2011; Kryukov *et al.*, 2009). To our knowledge, the consequences of such a design for developing an imputation reference panel have not been previously evaluated. To this end, we constructed two ESP-derived reference panels: 'ESP.extreme' and 'ESP.normal' each of size H=853×2. The former included 254 African Americans from LDL cholesterol extremes, 247 from blood pressure extremes, 40 stroke cases, 39 early onset MI (EOMI) cases and 273 with extremely high BMI. The latter included 85 samples with high LDL, 85 with low LDL and 683 from the 'middle' of the phenotype distributions. We observed no loss of imputation quality using the 'Extreme' panel. (**Figure 4.3** and **Table 4.5**).

4.3.3 Alternative options to use or combine reference panels

Although our results suggested that the ESP panel led to substantially improved imputation accuracy of rare coding variants compared with the 1000G panel, the combination of the two panels could potentially result in even better performance than either one individually. We considered the following four options. The default option, Option 0, was to select a single panel a priori based on reference panel size, marker density and ancestry match. In this case, Option 0 would be the ESP reference panel alone, as it contains more haplotypes (3384 over 2184 in 1000G), greater marker density in exons and a better ancestry match with the target African Americans. Option 1 was to first impute using each panel separately, and then for each marker to select the one with higher Rsq. Option 2 was to impute using a concatenated panel of

the two (ESP_U_1000G). Option 3 was to impute using IMPUTE2, which allows two separate reference panels (ESP + 1000G).

The best option among the four was the concatenation of the two panels (Option 2) with ESP alone (Option 0), a close second best. For example, the average dosage r^2 increased by 1.8%, 2.3% and 1.5%, respectively, for markers with MAF<0.2, 0.2–0.5 and 0.5–1% using Option 2 over Option 0 (**Figure 4.3** and **Table 4.6**). We observed no noticeable gains using Option 1 compared with Option 0 with differences in dosage r^2 in the range of 0.02–1.5% (**Figure 4.5** and **Table 4.7**). Therefore, we would not recommend using Option 1, the Rsq-based selection, because higher Rsq does not guarantee better imputation quality. In fact a low quality reference panel could lead to poorly estimated Rsq values. Finally, IMPUTE2's ability to combine two reference panels (Option 3), led to decreased imputation quality compared with Option 0. For example, dosage r^2 decreased by an average of 7.3, 4.3 and 3.9% for markers with MAF<0.2, 0.2–0.5 and 0.5–1% (**Figure 4.6** and **Table 4.8**). Although less accurate, the convenience provided by IMPUTE2's approach warrants closer consideration. Decreases in quality could be due to software implementation because we used minimac for options 0–2 and IMPUTE2 for option 3. But importantly, our recommendation of concatenation of the two or ESP alone over 1000G alone or post-imputation Rsq-based selection holds when IMPUTE2 was used for all four options (see 'Imputation using IMPUTE2' in Materials and Methods, **Figure 4.7** and **Table 4.9**).

4.4 Discussion

We note that ESP is heavily enriched for extremes from several phenotypes rather than a single phenotype. Thus, it is unclear whether these results generalize to a design where sequenced subjects are selected based on extremes for a single phenotype. We did not attempt to

select one phenotype for evaluation, as doing so would reduce our reference size to below 300, which we view as of little value for the imputation of rare variants. We expect such 'Extreme' panels to make little difference for imputation overall and may affect imputation in the specific trait associated regions when the causal variant(s) exert large effect(s).

Although we recommend the concatenation of ESP and 1000G, we observed only modest gains in imputation quality by combining the two. Previous studies suggest that these gains may depend in part on the ethnic make-up of the study subjects (Browning & Yu, 2009) and whether 1000G data add substantial haplotype diversity. These gains should be weighed against the logistical challenges of combining data from multiple sources to avoid batch effects (e.g. mismatched strands, inconsistent marker naming schemes or systematic differences in genotype calling, QC or phasing).

In summary, we found that the ESP African American reference panel outperformed the 1000G reference panel for the imputation of rare coding variants in African Americans, both in terms of imputation quality, the number of imputable markers and consequently power for trait association testing. The finding was robust to adjustment of reference size and matching on ethnicity. We did not observe loss of imputation quality because of the ESP design for enriched sequencing of subjects selected for phenotypic extremes. Regarding the optimal way to combine the two panels, our evaluations suggested that ESP alone or concatenation of the ESP and 1000G reference panels was superior to either post-imputation selection based on Rsq or IMPUTE2's implementation of two separate reference panels. We focused here on imputation of coding variants from ESP. However, we believe that the conclusions drawn here apply to rare variants across the genome as recently reported by several whole-genome sequencing-based studies (Fuchsberger *et al.*, 2012; Sanna, 2012) in individuals of European ancestry. These studies and

our present work strongly suggest that population matched samples, even in diverse populations such as African Americans, can clearly outperform 1000G imputation performance. Therefore, we recommend investigators routinely consider sequencing for the design (Kang *et al.*, 2013) and analysis of the study samples.

Figure 4.1. Comparison of dosage r^2 between ESP-based and 1000G-based imputation. The *x*-axis is the proportion of SNPs that were removed based on elevated Rsq threshold (QC). The *y*-axis is the mean dosage r^2 (squared Pearson correlation between imputed dosages and experimental genotypes).

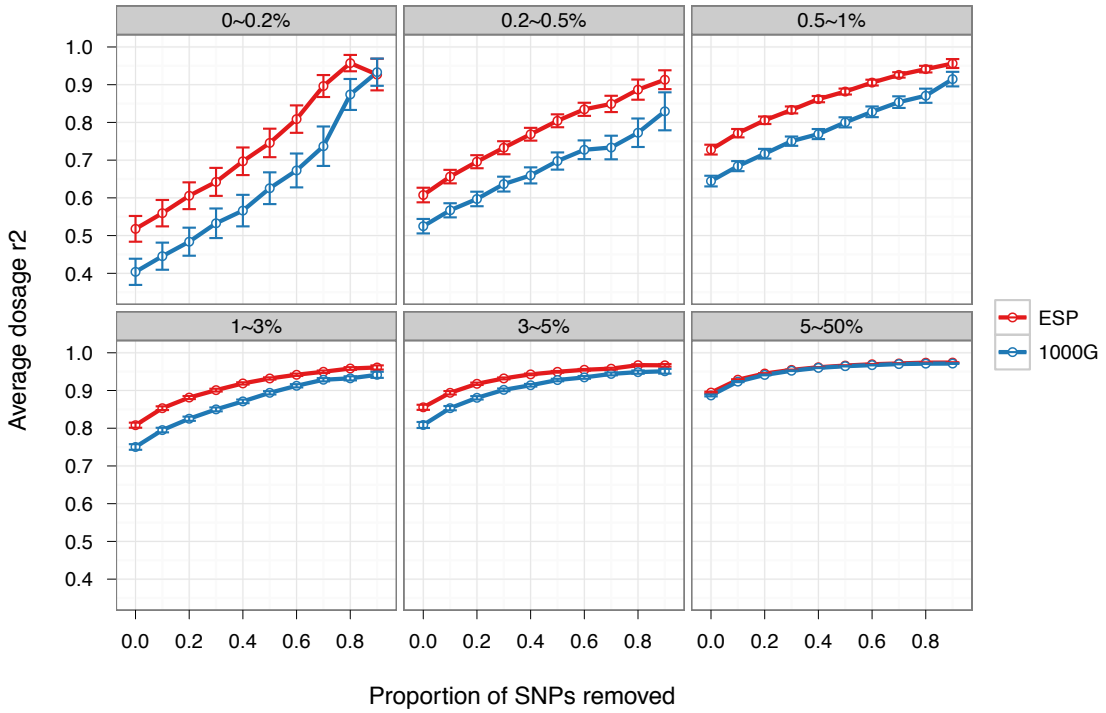


Figure 4.2. Comparison of dosage r^2 between ESP and 1000G full/relevant panel imputation. The *x*-axis is the proportion of SNPs that were removed based on elevated Rsq threshold (QC). The *y*-axis is the mean dosage r^2 (squared Pearson correlation between imputed dosages and experimental genotypes).

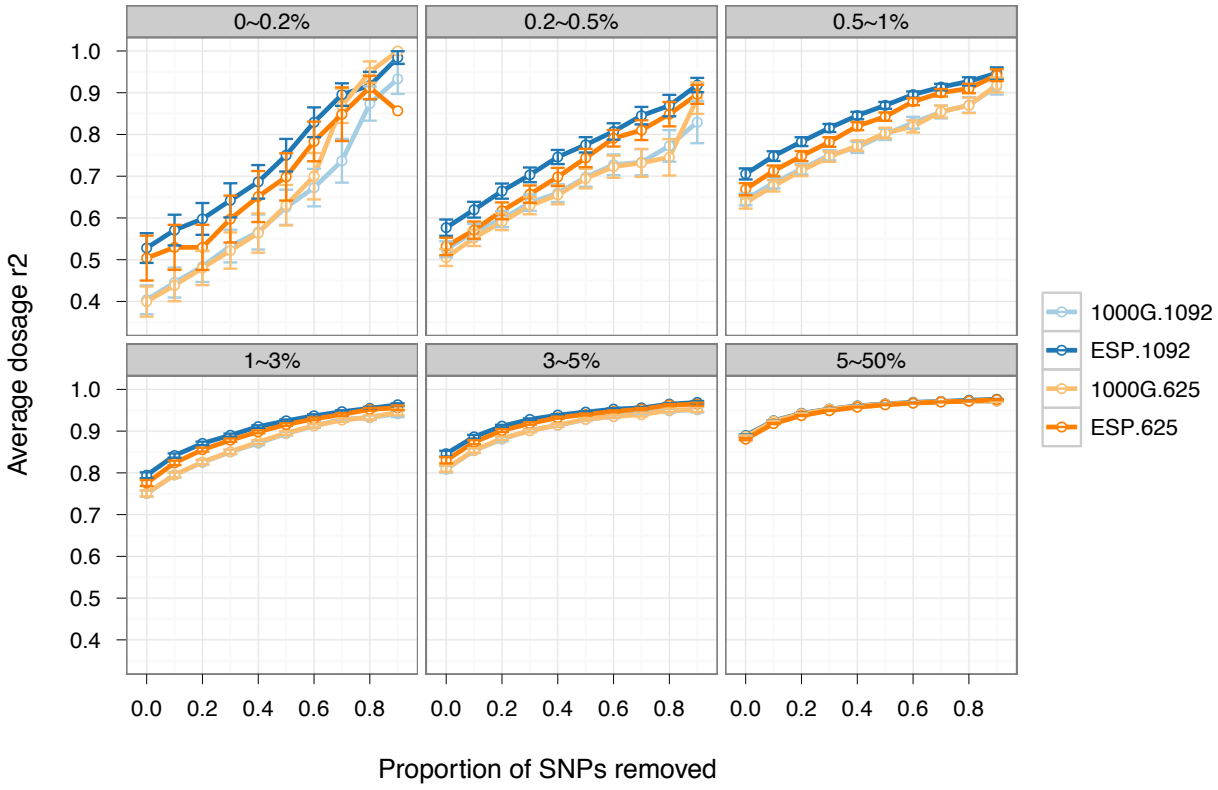


Figure 4.3. Comparison of dosage r^2 between ESP.extreme and ESP.normal imputation. The *x*-axis is the proportion of SNPs that were removed based on elevated Rsq threshold (QC). The *y*-axis is the mean dosage r^2 (squared Pearson correlation between imputed dosages and experimental genotypes)

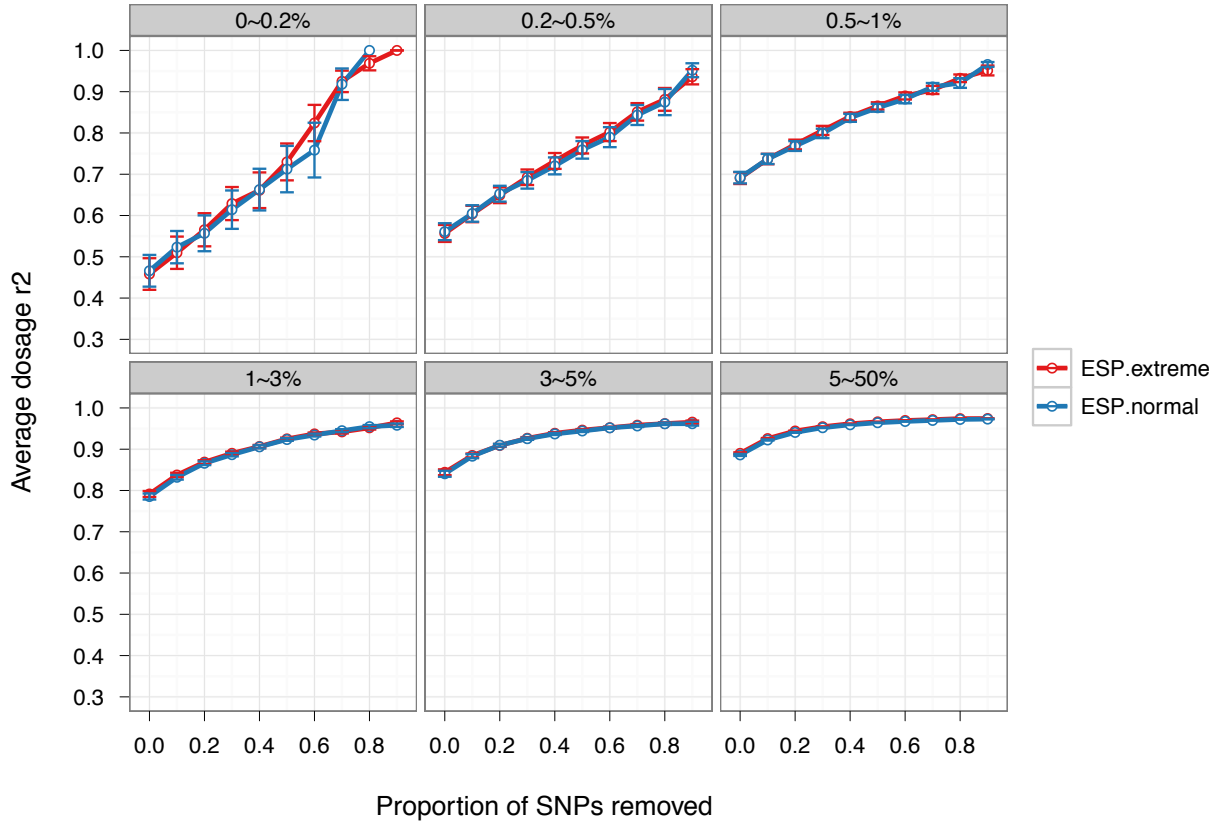


Figure 4.4. Comparison of dosage r^2 between using option 2 and option 0.

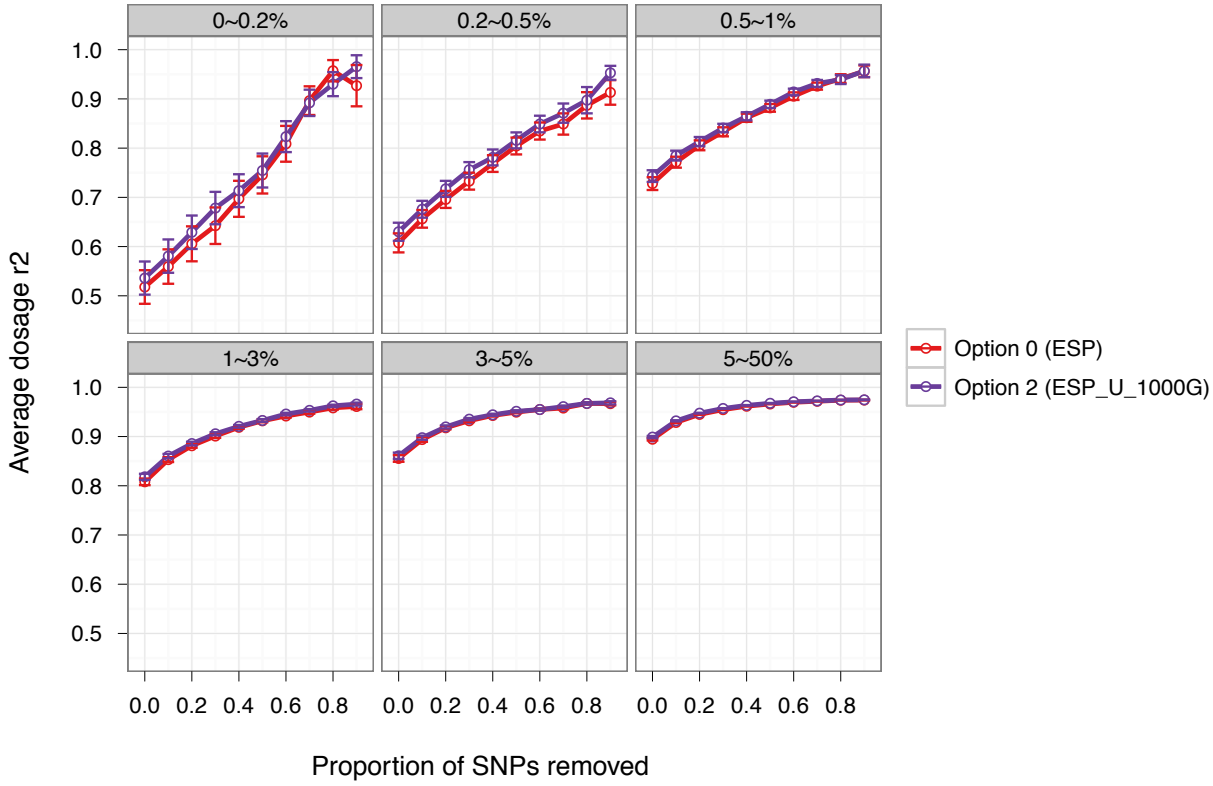


Figure 4.5. Comparison of dosage r^2 between using option 1 and option 0.

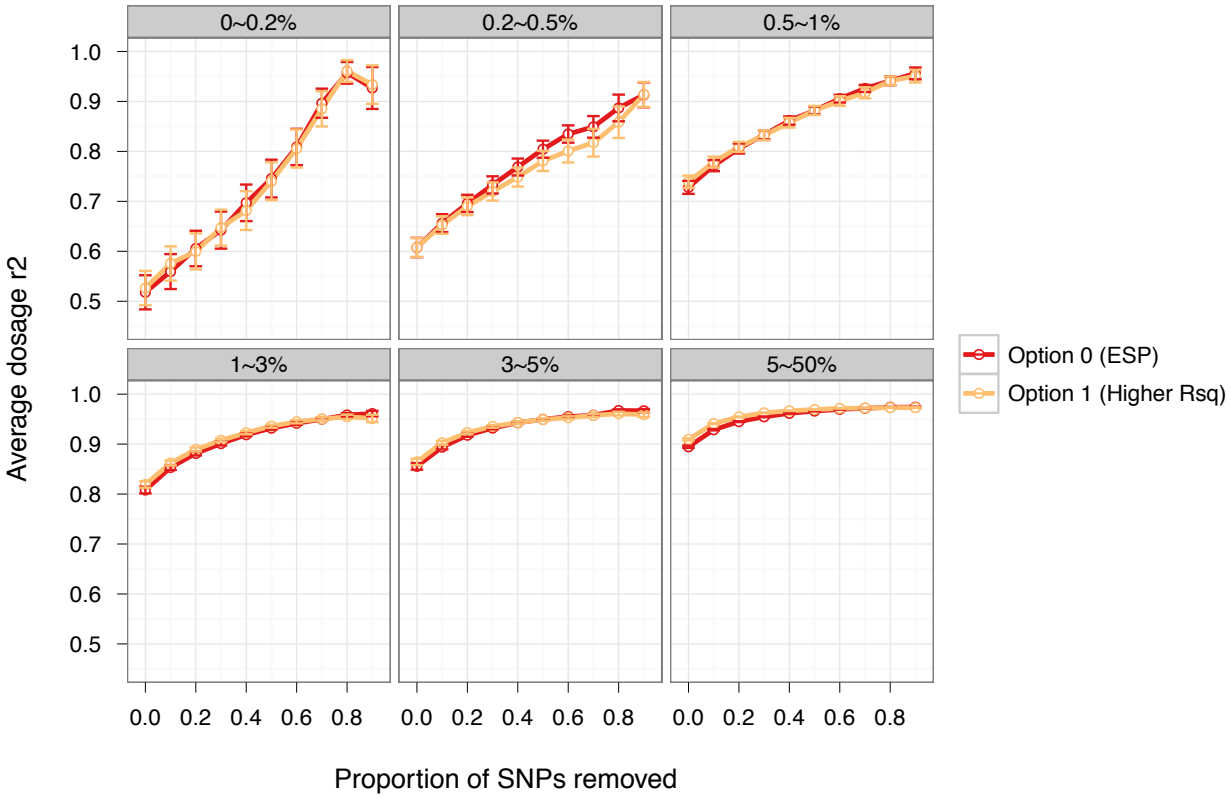


Figure 4.6. Comparison of dosage r^2 between Option 0 (ESP) and Option 3 (ESP+1000G) imputation. The *x*-axis is the proportion of SNPs that were removed based on elevated Rsq threshold (QC). The *y*-axis is the mean dosage r^2 (squared Pearson correlation between imputed dosages and experimental genotypes)

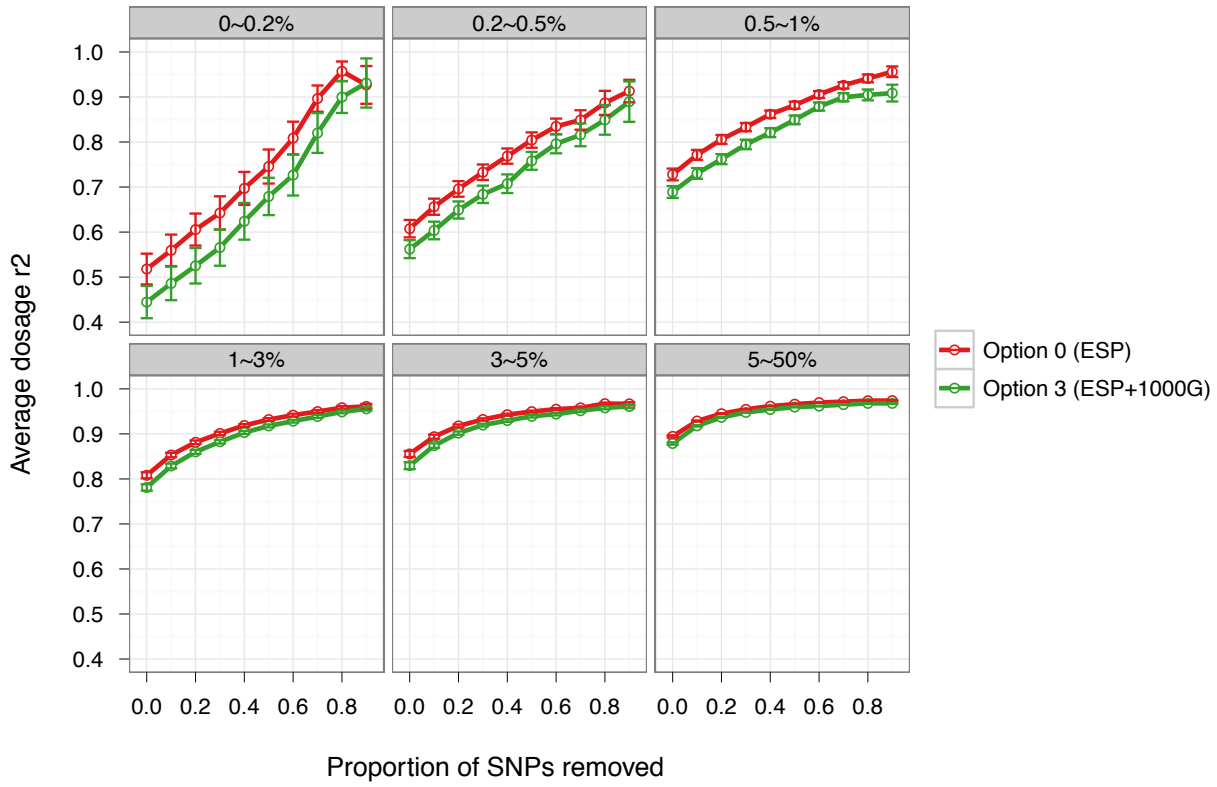


Figure 4.7. Comparison of dosage r^2 between IMPUTE2 results as well as minimac results.

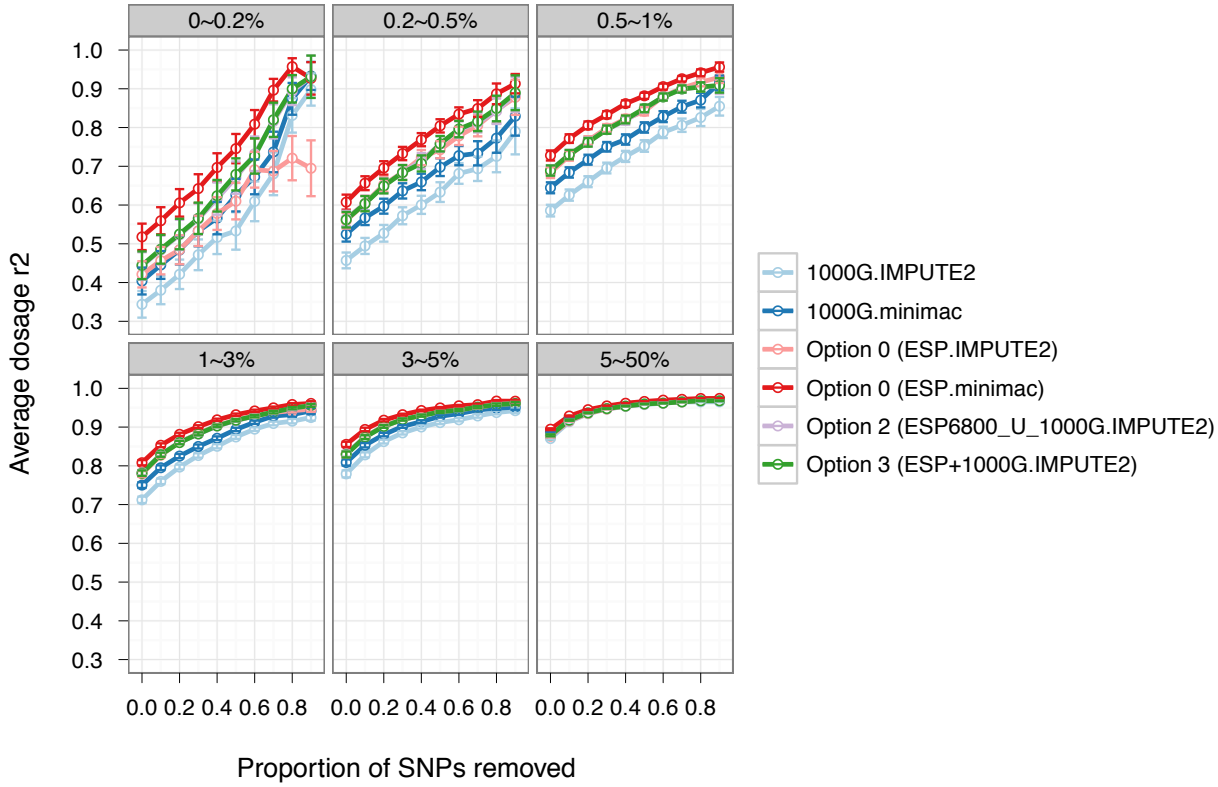


Table 4.1. Comparison of dosage r^2 between ESP imputation and 1000G imputation[a].

	1000G as reference		ESP as reference		Difference (ESP – 1000G)		Two-sided test	
MAF	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**	**Paired t-test**	**Paired Wilcoxon signed-rank test**
0-0.2%	0.404	3.49E-03	0.518	3.43E-03	0.114	2.92E-03	1.62E-04	9.20E-05
0.2-0.5%	0.525	1.33E-03	0.608	1.34E-03	0.083	9.59E-04	9.18E-09	2.70E-08
0.5-1%	0.644	8.41E-04	0.728	7.74E-04	0.083	6.92E-04	6.85E-12	3.68E-16
1-3%	0.750	2.59E-04	0.808	2.35E-04	0.058	1.92E-04	2.90E-24	1.31E-46
3-5%	0.809	3.39E-04	0.855	2.88E-04	0.047	2.31E-04	5.51E-17	2.42E-37
5-50%	0.887	4.58E-05	0.895	4.30E-05	0.008	2.88E-05	1.72E-06	4.48E-52

[a]All variants were included, irrespective of imputation quality Rsq.

Table 4.2. Comparison of dosage r^2 between ESP and 1000G full / relevant panels imputation[a].

MAF	Mean dosage r^2			Two–sided test	Mean dosage r^2			Two–sided test
	ESP (1092)	1000G (1092)	Difference (ESP-1000G)	Paired Wilcoxon signed-rank test	ESP (625)	1000G (625)	Difference (ESP-1000G)	Paired Wilcoxon signed-rank test
0-0.2%	0.528	0.404	0.113	**5.78E-05**	0.504	0.400	0.139	**1.50E-03**
0.2-0.5%	0.577	0.525	0.046	**4.15E-03**	0.532	0.505	0.010	8.85E-01
0.5-1%	0.705	0.644	0.061	**4.28E-09**	0.669	0.637	0.031	**2.62E-03**
1-3%	0.794	0.750	0.044	**7.81E-34**	0.776	0.751	0.025	**1.04E-14**
3-5%	0.846	0.809	0.037	**8.78E-32**	0.830	0.810	0.020	**8.02E-19**
5-50%	0.889	0.887	0.003	**4.69E-34**	0.881	0.887	-0.007	**6.52E-06**

[a]All variants were included, irrespective of imputation quality Rsq

Table 4.3. Number and percentage of well-imputed exonic variants.

MAF	Number (%) of well-imputed[a] markers		ESP:1000G ratio (Number of well-imputed)
	ESP	1000G	
0–0.2%	17 606 (31.8)	7713 (3.0)	2.28
0.2–0.5%	26 255 (70.0)	9283 (26.9)	2.83
0.5–1%	21 377 (92.1)	13 882 (62.9)	1.54
1–3%	29 784 (96.7)	26 466 (90.7)	1.13
3–5%	11 490 (96.9)	11 043 (96.0)	1.04
5–50%	40 500 (98.0)	39 849 (96.3)	1.02

[a]Well-imputed is defined such that the average Rsq of the QC + markers within each MAF category is >0.8.

Table 4.4. Imputability of blood trait associated variants reported in Auer et al.

					Rsq		
rs ID	**chr**	**position (b37)**	**Trait**	**Function**	**Early ESP***	**ESP**	**1000G**
rs334	11	5,248,232	hematocrit	missense	0.754	0.819	NA
rs9924561	16	314,780	hemoglobin	intron	0.535	0.869	0.919
rs13335497	16	310,005	hemoglobin	synonymous	0.595	0.899	0.949
rs11863726	16	230,578	hemoglobin	synonymous	0.518	0.633	0.653
rs35837297	2	136,594,439	WBC	missense	0.982	0.993	0.993
rs35940156	2	136,575,300	WBC	missense	0.981	0.994	0.990
rs17292650	1	43,803,807	platelet	missense	0.927	0.956	0.935
rs513349	6	33,541,719	platelet	intron	0.988	0.984	NA

*As reported in Auer et al, these Rsq's were from imputation using an earlier version of ESP containing 761 African Americans.

Table 4.5. Comparison of dosage r^2 between ESP.extreme and ESP.normal imputation.

	ESP.normal as reference		ESP.extreme as reference		Difference (ESP.normal – ESP.extreme)		Two-sided test	
MAF	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**	**Paired t-test**	**Paired Wilcoxon signed-rank test**
0-0.2%	0.458	3.97E-03	0.466	4.30E-03	0.041	3.26E-03	1.51E-01	**4.23E-02**
0.2-0.5%	0.557	1.44E-03	0.561	1.43E-03	-0.001	8.63E-04	9.35E-01	7.89E-01
0.5-1%	0.691	8.54E-04	0.692	8.30E-04	0.001	3.95E-04	8.63E-01	9.53E-01
1-3%	0.792	2.49E-04	0.785	2.54E-04	-0.006	7.84E-05	**6.18E-03**	**1.62E-02**
3-5%	0.844	3.02E-04	0.840	3.12E-04	-0.004	6.53E-05	**1.85E-02**	8.44E-02
5-50%	0.890	4.48E-05	0.886	4.55E-05	-0.004	6.47E-06	**2.25E-26**	**5.71E-30**

[a]All variants were included, irrespective of imputation quality Rsq.

Table 4.6. Comparison of dosage r^2 between option 2 and option 0[a].

	Option 0		Option 2		Difference (Option 2-Option 0)		Two-sided test	
MAF	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**	**Paired t-test**	**Paired Wilcoxon signed-rank test**
0-0.2%	0.518	3.43E-03	0.536	3.39E-03	0.018	1.94E-03	3.50E-01	1.54E-01
0.2-0.5%	0.608	1.34E-03	0.630	1.27E-03	0.023	4.40E-04	**4.57E-04**	**3.54E-02**
0.5-1%	0.728	7.74E-04	0.743	7.05E-04	0.015	2.01E-04	**8.88E-06**	**8.26E-05**
1-3%	0.808	2.35E-04	0.818	2.15E-04	0.010	6.04E-05	**7.75E-09**	**3.44E-07**
3-5%	0.855	2.88E-04	0.861	2.74E-04	0.005	5.84E-05	**7.38E-05**	**4.45E-05**
5-50%	0.895	4.30E-05	0.899	4.11E-05	0.004	5.77E-06	**8.46E-41**	**5.70E-80**

[a]All variants were included, irrespective of imputation quality Rsq.

Table 4.7. Comparison of dosage r^2 between option 1 and option 0[a].

	Option 0		Option 1		Difference (Option 1-Option 0)		Two-sided test	
MAF	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**	**Paired t-test**	**Paired Wilcoxon signed-rank test**
0-0.2%	0.518	3.43E-03	0.526	3.47E-03	0.008	1.26E-03	5.01E-01	4.57E-01
0.2-0.5%	0.608	1.34E-03	0.608	1.30E-03	0.0002	4.33E-04	9.78E-01	7.87E-01
0.5-1%	0.728	7.74E-04	0.739	7.35E-04	0.011	3.71E-04	6.97E-02	2.75E-01
1-3%	0.808	2.35E-04	0.819	2.21E-04	0.011	1.14E-04	**7.56E-04**	**4.30E-02**
3-5%	0.855	2.88E-04	0.864	2.76E-04	0.009	1.61E-04	**1.77E-02**	5.76E-01
5-50%	0.895	4.30E-05	0.909	3.90E-05	0.015	2.32E-05	**2.34E-30**	**4.94E-39**

[a]All variants were included, irrespective of imputation quality Rsq.

Table 4.8. Comparison of dosage r^2 between option 3 and option 0[a].

	Option 0		Option 3		Difference (Option 3-Option 0)		Two-sided test	
MAF	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**	**Paired t-test**	**Paired Wilcoxon signed-rank test**
0-0.2%	0.518	3.43E-03	0.444	3.60E-03	-0.073	2.12E-03	**7.70E-04**	**1.87E-04**
0.2-0.5%	0.608	1.34E-03	0.562	1.38E-03	-0.043	6.24E-04	**3.60E-06**	**2.92E-06**
0.5-1%	0.728	7.74E-04	0.689	7.83E-04	-0.039	3.04E-04	**4.09E-13**	**8.74E-16**
1-3%	0.808	2.35E-04	0.781	2.50E-04	-0.027	1.02E-04	**4.16E-19**	**7.89E-48**
3-5%	0.855	2.88E-04	0.829	3.27E-04	-0.028	9.61E-05	**1.17E-30**	**3.27E-47**
5-50%	0.895	4.30E-05	0.879	4.83E-05	-0.016	1.14E-05	**4.23E-127**	**3.59E-208**

[a]All variants were included, irrespective of imputation quality Rsq.

Table 4.9. Comparison of dosage r^2 between IMPUTE2 results[a].

MAF	Option 0 (ESP IMPUTE2) Mean	Option 2 (ESP_U_1000G IMPUTE2) Mean	Option3 (ESP+1000G IMPUTE2) Mean	Option 2 – Option 0		Option 3 – Option 0	
				Difference	Two-sided paired Wilcoxon signed-rank test	Difference	Two-sided paired Wilcoxon signed-rank test
0-0.2%	0.421	0.443	0.444	0.0215	3.99E-01	0.023	4.21E-01
0.2-0.5%	0.560	0.565	0.562	0.0046	9.63E-01	0.005	9.27E-01
0.5-1%	0.684	0.689	0.689	0.0057	6.48E-01	0.005	6.83E-01
1-3%	0.779	0.782	0.781	0.0034	1.27E-01	0.003	1.56E-01
3-5%	0.830	0.831	0.829	0.0013	1.08E-01	-0.0005	5.85E-02
5-50%	0.876	0.879	0.879	0.0026	**3.89E-06**	0.003	**1.16E-06**

[a]All variants were included, irrespective of imputation quality Rsq.

CHAPTER 5: ROBUST AND POWERFUL TWO-STEP TESTING PROCEDURE FOR ASSOCIATION ANALYSIS IN ADMIXED POPULATIONS

5.1 Introduction

Genome-wide association studies (GWAS) have been successful in improving our understanding of the genetic basis of numerous heritable diseases and quantitative traits (Visscher *et al.*, 2012). Although GWAS have initially been performed with individuals of European ancestry, the field has expanded to non-European populations (Rosenberg *et al.*, 2010). Performing genetic association analysis in diverse populations allows us to gain deeper understanding of the genetic architecture of human diseases and traits, through assessing the generalizability of risk variants (R. Chen *et al.*, 2012; Ioannidis *et al.*, 2004), narrowing down the location of the functional variants over the risk region (Helgason *et al.*, 2007; International HapMap, 2005) and identifying novel disease loci which are absent or in low frequency in European population (Rosenberg *et al.*, 2010). In the US, genetically admixed populations have been receiving increasing attention. Whereas, insufficient genetic association studies have been carried out.

Association studies in admixed populations impose challenges due to the unique LD patterns resulted from admixture process, where gene flow occurs between two or more distinct populations (ancestral populations). Consequently, admixed chromosomes can be viewed as mosaic segments (local ancestry) originating from each of the ancestral populations (Shriner, 2013). Therefore, admixed population exhibits two forms of LD – LD due to genetic linkage

(ancestry LD) and that due to admixture process (admixture LD) (Chakraborty & Weiss, 1988). The unique LD patterns enable admixture mapping, a method taking advantage of the extended admixture LD to scan for the association between chromosome segments of certain ancestry and traits, with the assumption that the functional variants leading to increased risk have higher frequencies in chromosomal segments inherited from the ancestral population with higher disease risk (Chakraborty & Weiss, 1988; Shriner, 2013). Admixture mapping is particularly effective in detecting genetic loci with differential risk in the parental populations and requires the use of only a small panel of ancestry informative makers (AIMs). In spite of the advantages, the coarse scale of the admixture LD results in low resolution for gene mapping (Winkler *et al.*, 2010).

With dense set of genotyped markers, GWAS provides high-resolution gene mapping. GWAS in admixed populations, however, exhibits special challenges from the presence of both ancestry LD and admixture LD, which may cause population stratification, thereby leading to spurious associations or false negatives (Kittles *et al.*, 2002; J. Liu *et al.*, 2013; Mao *et al.*, 2013; Qin *et al.*, 2010; X. Wang *et al.*, 2011; Zhang & Stram, 2014). To meet the challenges, it is important to have proper treatment of local ancestry when conducting GWAS in admixed populations.

As a result, advanced local ancestry inference methods have been developed to infer local ancestry of markers from high throughput genotyping data in admixed samples (Baran *et al.*, 2012; Patterson *et al.*, 2004; Price *et al.*, 2009). With the estimated local ancestry for each marker and the derived global ancestry, association studies have been performed by adjusting for local ancestry as a covariate (Levin *et al.*, 2014), by conducting SNP and admixture mapping

separately using the same set of markers (Z. Chen *et al.*, 2013; Reiner *et al.*, 2012) or by selecting a subset of African ancestry individuals for subsequent analysis (J. Li *et al.*, 2013).

Theoretically, genetic and ancestry effect contain independent information, thus combining the two types of information may lead to increased statistical power. Tang et al. demonstrated this in a joint testing procedure by integrating the two types of information in a family-based study (Tang *et al.*, 2010). Later, Pasaniuc et al. developed a joint test (MIXSCORE) in case-control GWAS setting (Pasaniuc *et al.*, 2011). This method has been applied to a fine-mapping study in African Americans (Levin *et al.*, 2014). Both methods show the usefulness of combining SNP and ancestry information assuming consistent SNP effect in the ancestral populations. However, effect heterogeneity may present when GWAS is conducted across diverse populations due to differential LD between ancestral populations (J. Liu *et al.*, 2013). For example, with a sample size of >80% power to detect significant association with equal effect size in a Caucasian cohort, association analysis in African Americans fails to replicate the validated associations identified in Caucasians (Frazier-Wood *et al.*, 2013). Thus, to guard against missing important signals, Liu et al. included SNP by local ancestry interaction in a logistic regression model and showed increased power when substantial differential LD between ancestral populations exists (J. Liu *et al.*, 2013).

In this study, we propose a robust and powerful two-step testing procedure for association studies in African Americans. In the first step of the testing procedure, we jointly test allele effect, ancestry effect and the existence of effect heterogeneity in a regression model. The joint test guards against missing important associations from any of the sources, as the true underlying genetic architecture at each locus is unknown. The significant signals are carried on to the second step where we narrow down the source of association through a one-time model selection

process. We model the interaction between allele and ancestry by taking advantage of the joint distribution of allele and ancestry. It not only captures the effect heterogeneity among ancestral populations, but is more powerful than simply modeling the interaction using a cross-product term, particularly when local ancestry is estimated. In the present study, we assess the power and type I error of the proposed testing procedure and existing ones by conducting extensive simulations mimicking a broad spectrum of real life scenarios ranging from one extreme of effect solely due to allele effect, to the other extreme of solely ancestry effect and anywhere in between, as well as differential allelic effect across local ancestry groups. In addition to using true simulated local ancestry, we demonstrate the robustness of our results with estimated local ancestry in simulated data. Finally, we applied the two-step testing procedure to a genome-wide association analysis with hemoglobin using African American data from CARe, where we further illustrated the usefulness of our proposed testing procedure through verifying previous findings.

5.2 Methods

5.2.1 Simulation of admixed samples and reference haplotypes

We simulated samples of admixed African and European ancestry and African and European reference haplotypes using COSI (Schaffner *et al.*, 2005), a genotypic data simulator based on coalescent population genetic model. It is well calibrated to closely resemble empirical data in allele frequency, linkage disequilibrium and population differentiation (Schaffner *et al.*, 2005). We first generated 3000 African haplotypes and 3000 European haplotypes, 50kb each, to serve as the parental ancestry populations. Then 1000 haplotypes from each of the parental populations was randomly selected and kept as reference haplotypes. The remaining 2000 African and 2000 European haplotypes were randomly combined to generate 1000 African

American samples. None of the parental ancestry haplotypes were reused. We constructed admixed samples based on the empirical estimate of 0.2125 switch points per 50kb region (Wegmann *et al.*, 2011). To be specific, we first generated 425 African American chromosome segment containing one switch point by joining one randomly selected African chromosome segment and one randomly selected European chromosome segment. The switch point assignment was weighted by the recombination rate from the crossover map generated by COSI. We used a binomial distribution with probability 0.5 to decide which ancestral population went first. Next, from the pool of unused chromosome segments, we randomly selected 1408 African and 167 European ones, together with the 425 African American chromosome segments, to generate a collection of 2000 African American chromosome segments. The rates of switch point occurrence and proportion African ancestry were consistent with the findings presented in (Parra *et al.*, 1998; Wegmann *et al.*, 2011). Afterwards, we randomly paired the 2000 haploid African American chromosome segments to yield 1000 diploids. 1000 replicates of this process were performed.

5.2.2 Simulation of quantitative traits

For each locus, we simulated quantitative traits for the 1000 African American samples based on a null model or a causal model. Our null model consists of two independent covariates, E1 and E2. Let $\mathrm{QT_i} = 0.5\mathrm{E_{1i}} + 0.5\mathrm{E_{2i}} + \epsilon_\mathrm{i}$, where $\mathrm{E_1} \sim \mathrm{Bernoulli}(0.5)$and $\mathrm{E_2} \sim \mathrm{Normal}(0,1)$. Error term has a standard normal distribution and is independent across individuals.

As summarized in **Table 5.1**, we simulated four scenarios for the causal model. In Scenario 1, only ancestry effect (γ) presents, which varies from 0.1 to 1 and the allele effect (β) equals 0. In Scenario 2, both ancestry and allele effect exist while allele effect is driven by ancestry effect by a factor of k. The value of k depends on the minor allele frequency difference

(δ). Scenario 3 has only allele effect (β), which varies from 0.1 to 1 and γ equals 0. Both ancestral populations have the same effect direction. In Scenario 4, γ equals 0 and β varies from 0.1 to 1. The effect direction is opposite in African and European ancestral populations.

Power was evaluated based on 10000 experiments. To be specific, in each of the 1000 regions, 10 markers were randomly selected with replacement for each allele frequency difference category. Then causal models were applied to generate quantitative traits under alternative hypothesis. Type I error was assessed based on 100000 experiments. Specifically, in each of the 1000 regions, 100 markers were randomly selected with replacement for each allele frequency difference category. Then the null model is used to generate the quantitative traits under null.

5.2.3 Inference of local ancestry using HAPMIX

We used HAPMIX to predict local ancestry with reference from the 1000 Genomes Project (Phase I, March 2013 release). HAPMIX provides highly accurate local ancestry estimates in two-way admixed sample by leveraging LD within populations based on Hidden Markov Model (Price *et al.*, 2009). We used the default parameters and "Diploid" mode. Instead of outputting, by default, the expected probability of 0, 1 or 2 copies of European ancestry at each SNP, we obtained the inferred joint distribution of local ancestry and genotype by setting "output_details" to "prob" (see **Figure 5.1** for an example). The probabilities from the joint distribution allow us to calculate the expected copies of reference alleles (i.e., genotype, ranging from 0 to 2), expected copies of African ancestry alleles (i.e., local ancestry, ranging from 0 to 2) and expected copies of African ancestry reference alleles (ranging from 0 to 2). Since the 16 probabilities of each marker may not sum up to 1, we did a conditional adjustment for each probability to make sure the summation is up to 1. Accuracy of HAPMIX local ancestry

estimation was evaluated by calculating the Pearson correlation between the estimated and true values.

5.2.4 Association tests

We evaluated the performance of the first step of our proposed testing procedure (T4) and three other existing methods (T1-3) by comparing power and type I error in the four simulated scenarios. In scenario 4 where effect heterogeneity presents, we compared our interaction model (T4) with the traditional one (T5) including the cross-product term.

- T1: $E(Y) = \alpha_0 + \alpha_{E_1}E_1 + \alpha_{E_2}E_2 + \alpha_G G + \beta X_{ref}$, where E_1 and E_2 are two simulated covariates, G is the estimated global African ancestry (half of the average African alleles per person) and X_{ref} is the number of reference alleles at each locus and we test for $\beta = 0$. T1 is the test that commonly used in GWAS to examine whether there is an allele effect on the phenotype, assuming an additive genetic model.
- T2: $E(Y) = \alpha_0 + \alpha_{E_1}E_1 + \alpha_{E_2}E_2 + \alpha_G G + \gamma X^{afr}$, where X^{afr} is the number of African ancestry alleles at each locus and we tests for $\gamma = 0$. T2 is the statistical test that is used in admixture mapping to scan for the ancestry effect of variants with frequency disparity between African and European populations.
- T3: $E(Y) = \alpha_0 + \alpha_{E_1}E_1 + \alpha_{E_2}E_2 + \alpha_G G + \beta X_{ref} + \gamma X^{afr}$ and we test for $\beta = \gamma = 0$. T3 jointly tests for allele and ancestry effect, which leverages the dense set of genotyped markers while does not sacrifice power due to multiple testing correction when testing for the two effects separately.

- T4: $E(Y) = \alpha_0 + \alpha_{E_1}E_1 + \alpha_{E_2}E_2 + \alpha_G G + \beta X_{ref} + \gamma X^{afr} + \eta X_{ref}^{afr}$, where X_{ref}^{afr} is the number of African ancestral reference allele at each locus. T4 is an extension of T3 by incorporating the interaction between genotype and ancestry. Taken advantage of the allele and ancestry joint distribution provided by HAPMIX, we parameterize the interaction as the allele-specific ancestry estimate to propose a two-step testing procedure. In the first step, we tests for $\beta = \gamma = \eta = 0$. Since the true underlying disease locus and LD between the tested marker and causal locus are unknown, testing all three terms simultaneously (T4) is especially powerful with little issue of false negatives.
- T5: $E(Y) = \alpha_0 + \alpha_{E_1}E_1 + \alpha_{E_2}E_2 + \alpha_G G + \beta X_{ref} + \gamma X^{afr} + \eta X_{ref} * X^{afr}$, which captures the present of effect heterogeneity using traditional cross-product interaction term.

We evaluated the power of our proposed test (T4) and other tests using both true and estimated local ancestry. Power was calculated by counting the number of times in the 10000 experiments when *P*-value is less than the GWAS significance threshold of 5×10^{-8} for model T1, T3, T4 and T5 or less than the significance threshold of 7×10^{-6} for admixture mapping model T2 (Reiner *et al.*, 2012).

Next, we evaluated the power of the second step of the proposed testing procedure, that is, the proportion of times the source of association can be correctly identified. To do so, we traced the source of association through a one-time model selection by comparing the absolute value of the test statistics associated with β, γ and η among the significant loci from the first step.

5.2.5 CARe data set

The Candidate-gene Association Resource (CARe) consortium has been previously described (Musunuru *et al.*, 2010). We applied our proposed testing procedure to study associations in 5711 African American participants from two CARe cohorts, Atherosclerosis Risk in Communities (ARIC) and Coronary Artery Risk Development in young Adults (CARDIA). These cohorts have been previously described (Bild *et al.*, 2002; Friedman *et al.*, 1988). All samples were genotyped using Affymetrix Genome-Wide Human SNP Array 6.0 Chip at the Broad Institute of MIT and Harvard. Markers with genotype call rates < 90%, or Hardy-Weinberg exact test P-value < 1×10^{-6}, or MAF < 1% were removed.

5.2.6 WHI-SHARe data set

We replicated our findings in a cohort of 8087 African American participants from Women's Health Initiative SNP Health Association Resource (WHI-SHARe) study. All samples were genotyped using the Affymetrix 6.0 genotyping platform. Prior to local ancestry inference, we removed Affymetrix 6.0 SNPs with genotype call rates < 90%, or Hardy-Weinberg exact test P-value < 1E-06, or MAF < 1%. Quality control details were described previously (Reiner *et al.*, 2012; Reiner *et al.*, 2011).

5.3 Results

5.3.1 Power evaluation with simulated data using true local ancestry

An advantage of performing association analysis in admixed populations is that it allows the identification of risk variants leading to disease disparity, which have substantial allele frequency difference, even monomorphic with different alleles, between two parental populations. For example, DARC null allele for white blood cells (Lautenberger *et al.*, 2000),

SLC24A5 for skin pigmentation (Lamason *et al.*, 2005) and APOL1 for kidney disease (Genovese *et al.*, 2010). Such association may be missed if genetic analysis is only conducted in one homogeneous population.

Bearing this in mind, we stimulated Scenario 1, where association is found only in one parental population. To be specific, we assumed that the causal allele presents solely in African ancestral populations with a fixed allele frequency of 1, leading to raised mean trait value, and the non-risk allele is monomorphic in European descendent populations. We also assumed that the tested SNP, which is not causal, locates in close vicinity of the causal variant (with strong admixture LD) but in complete ancestry linkage equilibrium with it, so that the SNP-trait association is driven by the local ancestry of the tested SNP which can hardly be captured by testing for SNP effects alone. As expected, T2, T3 and T4, which designed to test for the ancestry effect, are well powered to detect this association at modest effect size, regardless of the allele frequency differences (**Figure 5.2**). T2 shows advantage in this setting mainly due to lower multiple testing burden. As to control an overall type I error rate of 5%, the significance threshold of admixture mapping is set at 7×10^{-6} rather than a typical GWAS significant threshold of 5×10^{-8}, which is used by all the other tests. In this scenario, statistical power is not affected much by allele frequency difference, as the association is driven by local ancestry rather than genotype.

Based upon Scenario 1, we simulated a more realistic case (Scenario 2) by allowing allele effect. Similar to the previous scenario, the causal allele presents only in African ancestral populations with a nearly fixed allele frequency of 1, contributing to the elevated mean trait value. Again, the tested SNP, non-causal, is in strong admixture LD with the causal variant. In Scenario 2, we assumed that the tested SNP is in low to moderate ancestry LD with the causal

allele. The strength of the ancestry LD is reflected by the allele frequency difference of the tested SNP in African and European populations. To be specific, similar allele frequency in the two parental populations indicates low LD between the causal and the tested SNP, a case similar to that in Scenario 1. As the frequency difference increases, the ancestry LD between the causal and the tested SNP becomes stronger, where we would observe both ancestry and allele effect. Thus, whether or not the allele effect can be captured depends on the allele frequency difference. As shown in **Figure 5.3**, when allele frequency difference is small (0-0.03), the observed pattern is similar to that in Scenario 1. The power of T2 does not change with the increase of allele frequency difference, as it only tests for ancestry effect. T3 and T4, which tests for both SNP and ancestry effect, have the greatest power gain with the growing allele frequency difference and become more powerful than T2 for moderate effect size when allele frequency difference is greater than 0.2. T3 and T4 have comparable power in this scenario.

Often times, the risk variants identified in European population are transferable to other ethnicities with consistent effect direction and magnitude (Loth *et al.*, 2014; Teslovich *et al.*, 2010). Therefore, in Scenario 3, we simulated a case where African and European ancestral populations share the same risk allele with similar effect size and direction. It happens when the LD between the causal and the tested SNP share the same direction in the parental populations. Difference in mean trait values in the two populations is caused by the difference in allele frequency. **Figure 5.4** shows the power comparison in Scenario 3 for effect size ranging from 0.1 to 1 stratified by allele frequency difference. T1, T3 and T4 have comparable performance with the increase of effect size. The 1-df test (T1) is slightly more powerful than the 2- and 3-df test (T3 and T4). As expected, T2 has dramatic power loss, as it fails to capture the SNP effect. For a fixed effect size, power tends to increase with the growing differential allele frequencies.

On the contrary, effect heterogeneity has been reported across populations, which motivated us to simulate Scenario 4 (Frazier-Wood *et al.*, 2013). In this scenario, association can be found in both ancestral populations while the LD between the causal and the tested SNP in the two ancestral populations is in the opposite direction, leading to effect heterogeneity in the two populations. Although this may only happen to a small fraction of loci, it is worth being taken into account to ensure full understanding of the underlying genetic structure in worldwide populations. We compared the power across a broad spectrum of effect sizes and allele frequency differences (**Figure 5.5**). In this scenario, T4 outperforms all three other tests. With a moderate effect size ($\beta = 0.5$), T4 achieves a power gain of 168%, 589% and 150% for markers with frequency difference between 0.05 and 0.1.

Once a significant signal is found by T4, it is necessary to identify which component contributes to the association. Thus, as the second step of the proposed testing procedure, we aims to narrowing down the primary source of association by performing a single model selection by comparing the absolute values of the test statistics of the allele, ancestry and ancestry-specific allele effect. To examine the power of this second step, we calculated the fraction of times that an effect is correctly identified in each of the four scenarios, given the tested SNP passes the significance threshold in the first step. **Table 5.2** shows the average proportion of times that the source of association is correctly identified across small ($\beta = 0.3$) to large ($\beta = 1$) effect size. It is noteworthy that when effect heterogeneity exists (Scenario 4), this step is powerful to identify the ancestry-specific risk allele achieving a power of over 90% with mean 93.3%. In addition, it is efficient (> 95%) to identify ancestry effect as the only source of association (Scenario 1) and ancestry with allele effect (Scenario 2). The power to detect allele effect as the only source of association (Scenario 3) is slightly lower but above 80%.

Importantly, modeling interaction term by taking advantage of the joint distribution of ancestry and allele (T4) achieves higher power than using traditional interaction model (T5). In T4, effect heterogeneity between two parental populations is captured by the number of African ancestry reference alleles, rather than the cross-product of number of reference alleles and African ancestry alleles. Thus T4 contains more precise information leading to enhanced power in testing interactions. **Table 5.3** displays the ratio of the power of T4 to that of T5 in Scenario 4. T4 has higher power than T5 across different beta values and allele frequency differences, particularly when beta is small. This observation holds in both Step 1 and Step 2.

5.3.2 Type I error

We used 100000 experiments to evaluate the type I error at the nominal significance level $\alpha = 0.05, 0.01, 0.001$ in the null model ($\beta = \gamma = \eta = 0$). Results are summarized in **Table 5.4**. Across allele frequency difference categories, all tests appropriately control the type I errors.

5.3.3 Power in simulated data with estimated local ancestry

In the above four scenarios, particularly Scenario 4, we show using simulated data and true local ancestry that our testing procedure incorporating allele specific ancestry is powerful. In real data analysis, however, the local ancestry and allele specific ancestry are unknown. Thus the usefulness of the proposed test largely depends on the accuracy of the estimated ancestry. For this reason, next we examine whether the observed patterns remain if we use inferred ancestry and how much power loss would there be.

We first evaluated the accuracy of the estimated local ancestry and allele specific ancestry by calculating the Pearson correlation coefficient between the true and inferred ones. As shown in **Table 5.5**, the median correlation for the estimated genotype, African allele and African reference allele is above 0.8. Note that the inferred African reference allele is highly

accurate. For markers with MAF less than 0.05 in African population, we notice decrease in the accuracy of allele specific ancestry estimates but the median correlation is still above 0.8.

Having confirmed that local ancestry can be inferred with adequate accuracy at both marker and allele level, we performed power analysis using inferred local ancestry with the same simulated data set. Then we compared the results with what obtained earlier using true ancestry. Overall, the pattern remains using inferred local ancestry in Scenario 1 through 4. **Figure 5.6** shows the evaluation performed in Scenario 4. As expected, although using inferred ancestry incurs slightly power loss, the pattern observed earlier remains.

We also examine the fraction of times a source of association is correctly identified using inferred local ancestry and allele specific ancestry in the second step of our proposed testing procedure. Similar results are obtained as those using true ancestry.

When comparing with T5, T4 achieves higher power using estimated local ancestry (**Table 5.6**). Note that the better performance of T4 is more evident when using estimated local ancestry than true local ancestry. This may be due to the more accurate estimation of African reference allele than African ancestry as shown in **Table 5.5**.

5.3.4 Application to real phenotypes

We applied the proposed 2-step testing procedure genome-wide to test for the association between one of the hematological traits, hemoglobin, and genetic variants in African American samples from CARe project. Many genetic loci has been reported to be associated with hemoglobin in Europeans (van der Harst *et al.*, 2012), East Asians (Kamatani *et al.*, 2010) and African Americans (Auer *et al.*, 2012; Auer *et al.*, 2014; Z. Chen *et al.*, 2013; Lo *et al.*, 2011).

In this analysis, we first perform a genome-wide 3-df joint test (T4) adjusting for age, proportion of global African ancestry, cohort and smoking status. Then in the second step, we

conduct a single model selection by comparing the test statistics of allele, ancestry and ancestry specific allele effect. We consider a suggestive P-value threshold of 1×10^{-6}. Our analysis strategy was validated by the replication of previously reported hemoglobin associated loci. As shown in **Table 5.7**, SNP rs9940149 and rs7199221 on chromosome 16 show a strong allele effect, indicating similar effect in African and European ancestral populations. It is worth noting that rs1211375 on chromosome 16 shows a strong ancestry-specific allele effect, suggesting effect heterogeneity at this locus in African and European ancestral populations. This differential effect is replicated in WHI-SHARe data set (P-value = 5.78×10^{-6}, allele effect = 0.54, ancestry effect = 0.50 and African ancestry allele effect = -1.04).

5.4 Discussion

Association studies in admixed populations bring opportunities to gain deeper understanding of genetic architecture of complex diseases and traits, whereas require special treatment of the local ancestry due to differential origin of the chromosomal segments. In this study, we propose a robust and powerful two-step testing procedure for association analysis in African Americans. We demonstrate its usefulness in extensive numerical simulations and real data analysis.

The underlying genetic structures are unknown and can be complex in admixed populations. Therefore, we simulated data covering 4 scenarios that reflect the complexity of the real admixed data set, which includes only ancestry effect (Scenario 1), both ancestry and allele effect (Scenario 2), only allele effect but homogeneous in the two parental populations (Scenario 3) and only allele effect but heterogeneous (Scenario 4). We also provided justifications of the relationship between the tested SNP and causal variant. As shown in the results, without prior knowledge of the form of association, T4, which tests for allele, ancestry and their interaction

simultaneously, achieves competitive performance in all scenarios and is superior when effect heterogeneity exists. As expected, power of the tests depends on the underlying genetic structure. Generally, power is slightly greater when the test matches the simulated underlying model, although the power loss for mis-specifying the model is small. When ancestry effect is the only source of association, T2 is more powerful than other tests which may be due to 1) its matching with the simulated model, 2) the lower significant threshold (7×10^{-6}). When homogeneous allele effect is the only source of association, T1 has the best performance due to its matching with the simulated model and T4 has slightly power loss.

The second step in the 2-step testing procedure can help tease apart the source of association. From our results, we show that it is highly powerful at the presence of a strong ancestry by allele interaction. The advantage of our parameterization of the interaction term as the number of African reference alleles is that it is more powerful to detect the presence of effect heterogeneity than the traditional cross-product interaction term. The enhanced power comes from more precise capture of the local ancestry information down to the allele level. This advantage is more evident when local ancestry is estimated, as the estimation accuracy of the number of African reference alleles is higher than that of the number of African alleles, which is used in the cross-product term.

We use a GWAS significance threshold of 5×10^{-8} for all tests involving allele effect (T1, T3 and T4) and 7×10^{-6} for T2, the test used in admixture mapping. The rational of using a substantially lower critical value for genome-wide significance of admixture mapping is based on previous theoretical analysis and simulation results that a threshold of 7×10^{-6} provides a genome-wide type I error of 0.05, because of the extensive correlation in local ancestry in admixed populations (Tang *et al.*, 2006). However, it is still unclear whether we should adjust

the GWAS significance threshold accordingly for other tests because of the reduced number of independent tests in admixed populations.

The local ancestry and allele-specific ancestry are estimated using HAPMIX and our assessment shows the estimation is highly accurate. However, the accuracy comes with the cost of high computation demand. To cope with the computational burden, we restricted our reference to the overlapping markers between the reference and genotyped markers, allowing only imputation for sporadic missingness among the genotypes. As a result, our analysis is limited to typed markers. To obtain a finer resolution, one may apply a suggestive P-value, e.g., 1×10^{-6}, to obtain the candidate region and follow it up by fine mapping. Through restricting to a small region, it is more computationally feasible to infer the local ancestry for both typed and imputed markers. We learned from our practice that it is necessary to raise the mutation rate parameter by a factor of 10 when using sequencing based reference haplotypes from the 1000 Genome project, as is recommended in the HAPMIX tutorial. After using a higher mutation rate, we obtain similar local ancestry estimation as is reported previously (Reiner *et al.*, 2012). A caveat of using the joint distribution generated by HAPMIX is that all samples need to be included to ensure the same reference allele for each individual when allele-specific ancestry is obtained. We paralleled the process by modifying HAPMIX into three separate steps including 1) generate reference related and genotype related files; 2) infer local ancestry for each sample and 3) clean up the temporary files keeping only the output files. The first step needs to include all samples to ensure the same reference allele is used, which is the most memory-consuming step. The second step can be run in parallel individually, which dramatically speeds up the inference process if one has thousands of samples.

We demonstrated the usefulness of our two-step testing procedure through testing for the association with hemoglobin using data from CARe. In the first step, we replicated previous findings and in the second step, we are able to pinpoint the primary source of association. When a SNP passing GWAS significance in the first step, it is advisable to check the MAF of the tested SNP in each of the reference populations (e.g., AFR and EUR in the 1000 Genomes project). We noticed that a marker with MAF zero in one reference population may be identified to be significant with a strong interaction term, which is actually an artifact due to local ancestry estimation error.

Figure 5.1. HapMix output (per marker per subject) as joint probability of genotype and local ancestry. "A" and "E" represent allele ancestry of AFR and EUR, respectively. "1" and "0" represent minor and major alleles, respectively. Copies of minor alleles are calculated by summing up the number of minor alleles of African and European ancestry. Copies of African ancestry alleles are calculated by summing up the number of minor and major alleles of African ancestry.

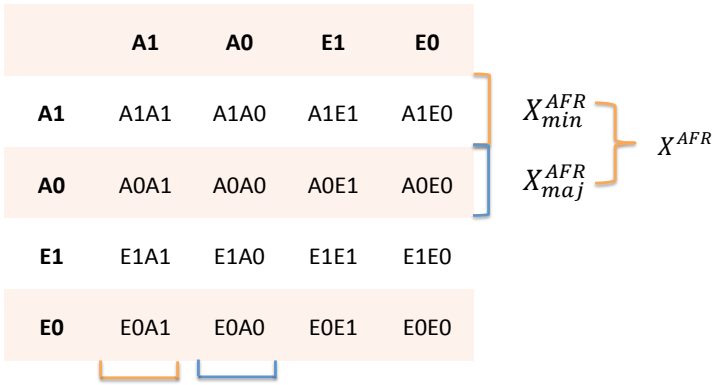


Figure 5.2. Statistical power of the four tests in Scenario 1 where African ancestral alleles associated with trait. Power is plotted as a function of effect size stratified by allele frequency differences between African and European ancestral populations.

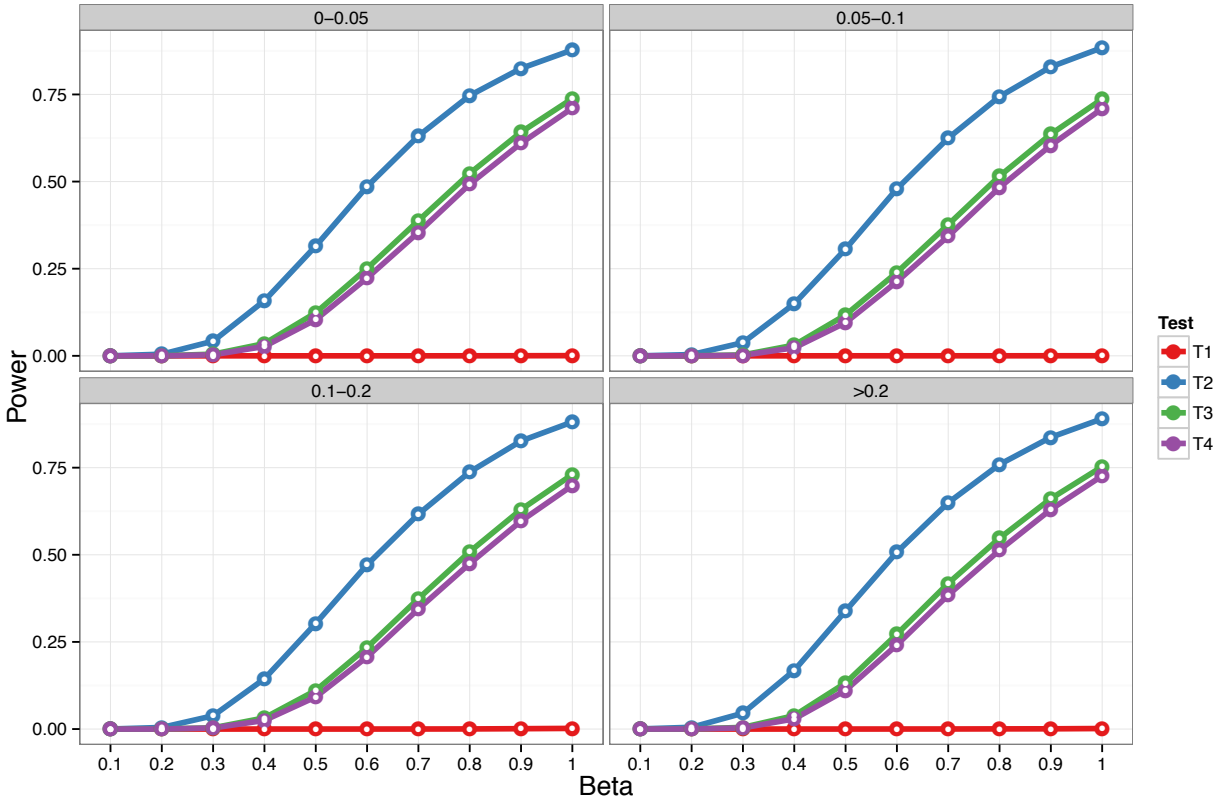


Figure 5.3. Statistical power of the four tests in Scenario 2 where risk alleles presents only in African ancestral population with weak to moderate ancestry LD with the tested SNP. Power is plotted as a function of effect size stratified by allele frequency difference between African and European ancestral populations.

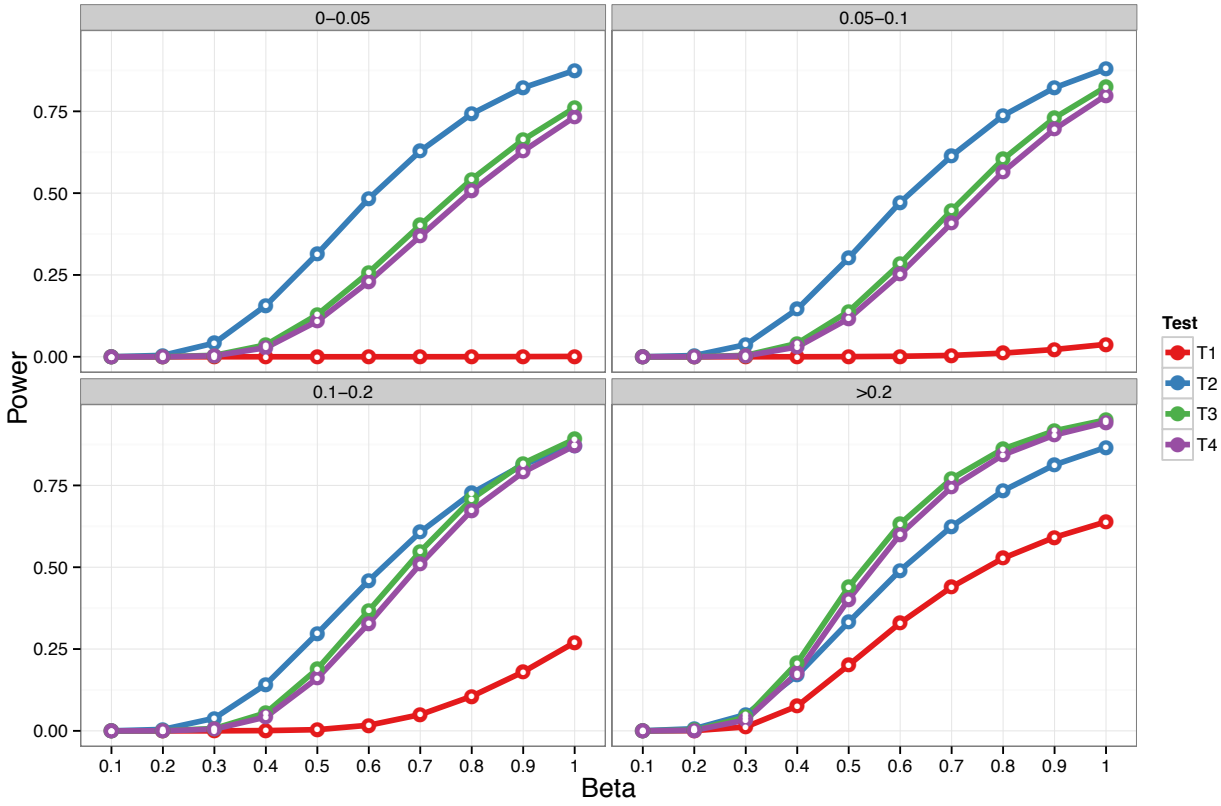


Figure 5.4. Statistical power of the four tests in Scenario 3 where risk alleles associate with trait both in African and European ancestral population with consistent effect size and direction. Power is plotted as a function of effect size stratified by allele frequency differences between African and European ancestral population.

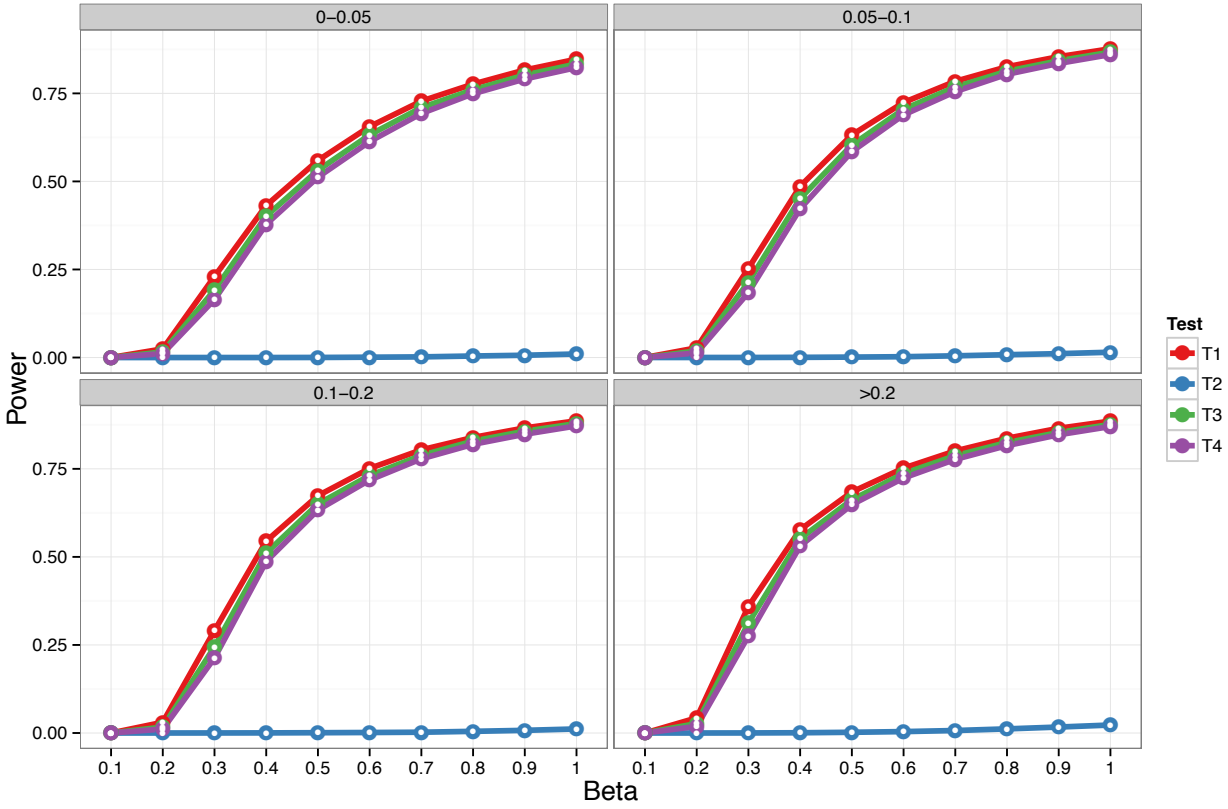


Figure 5.5. Statistical power of the four tests in Scenario 4 where risk alleles associate with trait both in African and European ancestral populations with inconsistent effect direction. Power is plotted as a function of effect size stratified by allele frequency difference between African and European ancestral populations.

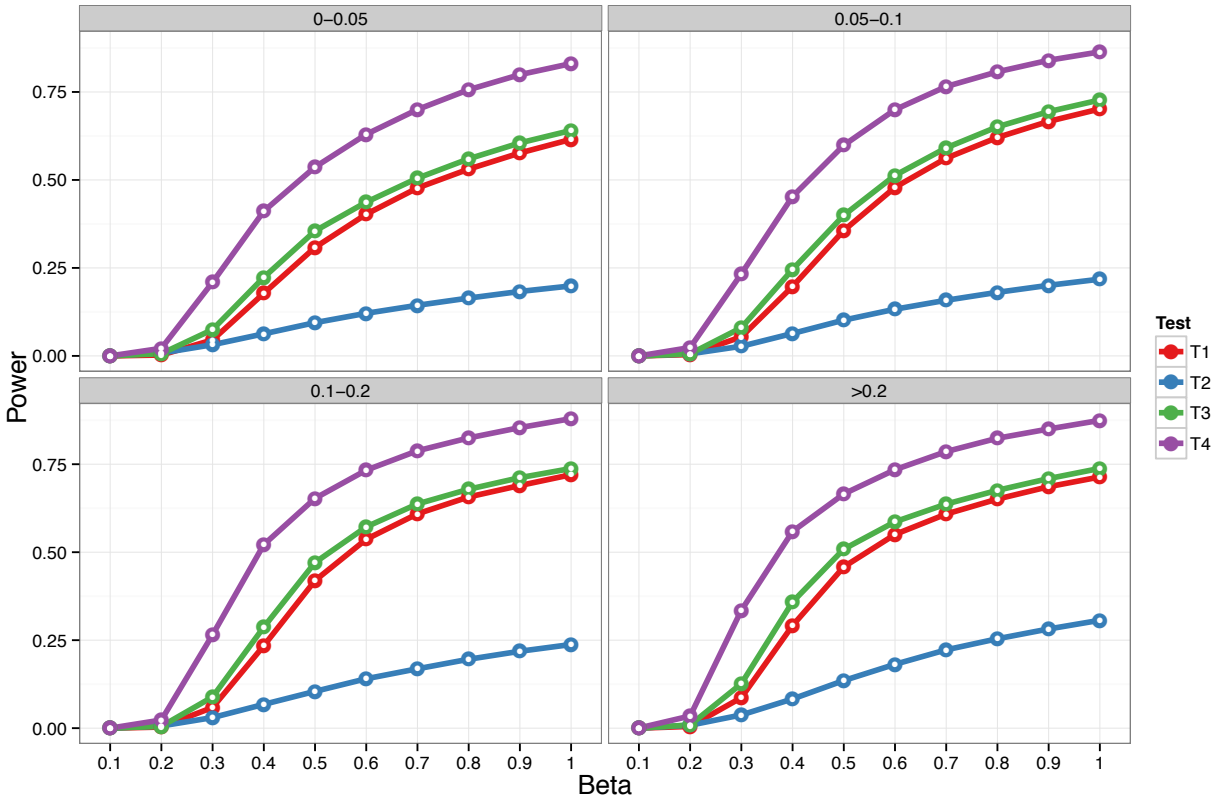


Figure 5.6. Power comparison among four tests with simulated true and inferred ancestry under Scenario 4.

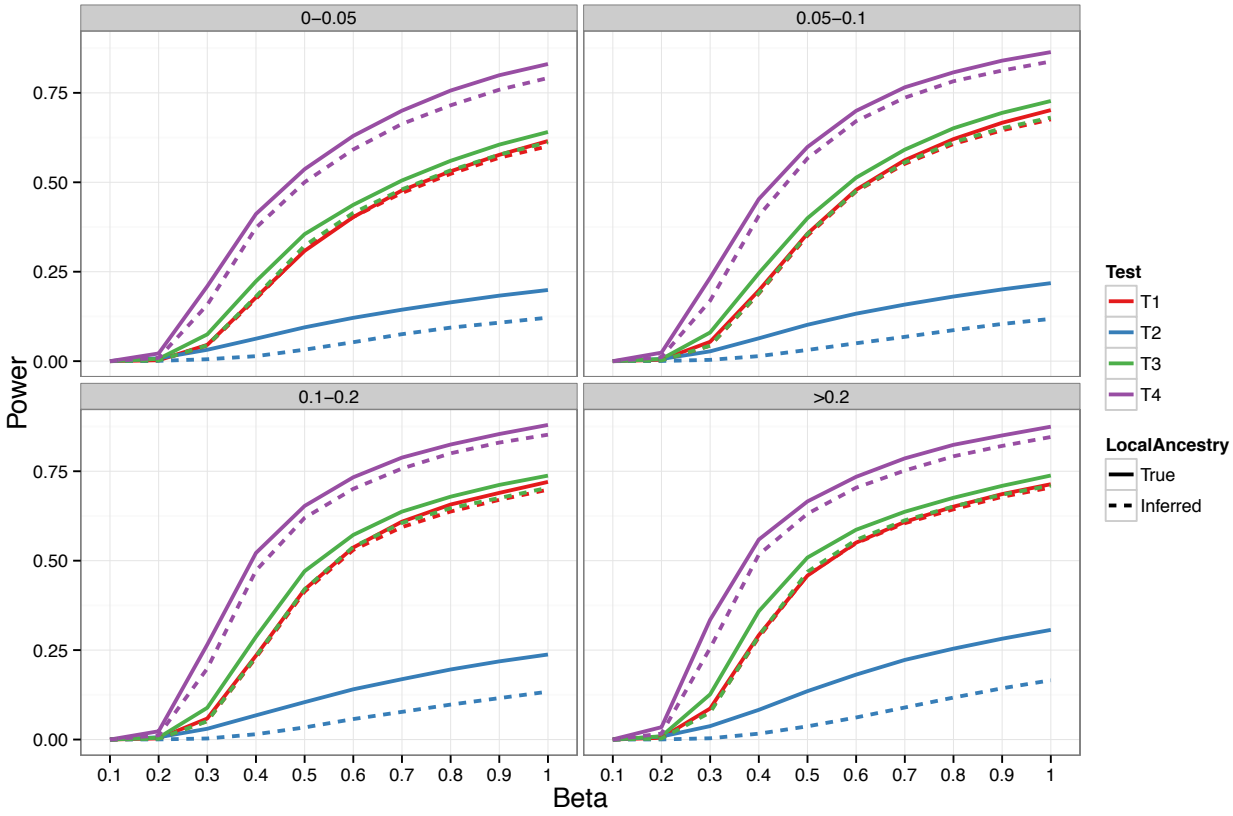


Table 5.1. Four simulation settings

Scenario	Ancestry effect (γ)	Allele effect (β)	Effect direction (AFR/EUR)	Allele frequency difference (δ)
1	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1	0	NA/NA	0-0.05, 0.05-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4 and >0.4
2	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1	$\gamma \times k$ [1]	+ / NA	0-0.05, 0.05-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4 and >0.4
3	0	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1	+ / +	0-0.05, 0.05-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4 and >0.4
4	0	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1	+ / -	0-0.05, 0.05-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4 and >0.4

[1] $k = 0.1$, when $0 < \delta < 0.05$; $k = 0.2$, when $0.05 < \delta < 0.1$; $k = 0.3$, when $0.1 < \delta < 0.2$; $k = 0.4$, when $0.2 < \delta < 0.3$; $k = 0.5$, when $0.3 < \delta < 0.4$; and $k = 0.6$, when $\delta > 0.4$.

Table 5.2. Average proportion of times that the source of association is correctly identified.

	Allele frequency difference			
	0-0.05	**0.05-0.1**	**0.1-0.2**	**>0.2**
Scenario 1	98.7%	99.6%	99.6%	99.1%
Scenario 2	99.5%	99.3%	98.1%	90.8%
Scenario 3	88.1%	87.8%	87.9%	89.3%
Scenario 4	93.5%	93.0%	93.2%	94.0%

Table 5.3. Ratio of the statistical power of T4 to that of T5 in Scenario 4 using two-step testing procedure

STEP 1				
	Allele frequency difference			
Beta	**0-0.05**	**0.05-0.1**	**0.1-0.2**	**>0.2**
0.1	NA	NA	NA	10.29
0.2	2.53	2.63	2.82	2.46
0.3	1.75	1.80	1.89	1.69
0.4	1.31	1.30	1.27	1.19
0.5	1.19	1.16	1.12	1.11
0.6	1.14	1.11	1.09	1.09
0.7	1.11	1.08	1.07	1.07
0.8	1.08	1.06	1.05	1.06
0.9	1.07	1.05	1.04	1.04
1.0	1.06	1.04	1.04	1.04
STEP 2				
Average	1.20	1.22	1.21	1.18

Table 5.4. Ratio of type I error to the nominal significance level of 0.05, 0.01 and 0.001.

Allele Frequency difference	0.05				0.01				0.001			
	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4
0-0.05	1.01	0.99	1.00	1.00	1.04	0.99	1.01	1.00	0.99	1.00	1.01	1.00
0.05-0.1	1.00	1.00	0.99	1.00	0.96	1.02	1.00	1.01	0.88	0.97	0.99	0.89
0.1-0.2	0.99	1.00	0.99	1.00	0.99	1.01	1.00	1.02	0.97	0.98	0.83	0.84
0.2-0.3	0.99	1.00	1.01	1.00	0.99	1.00	1.00	0.97	0.89	0.93	0.88	0.82
0.3-0.4	1.00	1.00	1.00	1.01	1.03	1.00	1.01	1.01	0.98	0.98	0.86	0.95
>0.4	1.01	0.99	1.01	1.01	1.01	0.96	0.96	0.98	0.82	0.92	0.85	0.87

Table 5.5. Median Pearson correlation coefficient between true and inferred local ancestry

Correlation	Genotype	African allele	African ref allele
All markers	1.00	0.81	0.97
Markers with overall MAF < 0.05	1.00	0.81	0.94
Markers with AFR MAF < 0.05	1.00	0.82	0.88

Table 5.6. Ratio of the statistical power of T4 to that of T5 in Scenario 4 using two-step testing procedure and estimated ancestry

STEP 1				
	Allele frequency difference			
Beta	**0-0.05**	**0.05-0.1**	**0.1-0.2**	**>0.2**
0.1	NA	NA	NA	NA
0.2	3.00	4.15	3.50	3.50
0.3	2.09	2.14	2.18	1.94
0.4	1.38	1.41	1.37	1.27
0.5	1.19	1.22	1.17	1.13
0.6	1.15	1.14	1.12	1.10
0.7	1.12	1.11	1.10	1.07
0.8	1.09	1.09	1.09	1.07
0.9	1.07	1.08	1.08	1.06
1.0	1.06	1.07	1.06	1.05
STEP 2				
Average	1.16	1.15	1.16	1.15

Table 5.7. Replication of published loci that are associated with hemoglobin using two-step testing procedure

SNP	Chr	Pos[b]	*P*-value	Allele effect	Ancestry effect	African ancestry allele effect	N	Ref.
rs9940149	16	300641	7.23×10^{-7}	**1.87**	0.62	-1.24	5711	(Li, et al., 2013)
rs7199221	16	3101639	1.74×10^{-7}	**-3.89**	-0.73	2.59	5711	(Lo, et al., 2011)
rs1211375	16	240280	7.32×10^{-9}	0.86	-0.43	**-1.53**	5711	(Lo, et al., 2011)

CHAPTER 6: CONCLUDING REMARKS

In this document, I present computational resource and statistical approaches that facilitate genetic association studies with complex human traits in admixed populations. Three studies are performed that enhance genotype imputation quality with a focus on rare variants and account for local ancestry information in association analysis. In the first study (Chapter 3), I built a database containing the marker imputability information using reference panels from the public domain for four continental groups, respectively. This study facilitates genetic association studies in diverse populations including admixed populations from the study design stage to the post-imputation stage where marker imputation information is desirable. In the second study (Chapter 4), I focus particularly on enhancing imputation accuracy in African American samples. I show as a proof of principle that imputation accuracy, particular that of rare variants, can be enhanced by using an internal reference panel with similar ancestral makeup. I also provide the optimal approach to combine internal and external reference panels for better imputation. In the third study (Chapter 5), I propose a two-step testing procedure for genetic association studies in African Americans. The first step of the testing procedure could capture genetic associations due to only an allele effect, only an ancestry effect and anywhere in between. In particularly, it is powerful when effect heterogeneity exists between the ancestral populations. Through a second step, the source of association, i.e., allele effect, ancestry effect or ancestry-specific allele effect can be identified. Taking advantage of the joint distribution of allele and local ancestry, we use a novel parameterization of the interaction term, which is found to be more powerful to capture the

presence of effect heterogeneity than the traditional cross-product interaction term. Further studies are needed to extend the testing procedure developed in a two-way admixed population to three-way or multi-way admixed populations. In addition, the current study focuses on common variants. Rare variants association methods tailored for admixed populations are needed.

In the U.S., admixed populations have been drawing increasing attention and genetic and phenotypic data are growing for African American and Hispanic populations. Advanced statistical and computational methods addressing many challenging issues pave the way for genetic association studies in admixed populations. In general, the success of GWAS in Caucasians has advanced our understanding of disease etiology. In diverse populations, GWAS will provide a more complete understanding of the genetic basis of complex traits, thereby, potentially reducing health disparities due to the bias towards GWAS in European populations. The knowledge gained from GWAS, including the discovery of risk variants and their characterization, will eventually allow all people to benefit from the improvement in more precise disease prevention, clinical diagnostics and medical treatment.

REFERENCES

Adair, L. S., Popkin, B. M., Akin, J. S., Guilkey, D. K., Gultiano, S., Borja, J., Perez, L., Kuzawa, C. W., McDade, T., & Hindin, M. J. (2011). Cohort profile: the Cebu longitudinal health and nutrition survey. *Int J Epidemiol, 40*(3), 619-625. doi: 10.1093/ije/dyq085

Auer, P. L., Johnsen, J. M., Johnson, A. D., Logsdon, B. A., Lange, L. A., Nalls, M. A., Zhang, G., Franceschini, N., Fox, K., Lange, E. M., Rich, S. S., O'Donnell, C. J., Jackson, R. D., Wallace, R. B., Chen, Z., Graubert, T. A., Wilson, J. G., Tang, H., Lettre, G., Reiner, A. P., Ganesh, S. K., & Li, Y. (2012). Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. *Am J Hum Genet, 91*(5), 794-808. doi: 10.1016/j.ajhg.2012.08.031

Auer, P. L., Teumer, A., Schick, U., O'Shaughnessy, A., Lo, K. S., Chami, N., Carlson, C., de Denus, S., Dube, M. P., Haessler, J., Jackson, R. D., Kooperberg, C., Perreault, L. P., Nauck, M., Peters, U., Rioux, J. D., Schmidt, F., Turcot, V., Volker, U., Volzke, H., Greinacher, A., Hsu, L., Tardif, J. C., Diaz, G. A., Reiner, A. P., & Lettre, G. (2014). Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. *Nat Genet, 46*(6), 629-634. doi: 10.1038/ng.2962

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J. G., Avila, P. C., Rodriguez-Santana, J., Burchard, E. G., & Halperin, E. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics, 28*(10), 1359-1367. doi: 10.1093/bioinformatics/bts144

Barnett, I. J., Lee, S., & Lin, X. (2013). Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genet Epidemiol, 37*(2), 142-151. doi: 10.1002/gepi.21699

Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Diez Roux, A. V., Folsom, A. R., Greenland, P., Jacob, D. R., Jr., Kronmal, R., Liu, K., Nelson, J. C., O'Leary, D., Saad, M. F., Shea, S., Szklo, M., & Tracy, R. P. (2002). Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am J Epidemiol, 156*(9), 871-881.

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J. G., & Bustamante, C. D. (2012). PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum Biol, 84*(4), 343-364. doi: 10.3378/027.084.0401

Browning, B. L., & Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet, 85*(6), 847-861. doi: 10.1016/j.ajhg.2009.11.004

Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., Froment, A., Bodo, J. M., Wambebe, C., Tishkoff, S. A., & Bustamante, C. D. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A, 107*(2), 786-791. doi: 10.1073/pnas.0909559107

Bunker, C. H., Patrick, A. L., Konety, B. R., Dhir, R., Brufsky, A. M., Vivas, C. A., Becich, M. J., Trump, D. L., & Kuller, L. H. (2002). High prevalence of screening-detected prostate cancer among Afro-Caribbeans: the Tobago Prostate Cancer Survey. *Cancer Epidemiol Biomarkers Prev, 11*(8), 726-729.

Buyske, S., Wu, Y., Carty, C. L., Cheng, I., Assimes, T. L., Dumitrescu, L., Hindorff, L. A., Mitchell, S., Ambite, J. L., Boerwinkle, E., Buzkova, P., Carlson, C. S., Cochran, B., Duggan, D., Eaton, C. B., Fesinmeyer, M. D., Franceschini, N., Haessler, J., Jenny, N., Kang, H. M., Kooperberg, C., Lin, Y., Le Marchand, L., Matise, T. C., Robinson, J. G., Rodriguez, C., Schumacher, F. R., Voight, B. F., Young, A., Manolio, T. A., Mohlke, K. L., Haiman, C. A., Peters, U., Crawford, D. C., & North, K. E. (2012). Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: the PAGE study. *PLoS One, 7*(4), e35651. doi: 10.1371/journal.pone.0035651

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., & Nickerson, D. A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet, 74*(1), 106-120. doi: 10.1086/381000

Chakraborty, R., & Weiss, K. M. (1988). Admixture as a Tool for Finding Linked Genes and Detecting That Difference from Allelic Association between Loci. *Proc Natl Acad Sci U S A, 85*(23), 9119-9123. doi: Doi 10.1073/Pnas.85.23.9119

Chen, G. K., Millikan, R. C., John, E. M., Ambrosone, C. B., Bernstein, L., Zheng, W., Hu, J. J., Chanock, S. J., Ziegler, R. G., Bandera, E. V., Henderson, B. E., Haiman, C. A., & Stram, D. O. (2010). The potential for enhancing the power of genetic association studies in African Americans through the reuse of existing genotype data. *PLoS Genet, 6*(9), e1001096. doi: 10.1371/journal.pgen.1001096

Chen, R., Corona, E., Sikora, M., Dudley, J. T., Morgan, A. A., Moreno-Estrada, A., Nilsen, G. B., Ruau, D., Lincoln, S. E., Bustamante, C. D., & Butte, A. J. (2012). Type 2 diabetes

risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases. *PLoS Genet, 8*(4), e1002621. doi: 10.1371/journal.pgen.1002621

Chen, Z., Tang, H., Qayyum, R., Schick, U. M., Nalls, M. A., Handsaker, R., Li, J., Lu, Y., Yanek, L. R., Keating, B., Meng, Y., van Rooij, F. J., Okada, Y., Kubo, M., Rasmussen-Torvik, L., Keller, M. F., Lange, L., Evans, M., Bottinger, E. P., Linderman, M. D., Ruderfer, D. M., Hakonarson, H., Papanicolaou, G., Zonderman, A. B., Gottesman, O., BioBank Japan, Project, Consortium, Charge, Thomson, C., Ziv, E., Singleton, A. B., Loos, R. J., Sleiman, P. M., Ganesh, S., McCarroll, S., Becker, D. M., Wilson, J. G., Lettre, G., & Reiner, A. P. (2013). Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network. *Hum Mol Genet, 22*(12), 2529-2538. doi: 10.1093/hmg/ddt087

Cheng, C. Y., Kao, W. H., Patterson, N., Tandon, A., Haiman, C. A., Harris, T. B., Xing, C., John, E. M., Ambrosone, C. B., Brancati, F. L., Coresh, J., Press, M. F., Parekh, R. S., Klag, M. J., Meoni, L. A., Hsueh, W. C., Fejerman, L., Pawlikowska, L., Freedman, M. L., Jandorf, L. H., Bandera, E. V., Ciupak, G. L., Nalls, M. A., Akylbekova, E. L., Orwoll, E. S., Leak, T. S., Miljkovic, I., Li, R., Ursin, G., Bernstein, L., Ardlie, K., Taylor, H. A., Boerwinckle, E., Zmuda, J. M., Henderson, B. E., Wilson, J. G., & Reich, D. (2009). Admixture mapping of 15,280 African Americans identifies obesity susceptibility loci on chromosomes 5 and X. *PLoS Genet, 5*(5), e1000490. doi: 10.1371/journal.pgen.1000490

Cho, Y. S., Chen, C. H., Hu, C., Long, J., Ong, R. T., Sim, X., Takeuchi, F., Wu, Y., Go, M. J., Yamauchi, T., Chang, Y. C., Kwak, S. H., Ma, R. C., Yamamoto, K., Adair, L. S., Aung, T., Cai, Q., Chang, L. C., Chen, Y. T., Gao, Y., Hu, F. B., Kim, H. L., Kim, S., Kim, Y. J., Lee, J. J., Lee, N. R., Li, Y., Liu, J. J., Lu, W., Nakamura, J., Nakashima, E., Ng, D. P., Tay, W. T., Tsai, F. J., Wong, T. Y., Yokota, M., Zheng, W., Zhang, R., Wang, C., So, W. Y., Ohnaka, K., Ikegami, H., Hara, K., Cho, Y. M., Cho, N. H., Chang, T. J., Bao, Y., Hedman, A. K., Morris, A. P., McCarthy, M. I., Consortium, Diagram, Mu, Ther Consortium, Takayanagi, R., Park, K. S., Jia, W., Chuang, L. M., Chan, J. C., Maeda, S., Kadowaki, T., Lee, J. Y., Wu, J. Y., Teo, Y. Y., Tai, E. S., Shu, X. O., Mohlke, K. L., Kato, N., Han, B. G., & Seielstad, M. (2012). Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat Genet, 44*(1), 67-72. doi: 10.1038/ng.1019

Coram, M. A., Duan, Q., Hoffmann, T. J., Thornton, T., Knowles, J. W., Johnson, N. A., Ochs-Balcom, H. M., Donlon, T. A., Martin, L. W., Eaton, C. B., Robinson, J. G., Risch, N. J., Zhu, X., Kooperberg, C., Li, Y., Reiner, A. P., & Tang, H. (2013). Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. *Am J Hum Genet, 92*(6), 904-916. doi: 10.1016/j.ajhg.2013.04.025

Croteau-Chonka, D. C., Wu, Y., Li, Y., Fogarty, M. P., Lange, L. A., Kuzawa, C. W., McDade, T. W., Borja, J. B., Luo, J., AbdelBaky, O., Combs, T. P., Adair, L. S., Lange, E. M., & Mohlke, K. L. (2012). Population-specific coding variant underlies genome-wide association with adiponectin level. *Hum Mol Genet, 21*(2), 463-471. doi: 10.1093/hmg/ddr480

Cunnington, M. S., Santibanez Koref, M., Mayosi, B. M., Burn, J., & Keavney, B. (2010). Chromosome 9p21 SNPs Associated with Multiple Disease Phenotypes Correlate with ANRIL Expression. *PLoS Genet, 6*(4), e1000899. doi: 10.1371/journal.pgen.1000899

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., & Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat Genet, 29*(2), 229-232. doi: 10.1038/ng1001-229

Dastani, Z., Hivert, M. F., Timpson, N., Perry, J. R., Yuan, X., Scott, R. A., Henneman, P., Heid, I. M., Kizer, J. R., Lyytikainen, L. P., Fuchsberger, C., Tanaka, T., Morris, A. P., Small, K., Isaacs, A., Beekman, M., Coassin, S., Lohman, K., Qi, L., Kanoni, S., Pankow, J. S., Uh, H. W., Wu, Y., Bidulescu, A., Rasmussen-Torvik, L. J., Greenwood, C. M., Ladouceur, M., Grimsby, J., Manning, A. K., Liu, C. T., Kooner, J., Mooser, V. E., Vollenweider, P., Kapur, K. A., Chambers, J., Wareham, N. J., Langenberg, C., Frants, R., Willems-Vandijk, K., Oostra, B. A., Willems, S. M., Lamina, C., Winkler, T. W., Psaty, B. M., Tracy, R. P., Brody, J., Chen, I., Viikari, J., Kahonen, M., Pramstaller, P. P., Evans, D. M., St Pourcain, B., Sattar, N., Wood, A. R., Bandinelli, S., Carlson, O. D., Egan, J. M., Bohringer, S., van Heemst, D., Kedenko, L., Kristiansson, K., Nuotio, M. L., Loo, B. M., Harris, T., Garcia, M., Kanaya, A., Haun, M., Klopp, N., Wichmann, H. E., Deloukas, P., Katsareli, E., Couper, D. J., Duncan, B. B., Kloppenburg, M., Adair, L. S., Borja, J. B., Consortium, Diagram, Consortium, Magic, Investigators, Glgc, Mu, Ther Consortium, Wilson, J. G., Musani, S., Guo, X., Johnson, T., Semple, R., Teslovich, T. M., Allison, M. A., Redline, S., Buxbaum, S. G., Mohlke, K. L., Meulenbelt, I., Ballantyne, C. M., Dedoussis, G. V., Hu, F. B., Liu, Y., Paulweber, B., Spector, T. D., Slagboom, P. E., Ferrucci, L., Jula, A., Perola, M., Raitakari, O., Florez, J. C., Salomaa, V., Eriksson, J. G., Frayling, T. M., Hicks, A. A., Lehtimaki, T., Smith, G. D., Siscovick, D. S., Kronenberg, F., van Duijn, C., Loos, R. J., Waterworth, D. M., Meigs, J. B., Dupuis, J., Richards, J. B., Voight, B. F., Scott, L. J., Steinthorsdottir, V., Dina, C., Welch, R. P., Zeggini, E., Huth, C., Aulchenko, Y. S., Thorleifsson, G., McCulloch, L. J., Ferreira, T., Grallert, H., Amin, N., Wu, G., Willer, C. J., Raychaudhuri, S., McCarroll, S. A., Hofmann, O. M., Segre, A. V., van Hoek, M., Navarro, P., Ardlie, K., Balkau, B., Benediktsson, R., Bennett, A. J., Blagieva, R., Boerwinkle, E., Bonnycastle, L. L., Bostrom, K. B., Bravenboer, B., Bumpstead, S., Burtt, N. P., Charpentier, G., Chines, P. S., Cornelis, M., Crawford, G., Doney, A. S., Elliott, K. S., Elliott, A. L., Erdos, M. R., Fox, C. S., Franklin, C. S., Ganser, M., Gieger, C., Grarup, N., Green, T., Griffin, S., Groves, C. J., Guiducci, C., Hadjadj, S., Hassanali, N., Herder, C., Isomaa, B., Jackson, A. U., Johnson, P. R., Jorgensen, T., Kao, W. H., Kong, A., Kraft, P., Kuusisto, J., Lauritzen, T., Li, M., Lieverse, A., Lindgren, C. M., Lyssenko, V., Marre, M., Meitinger,

T., Midthjell, K., Morken, M. A., Narisu, N., Nilsson, P., Owen, K. R., Payne, F., Petersen, A. K., Platou, C., Proenca, C., Prokopenko, I., Rathmann, W., Rayner, N. W., Robertson, N. R., Rocheleau, G., Roden, M., Sampson, M. J., Saxena, R., Shields, B. M., Shrader, P., Sigurdsson, G., Sparso, T., Strassburger, K., Stringham, H. M., Sun, Q., Swift, A. J., Thorand, B., Tichet, J., Tuomi, T., van Dam, R. M., van Haeften, T. W., van Herpt, T., van Vliet-Ostaptchouk, J. V., Walters, G. B., Weedon, M. N., Wijmenga, C., Witteman, J., Bergman, R. N., Cauchi, S., Collins, F. S., Gloyn, A. L., Gyllensten, U., Hansen, T., Hide, W. A., Hitman, G. A., Hofman, A., Hunter, D. J., Hveem, K., Laakso, M., Morris, A. D., Palmer, C. N., Rudan, I., Sijbrands, E., Stein, L. D., Tuomilehto, J., Uitterlinden, A., Walker, M., Watanabe, R. M., Abecasis, G. R., Boehm, B. O., Campbell, H., Daly, M. J., Hattersley, A. T., Pedersen, O., Barroso, I., Groop, L., Sladek, R., Thorsteinsdottir, U., Wilson, J. F., Illig, T., Froguel, P., van Duijn, C. M., Stefansson, K., Altshuler, D., Boehnke, M., McCarthy, M. I., Soranzo, N., Wheeler, E., Glazer, N. L., Bouatia-Naji, N., Magi, R., Randall, J., Elliott, P., Rybin, D., Dehghan, A., Hottenga, J. J., Song, K., Goel, A., Lajunen, T., Doney, A., Cavalcanti-Proenca, C., Kumari, M., Timpson, N. J., Zabena, C., Ingelsson, E., An, P., O'Connell, J., Luan, J., Elliott, A., McCarroll, S. A., Roccasecca, R. M., Pattou, F., Sethupathy, P., Ariyurek, Y., Barter, P., Beilby, J. P., Ben-Shlomo, Y., Bergmann, S., Bochud, M., Bonnefond, A., Borch-Johnsen, K., Bottcher, Y., Brunner, E., Bumpstead, S. J., Chen, Y. D., Chines, P., Clarke, R., Coin, L. J., Cooper, M. N., Crisponi, L., Day, I. N., de Geus, E. J., Delplanque, J., Fedson, A. C., Fischer-Rosinsky, A., Forouhi, N. G., Franzosi, M. G., Galan, P., Goodarzi, M. O., Graessler, J., Grundy, S., Gwilliam, R., Hallmans, G., Hammond, N., Han, X., Hartikainen, A. L., Hayward, C., Heath, S. C., Hercberg, S., Hillman, D. R., Hingorani, A. D., Hui, J., Hung, J., Kaakinen, M., Kaprio, J., Kesaniemi, Y. A., Kivimaki, M., Knight, B., Koskinen, S., Kovacs, P., Kyvik, K. O., Lathrop, G. M., Lawlor, D. A., Le Bacquer, O., Lecoeur, C., Li, Y., Mahley, R., Mangino, M., Martinez-Larrad, M. T., McAteer, J. B., McPherson, R., Meisinger, C., Melzer, D., Meyre, D., Mitchell, B. D., Mukherjee, S., Naitza, S., Neville, M. J., Orru, M., Pakyz, R., Paolisso, G., Pattaro, C., Pearson, D., Peden, J. F., Pedersen, N. L., Pfeiffer, A. F., Pichler, I., Polasek, O., Posthuma, D., Potter, S. C., Pouta, A., Province, M. A., Rayner, N. W., Rice, K., Ripatti, S., Rivadeneira, F., Rolandsson, O., Sandbaek, A., Sandhu, M., Sanna, S., Sayer, A. A., Scheet, P., Seedorf, U., Sharp, S. J., Shields, B., Sigurethsson, G., Sijbrands, E. J., Silveira, A., Simpson, L., Singleton, A., Smith, N. L., Sovio, U., Swift, A., Syddall, H., Syvanen, A. C., Tonjes, A., Uitterlinden, A. G., van Dijk, K. W., Varma, D., Visvikis-Siest, S., Vitart, V., Vogelzangs, N., Waeber, G., Wagner, P. J., Walley, A., Ward, K. L., Watkins, H., Wild, S. H., Willemsen, G., Witteman, J. C., Yarnell, J. W., Zelenika, D., Zethelius, B., Zhai, G., Zhao, J. H., Zillikens, M. C., Consortium, Diagram, Consortium, Giant, Global, B. Pgen Consortium, Borecki, I. B., Meneton, P., Magnusson, P. K., Nathan, D. M., Williams, G. H., Silander, K., Bornstein, S. R., Schwarz, P., Spranger, J., Karpe, F., Shuldiner, A. R., Cooper, C., Serrano-Rios, M., Lind, L., Palmer, L. J., Hu, F. B. st, Franks, P. W., Ebrahim, S., Marmot, M., Kao, W. H., Pramstaller, P. P., Wright, A. F., Stumvoll, M., Hamsten, A., Procardis, Consortium, Buchanan, T. A., Valle, T. T., Rotter, J. I., Penninx, B. W., Boomsma, D. I., Cao, A., Scuteri, A., Schlessinger, D., Uda, M., Ruokonen, A., Jarvelin, M. R., Peltonen, L., Mooser, V., Sladek, R., investigators, Magic, Consortium, Glgc, Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Chasman, D. I.,

Johansen, C. T., Fouchier, S. W., Peloso, G. M., Barbalic, M., Ricketts, S. L., Bis, J. C., Feitosa, M. F., Orho-Melander, M., Melander, O., Li, X., Li, M., Cho, Y. S., Go, M. J., Kim, Y. J., Lee, J. Y., Park, T., Kim, K., Sim, X., Ong, R. T., Croteau-Chonka, D. C., Lange, L. A., Smith, J. D., Ziegler, A., Zhang, W., Zee, R. Y., Whitfield, J. B., Thompson, J. R., Surakka, I., Spector, T. D., Smit, J. H., Sinisalo, J., Scott, J., Saharinen, J., Sabatti, C., Rose, L. M., Roberts, R., Rieder, M., Parker, A. N., Pare, G., O'Donnell, C. J., Nieminen, M. S., Nickerson, D. A., Montgomery, G. W., McArdle, W., Masson, D., Martin, N. G., Marroni, F., Lucas, G., Luben, R., Lokki, M. L., Lettre, G., Launer, L. J., Lakatta, E. G., Laaksonen, R., Kyvik, K. O., Konig, I. R., Khaw, K. T., Kaplan, L. M., Johansson, A., Janssens, A. C., Igl, W., Hovingh, G. K., Hengstenberg, C., Havulinna, A. S., Hastie, N. D., Harris, T. B., Haritunians, T., Hall, A. S., Groop, L. C., Gonzalez, E., Freimer, N. B., Erdmann, J., Ejebe, K. G., Doring, A., Dominiczak, A. F., Demissie, S., Deloukas, P., de Faire, U., Crawford, G., Chen, Y. D., Caulfield, M. J., Boekholdt, S. M., Assimes, T. L., Quertermous, T., Seielstad, M., Wong, T. Y., Tai, E. S., Feranil, A. B., Kuzawa, C. W., Taylor, H. A., Jr., Gabriel, S. B., Holm, H., Gudnason, V., Krauss, R. M., Ordovas, J. M., Munroe, P. B., Kooner, J. S., Tall, A. R., Hegele, R. A., Kastelein, J. J., Schadt, E. E., Strachan, D. P., Reilly, M. P., Samani, N. J., Schunkert, H., Cupples, L. A., Sandhu, M. S., Ridker, P. M., Rader, D. J., & Kathiresan, S. (2012). Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet, 8*(3), e1002607. doi: 10.1371/journal.pgen.1002607

Day-Williams, A. G., Southam, L., Panoutsopoulou, K., Rayner, N. W., Esko, T., Estrada, K., Helgadottir, H. T., Hofman, A., Ingvarsson, T., Jonsson, H., Keis, A., Kerkhof, H. J., Thorleifsson, G., Arden, N. K., Carr, A., Chapman, K., Deloukas, P., Loughlin, J., McCaskie, A., Ollier, W. E., Ralston, S. H., Spector, T. D., Wallis, G. A., Wilkinson, J. M., Aslam, N., Birell, F., Carluke, I., Joseph, J., Rai, A., Reed, M., Walker, K., arc, Ogen Consortium, Doherty, S. A., Jonsdottir, I., Maciewicz, R. A., Muir, K. R., Metspalu, A., Rivadeneira, F., Stefansson, K., Styrkarsdottir, U., Uitterlinden, A. G., van Meurs, J. B., Zhang, W., Valdes, A. M., Doherty, M., & Zeggini, E. (2011). A variant in MCF2L is associated with osteoarthritis. *Am J Hum Genet, 89*(3), 446-450. doi: 10.1016/j.ajhg.2011.08.001

de Bakker, P. I., Ferreira, M. A., Jia, X., Neale, B. M., Raychaudhuri, S., & Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet, 17*(R2), R122-128. doi: 10.1093/hmg/ddn288

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics, 55*(4), 997-1004.

Dhandapany, P. S., Sadayappan, S., Xue, Y., Powell, G. T., Rani, D. S., Nallari, P., Rai, T. S., Khullar, M., Soares, P., Bahl, A., Tharkan, J. M., Vaideeswar, P., Rathinavel, A., Narasimhan, C., Ayapati, D. R., Ayub, Q., Mehdi, S. Q., Oppenheimer, S., Richards, M. B., Price, A. L., Patterson, N., Reich, D., Singh, L., Tyler-Smith, C., & Thangaraj, K.

(2009). A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia. *Nat Genet, 41*(2), 187-191. doi: 10.1038/ng.309

Elbein, S. C., Das, S. K., Hallman, D. M., Hanis, C. L., & Hasstedt, S. J. (2009). Genome-wide linkage and admixture mapping of type 2 diabetes in African American families from the American Diabetes Association GENNID (Genetics of NIDDM) Study Cohort. *Diabetes, 58*(1), 268-274. doi: 10.2337/db08-0931

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics, 164*(4), 1567-1587.

Franceschini, N., Fox, E., Zhang, Z., Edwards, T. L., Nalls, M. A., Sung, Y. J., Tayo, B. O., Sun, Y. V., Gottesman, O., Adeyemo, A., Johnson, A. D., Young, J. H., Rice, K., Duan, Q., Chen, F., Li, Y., Tang, H., Fornage, M., Keene, K. L., Andrews, J. S., Smith, J. A., Faul, J. D., Guangfa, Z., Guo, W., Liu, Y., Murray, S. S., Musani, S. K., Srinivasan, S., Velez Edwards, D. R., Wang, H., Becker, L. C., Bovet, P., Bochud, M., Broeckel, U., Burnier, M., Carty, C., Chasman, D. I., Ehret, G., Chen, W. M., Chen, G., Chen, W., Ding, J., Dreisbach, A. W., Evans, M. K., Guo, X., Garcia, M. E., Jensen, R., Keller, M. F., Lettre, G., Lotay, V., Martin, L. W., Moore, J. H., Morrison, A. C., Mosley, T. H., Ogunniyi, A., Palmas, W., Papanicolaou, G., Penman, A., Polak, J. F., Ridker, P. M., Salako, B., Singleton, A. B., Shriner, D., Taylor, K. D., Vasan, R., Wiggins, K., Williams, S. M., Yanek, L. R., Zhao, W., Zonderman, A. B., Becker, D. M., Berenson, G., Boerwinkle, E., Bottinger, E., Cushman, M., Eaton, C., Nyberg, F., Heiss, G., Hirschhron, J. N., Howard, V. J., Karczewsk, K. J., Lanktree, M. B., Liu, K., Liu, Y., Loos, R., Margolis, K., Snyder, M., Asian Genetic Epidemiology Network, Consortium, Psaty, B. M., Schork, N. J., Weir, D. R., Rotimi, C. N., Sale, M. M., Harris, T., Kardia, S. L., Hunt, S. C., Arnett, D., Redline, S., Cooper, R. S., Risch, N. J., Rao, D. C., Rotter, J. I., Chakravarti, A., Reiner, A. P., Levy, D., Keating, B. J., & Zhu, X. (2013). Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations. *Am J Hum Genet, 93*(3), 545-554. doi: 10.1016/j.ajhg.2013.07.010

Frazier-Wood, A. C., Manichaikul, A., Aslibekyan, S., Borecki, I. B., Goff, D. C., Hopkins, P. N., Lai, C. Q., Ordovas, J. M., Post, W. S., Rich, S. S., Sale, M. M., Siscovick, D., Straka, R. J., Tiwari, H. K., Tsai, M. Y., Rotter, J. I., & Arnett, D. K. (2013). Genetic variants associated with VLDL, LDL and HDL particle size differ with race/ethnicity. *Hum Genet, 132*(4), 405-413. doi: 10.1007/s00439-012-1256-1

Freedman, M. L., Haiman, C. A., Patterson, N., McDonald, G. J., Tandon, A., Waliszewska, A., Penney, K., Steen, R. G., Ardlie, K., John, E. M., Oakley-Girvan, I., Whittemore, A. S., Cooney, K. A., Ingles, S. A., Altshuler, D., Henderson, B. E., & Reich, D. (2006).

Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc Natl Acad Sci U S A, 103*(38), 14068-14073. doi: 10.1073/pnas.0605832103

Fridley, B. L., Jenkins, G., Deyo-Svendsen, M. E., Hebbring, S., & Freimuth, R. (2010). Utilizing genotype imputation for the augmentation of sequence data. *PLoS One, 5*(6), e11018. doi: 10.1371/journal.pone.0011018

Friedman, G. D., Cutter, G. R., Donahue, R. P., Hughes, G. H., Hulley, S. B., Jacobs, D. R., Jr., Liu, K., & Savage, P. J. (1988). CARDIA: study design, recruitment, and some characteristics of the examined subjects. *J Clin Epidemiol, 41*(11), 1105-1116.

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Rieder, M. J., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., Project, Nhlbi Exome Sequencing, & Akey, J. M. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature, 493*(7431), 216-220. doi: 10.1038/nature11690

Fuchsberger, C., Howie, B., Laakso, M., Boehnke, M., Abecasis, G., & the Genetics of Type-2 Diabetes (Go-T2D), Consortium (2012). The value of population-specific reference panels for genotype imputation in the age of whole-genome sequencing. *Presented at the 62nd Annual Meeting of The American Society of Human Genetics. San Francisco, CA*.

Futema, M., Plagnol, V., Whittall, R. A., Neil, H. A., Simon Broome Register, Group, Humphries, S. E., & Uk10K. (2012). Use of targeted exome sequencing as a diagnostic tool for Familial Hypercholesterolaemia. *J Med Genet, 49*(10), 644-649. doi: 10.1136/jmedgenet-2012-101189

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., & Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science, 296*(5576), 2225-2229. doi: 10.1126/science.1069424

Genovese, G., Friedman, D. J., Ross, M. D., Lecordier, L., Uzureau, P., Freedman, B. I., Bowden, D. W., Langefeld, C. D., Oleksyk, T. K., Uscinski Knob, A. L., Bernhardy, A. J., Hicks, P. J., Nelson, G. W., Vanhollebeke, B., Winkler, C. A., Kopp, J. B., Pays, E., & Pollak, M. R. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science, 329*(5993), 841-845. doi: 10.1126/science.1193032

Gignoux, C. R. (2015). A Multi-Ethnic Genotyping Array for the Next Generation of Association Studies; (Program #1885W). *Presented at the Annual Meeting of The American Society of Human Genetics, Oct 7, Baltimore, MD*.

Guey, L. T., Kravic, J., Melander, O., Burtt, N. P., Laramie, J. M., Lyssenko, V., Jonsson, A., Lindholm, E., Tuomi, T., Isomaa, B., Nilsson, P., Almgren, P., Kathiresan, S., Groop, L., Seymour, A. B., Altshuler, D., & Voight, B. F. (2011). Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. *Genet Epidemiol, 35*(4), 236-246. doi: 10.1002/gepi.20572

Helgason, A., Palsson, S., Thorleifsson, G., Grant, S. F., Emilsson, V., Gunnarsdottir, S., Adeyemo, A., Chen, Y., Chen, G., Reynisdottir, I., Benediktsson, R., Hinney, A., Hansen, T., Andersen, G., Borch-Johnsen, K., Jorgensen, T., Schafer, H., Faruque, M., Doumatey, A., Zhou, J., Wilensky, R. L., Reilly, M. P., Rader, D. J., Bagger, Y., Christiansen, C., Sigurdsson, G., Hebebrand, J., Pedersen, O., Thorsteinsdottir, U., Gulcher, J. R., Kong, A., Rotimi, C., & Stefansson, K. (2007). Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet, 39*(2), 218-225. doi: 10.1038/ng1960

Holm, H., Gudbjartsson, D. F., Sulem, P., Masson, G., Helgadottir, H. T., Zanon, C., Magnusson, O. T., Helgason, A., Saemundsdottir, J., Gylfason, A., Stefansdottir, H., Gretarsdottir, S., Matthiasson, S. E., Thorgeirsson, G. M., Jonasdottir, A., Sigurdsson, A., Stefansson, H., Werge, T., Rafnar, T., Kiemeney, L. A., Parvez, B., Muhammad, R., Roden, D. M., Darbar, D., Thorleifsson, G., Walters, G. B., Kong, A., Thorsteinsdottir, U., Arnar, D. O., & Stefansson, K. (2011). A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet, 43*(4), 316-320. doi: 10.1038/ng.781

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet, 44*(8), 955-959. doi: 10.1038/ng.2354

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 (Bethesda), 1*(6), 457-470. doi: 10.1534/g3.111.001198

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet, 5*(6), e1000529. doi: 10.1371/journal.pgen.1000529

Hu, Y., Willer, C., Zhan, X., Kang, H. M., & Abecasis, G. R. (2013). Accurate local-ancestry inference in exome-sequenced admixed individuals via off-target sequence reads. *Am J Hum Genet, 93*(5), 891-899. doi: 10.1016/j.ajhg.2013.10.008

Huang, J., Ellinghaus, D., Franke, A., Howie, B., & Li, Y. (2012). 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur J Hum Genet, 20*(7), 801-805. doi: 10.1038/ejhg.2012.3

International HapMap, Consortium. (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299-1320. doi: 10.1038/nature04226

International HapMap, Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Gibbs, R. A., Muzny, D. M., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarrol, S. A., Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S. B., Pollack, S., Price, A. L., Soranzo, N., Bonnen, P. E., Gibbs, R. A., Gonzaga-Jauregui, C., Keinan, A., Price, A. L., Yu, F., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S. F., Zhang, Q., Ghori, M. J., McGinnis, R., McLaren, W., Pollack, S., Price, A. L., Schaffner, S. F., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Manca, M. C., Marshall, P. A., Matsuda, I., Ngare, D., Wang, V. O., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D., & McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature, 467*(7311), 52-58. doi: 10.1038/nature09298

International HapMap, Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M., Tsui, S. K., Xue, H., Wong, J. T., Galver, L. M., Fan, J. B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J. F., Phillips, M. S., Roumy, S., Sallee, C., Verner, A., Hudson, T. J., Kwok, P. Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L. C., Mak, W., Song, Y. Q., Tam, P. K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler,

D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Gibbs, R. A., Belmont, J. W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Wheeler, D. A., Yakub, I., Gabriel, S. B., Onofrio, R. C., Richter, D. J., Ziaugra, L., Birren, B. W., Daly, M. J., Altshuler, D., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archeveque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., & Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature, 449*(7164), 851-861. doi: 10.1038/nature06258

Ioannidis, J. P., Ntzani, E. E., & Trikalinos, T. A. (2004). 'Racial' differences in genetic effects for complex diseases. *Nat Genet, 36*(12), 1312-1318. doi: 10.1038/ng1474

Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., Kivinen, K., Bojang, K. A., Conway, D. J., Pinder, M., Sirugo, G., Sisay-Joof, F., Usen, S., Auburn, S., Bumpstead, S. J., Campino, S., Coffey, A., Dunham, A., Fry, A. E., Green, A., Gwilliam, R., Hunt, S. E., Inouye, M., Jeffreys, A. E., Mendy, A., Palotie, A., Potter, S., Ragoussis, J., Rogers, J., Rowlands, K., Somaskantharajah, E., Whittaker, P., Widden, C., Donnelly, P., Howie, B., Marchini, J., Morris, A., SanJoaquin, M., Achidi, E. A., Agbenyega, T., Allen, A., Amodu, O., Corran, P., Djimde, A., Dolo, A., Doumbo, O. K., Drakeley, C., Dunstan, S., Evans, J., Farrar, J., Fernando, D., Hien, T. T., Horstmann, R. D., Ibrahim, M., Karunaweera, N., Kokwaro, G., Koram, K. A., Lemnge, M., Makani, J., Marsh, K., Michon, P., Modiano, D., Molyneux, M. E., Mueller, I., Parker, M., Peshu, N., Plowe, C. V., Puijalon, O., Reeder, J., Reyburn, H., Riley, E. M., Sakuntabhai, A., Singhasivanon, P., Sirima, S., Tall, A., Taylor, T. E., Thera, M., Troye-Blomberg, M., Williams, T. N., Wilson, M., Kwiatkowski, D. P., Wellcome Trust Case Control, Consortium, & Malaria Genomic Epidemiology, Network. (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat Genet, 41*(6), 657-665. doi: 10.1038/ng.388

Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y., & Kamatani, N. (2010). Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat Genet, 42*(3), 210-215. doi: 10.1038/ng.531

Kang, J., Huang, K. C., Xu, Z., Wang, Y., Abecasis, G. R., & Li, Y. (2013). AbCD: arbitrary coverage design for sequencing-based genetic studies. *Bioinformatics, 29*(6), 799-801. doi: 10.1093/bioinformatics/btt041

Kittles, R. A., Chen, W., Panguluri, R. K., Ahaghotu, C., Jackson, A., Adebamowo, C. A., Griffin, R., Williams, T., Ukoli, F., Adams-Campbell, L., Kwagyan, J., Isaacs, W., Freeman, V., & Dunston, G. M. (2002). CYP3A4-V and prostate cancer in African Americans: causal or confounding association because of population stratification? *Hum Genet, 110*(6), 553-560. doi: 10.1007/s00439-002-0731-5

Kooner, J. S., Saleheen, D., Sim, X., Sehmi, J., Zhang, W., Frossard, P., Been, L. F., Chia, K. S., Dimas, A. S., Hassanali, N., Jafar, T., Jowett, J. B., Li, X., Radha, V., Rees, S. D., Takeuchi, F., Young, R., Aung, T., Basit, A., Chidambaram, M., Das, D., Grundberg, E., Hedman, A. K., Hydrie, Z. I., Islam, M., Khor, C. C., Kowlessur, S., Kristensen, M. M., Liju, S., Lim, W. Y., Matthews, D. R., Liu, J., Morris, A. P., Nica, A. C., Pinidiyapathirage, J. M., Prokopenko, I., Rasheed, A., Samuel, M., Shah, N., Shera, A. S., Small, K. S., Suo, C., Wickremasinghe, A. R., Wong, T. Y., Yang, M., Zhang, F., Diagram, MuTher, Abecasis, G. R., Barnett, A. H., Caulfield, M., Deloukas, P., Frayling, T. M., Froguel, P., Kato, N., Katulanda, P., Kelly, M. A., Liang, J., Mohan, V., Sanghera, D. K., Scott, J., Seielstad, M., Zimmet, P. Z., Elliott, P., Teo, Y. Y., McCarthy, M. I., Danesh, J., Tai, E. S., & Chambers, J. C. (2011). Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet, 43*(10), 984-989. doi: 10.1038/ng.921

Kryukov, G. V., Shpunt, A., Stamatoyannopoulos, J. A., & Sunyaev, S. R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci U S A, 106*(10), 3871-3876. doi: 10.1073/pnas.0812824106

Lamason, R. L., Mohideen, M. A., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., Jurynec, M. J., Mao, X., Humphreville, V. R., Humbert, J. E., Sinha, S., Moore, J. L., Jagadeeswaran, P., Zhao, W., Ning, G., Makalowska, I., McKeigue, P. M., O'Donnell, D., Kittles, R., Parra, E. J., Mangini, N. J., Grunwald, D. J., Shriver, M. D., Canfield, V. A., & Cheng, K. C. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science, 310*(5755), 1782-1786. doi: 10.1126/science.1116238

Lange, L. A., Croteau-Chonka, D. C., Marvelle, A. F., Qin, L., Gaulton, K. J., Kuzawa, C. W., McDade, T. W., Wang, Y., Li, Y., Levy, S., Borja, J. B., Lange, E. M., Adair, L. S., & Mohlke, K. L. (2010). Genome-wide association study of homocysteine levels in Filipinos provides evidence for CPS1 in women and a stronger MTHFR effect in young adults. *Hum Mol Genet, 19*(10), 2050-2058. doi: 10.1093/hmg/ddq062

Lautenberger, J. A., Stephens, J. C., O'Brien, S. J., & Smith, M. W. (2000). Significant admixture linkage disequilibrium across 30 cM around the FY locus in African Americans. *Am J Hum Genet, 66*(3), 969-978. doi: 10.1086/302820

Levin, A. M., Iannuzzi, M. C., Montgomery, C. G., Trudeau, S., Datta, I., Adrianto, I., Chitale, D. A., McKeigue, P., & Rybicki, B. A. (2014). Admixture fine-mapping in African Americans implicates XAF1 as a possible sarcoidosis risk gene. *PLoS One, 9*(3), e92646. doi: 10.1371/journal.pone.0092646

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi: 10.1093/bioinformatics/btp324

Li, J., Glessner, J. T., Zhang, H., Hou, C., Wei, Z., Bradfield, J. P., Mentch, F. D., Guo, Y., Kim, C., Xia, Q., Chiavacci, R. M., Thomas, K. A., Qiu, H., Grant, S. F., Furth, S. L., Hakonarson, H., & Sleiman, P. M. (2013). GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. *Hum Mol Genet, 22*(7), 1457-1464. doi: 10.1093/hmg/dds534

Li, L., Li, Y., Browning, S. R., Browning, B. L., Slater, A. J., Kong, X., Aponte, J. L., Mooser, V. E., Chissoe, S. L., Whittaker, J. C., Nelson, M. R., & Ehm, M. G. (2011). Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS One, 6*(9), e24945. doi: 10.1371/journal.pone.0024945

Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics, 165*(4), 2213-2233.

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res, 21*(6), 940-951. doi: 10.1101/gr.117259.110

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol, 34*(8), 816-834. doi: 10.1002/gepi.20533

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annu Rev Genomics Hum Genet, 10*, 387-406. doi: 10.1146/annurev.genom.9.081307.164242

Liu, E. Y., Buyske, S., Aragaki, A. K., Peters, U., Boerwinkle, E., Carlson, C., Carty, C., Crawford, D. C., Haessler, J., Hindorff, L. A., Marchand, L. L., Manolio, T. A., Matise, T., Wang, W., Kooperberg, C., North, K. E., & Li, Y. (2012). Genotype imputation of

Metabochip SNPs using a study-specific reference panel of ~4,000 haplotypes in African Americans from the Women's Health Initiative. *Genet Epidemiol, 36*(2), 107-117. doi: 10.1002/gepi.21603

Liu, E. Y., Li, M., Wang, W., & Li, Y. (2013). MaCH-admix: genotype imputation for admixed populations. *Genet Epidemiol, 37*(1), 25-37. doi: 10.1002/gepi.21690

Liu, J., Lewinger, J. P., Gilliland, F. D., Gauderman, W. J., & Conti, D. V. (2013). Confounding and heterogeneity in genetic association studies with admixed populations. *Am J Epidemiol, 177*(4), 351-360. doi: 10.1093/aje/kws234

Lo, K. S., Wilson, J. G., Lange, L. A., Folsom, A. R., Galarneau, G., Ganesh, S. K., Grant, S. F., Keating, B. J., McCarroll, S. A., Mohler, E. R., 3rd, O'Donnell, C. J., Palmas, W., Tang, W., Tracy, R. P., Reiner, A. P., & Lettre, G. (2011). Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. *Hum Genet, 129*(3), 307-317. doi: 10.1007/s00439-010-0925-1

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., Croteau-Chonka, D. C., Esko, T., Fall, T., Ferreira, T., Gustafsson, S., Kutalik, Z., Luan, J., Magi, R., Randall, J. C., Winkler, T. W., Wood, A. R., Workalemahu, T., Faul, J. D., Smith, J. A., Hua Zhao, J., Zhao, W., Chen, J., Fehrmann, R., Hedman, A. K., Karjalainen, J., Schmidt, E. M., Absher, D., Amin, N., Anderson, D., Beekman, M., Bolton, J. L., Bragg-Gresham, J. L., Buyske, S., Demirkan, A., Deng, G., Ehret, G. B., Feenstra, B., Feitosa, M. F., Fischer, K., Goel, A., Gong, J., Jackson, A. U., Kanoni, S., Kleber, M. E., Kristiansson, K., Lim, U., Lotay, V., Mangino, M., Mateo Leach, I., Medina-Gomez, C., Medland, S. E., Nalls, M. A., Palmer, C. D., Pasko, D., Pechlivanis, S., Peters, M. J., Prokopenko, I., Shungin, D., Stancakova, A., Strawbridge, R. J., Ju Sung, Y., Tanaka, T., Teumer, A., Trompet, S., van der Laan, S. W., van Setten, J., Van Vliet-Ostaptchouk, J. V., Wang, Z., Yengo, L., Zhang, W., Isaacs, A., Albrecht, E., Arnlov, J., Arscott, G. M., Attwood, A. P., Bandinelli, S., Barrett, A., Bas, I. N., Bellis, C., Bennett, A. J., Berne, C., Blagieva, R., Bluher, M., Bohringer, S., Bonnycastle, L. L., Bottcher, Y., Boyd, H. A., Bruinenberg, M., Caspersen, I. H., Ida Chen, Y. D., Clarke, R., Daw, E. W., de Craen, A. J., Delgado, G., Dimitriou, M., Doney, A. S., Eklund, N., Estrada, K., Eury, E., Folkersen, L., Fraser, R. M., Garcia, M. E., Geller, F., Giedraitis, V., Gigante, B., Go, A. S., Golay, A., Goodall, A. H., Gordon, S. D., Gorski, M., Grabe, H. J., Grallert, H., Grammer, T. B., Grassler, J., Gronberg, H., Groves, C. J., Gusto, G., Haessler, J., Hall, P., Haller, T., Hallmans, G., Hartman, C. A., Hassinen, M., Hayward, C., Heard-Costa, N. L., Helmer, Q., Hengstenberg, C., Holmen, O., Hottenga, J. J., James, A. L., Jeff, J. M., Johansson, A., Jolley, J., Juliusdottir, T., Kinnunen, L., Koenig, W., Koskenvuo, M., Kratzer, W., Laitinen, J., Lamina, C., Leander, K., Lee, N. R., Lichtner, P., Lind, L., Lindstrom, J., Sin Lo, K., Lobbens, S., Lorbeer, R., Lu, Y., Mach, F., Magnusson, P. K., Mahajan, A., McArdle, W. L., McLachlan, S., Menni, C., Merger, S., Mihailov, E., Milani, L., Moayyeri, A., Monda, K. L., Morken, M. A., Mulas, A., Muller, G., Muller-Nurasyid, M., Musk, A. W., Nagaraja,

R., Nothen, M. M., Nolte, I. M., Pilz, S., Rayner, N. W., Renstrom, F., Rettig, R., Ried, J. S., Ripke, S., Robertson, N. R., Rose, L. M., Sanna, S., Scharnagl, H., Scholtens, S., Schumacher, F. R., Scott, W. R., Seufferlein, T., Shi, J., Vernon Smith, A., Smolonska, J., Stanton, A. V., Steinthorsdottir, V., Stirrups, K., Stringham, H. M., Sundstrom, J., Swertz, M. A., Swift, A. J., Syvanen, A. C., Tan, S. T., Tayo, B. O., Thorand, B., Thorleifsson, G., Tyrer, J. P., Uh, H. W., Vandenput, L., Verhulst, F. C., Vermeulen, S. H., Verweij, N., Vonk, J. M., Waite, L. L., Warren, H. R., Waterworth, D., Weedon, M. N., Wilkens, L. R., Willenborg, C., Wilsgaard, T., Wojczynski, M. K., Wong, A., Wright, A. F., Zhang, Q., LifeLines Cohort, Study, Brennan, E. P., Choi, M., Dastani, Z., Drong, A. W., Eriksson, P., Franco-Cereceda, A., Gadin, J. R., Gharavi, A. G., Goddard, M. E., Handsaker, R. E., Huang, J., Karpe, F., Kathiresan, S., Keildson, S., Kiryluk, K., Kubo, M., Lee, J. Y., Liang, L., Lifton, R. P., Ma, B., McCarroll, S. A., McKnight, A. J., Min, J. L., Moffatt, M. F., Montgomery, G. W., Murabito, J. M., Nicholson, G., Nyholt, D. R., Okada, Y., Perry, J. R., Dorajoo, R., Reinmaa, E., Salem, R. M., Sandholm, N., Scott, R. A., Stolk, L., Takahashi, A., Tanaka, T., Van't Hooft, F. M., Vinkhuyzen, A. A., Westra, H. J., Zheng, W., Zondervan, K. T., Consortium, A. DIPOGen, Group, Agen-Bmi Working, Consortium, C. ARDIOGRAMplusC4D, Consortium, C. KDGen, Glgc, Icbp, Investigators, Magic, Mu, Ther Consortium, Consortium, M. IGen, Consortium, Page, ReproGen, Consortium, Consortium, Genie, International Endogene, Consortium, Heath, A. C., Arveiler, D., Bakker, S. J., Beilby, J., Bergman, R. N., Blangero, J., Bovet, P., Campbell, H., Caulfield, M. J., Cesana, G., Chakravarti, A., Chasman, D. I., Chines, P. S., Collins, F. S., Crawford, D. C., Cupples, L. A., Cusi, D., Danesh, J., de Faire, U., den Ruijter, H. M., Dominiczak, A. F., Erbel, R., Erdmann, J., Eriksson, J. G., Farrall, M., Felix, S. B., Ferrannini, E., Ferrieres, J., Ford, I., Forouhi, N. G., Forrester, T., Franco, O. H., Gansevoort, R. T., Gejman, P. V., Gieger, C., Gottesman, O., Gudnason, V., Gyllensten, U., Hall, A. S., Harris, T. B., Hattersley, A. T., Hicks, A. A., Hindorff, L. A., Hingorani, A. D., Hofman, A., Homuth, G., Hovingh, G. K., Humphries, S. E., Hunt, S. C., Hypponen, E., Illig, T., Jacobs, K. B., Jarvelin, M. R., Jockel, K. H., Johansen, B., Jousilahti, P., Jukema, J. W., Jula, A. M., Kaprio, J., Kastelein, J. J., Keinanen-Kiukaanniemi, S. M., Kiemeney, L. A., Knekt, P., Kooner, J. S., Kooperberg, C., Kovacs, P., Kraja, A. T., Kumari, M., Kuusisto, J., Lakka, T. A., Langenberg, C., Le Marchand, L., Lehtimaki, T., Lyssenko, V., Mannisto, S., Marette, A., Matise, T. C., McKenzie, C. A., McKnight, B., Moll, F. L., Morris, A. D., Morris, A. P., Murray, J. C., Nelis, M., Ohlsson, C., Oldehinkel, A. J., Ong, K. K., Madden, P. A., Pasterkamp, G., Peden, J. F., Peters, A., Postma, D. S., Pramstaller, P. P., Price, J. F., Qi, L., Raitakari, O. T., Rankinen, T., Rao, D. C., Rice, T. K., Ridker, P. M., Rioux, J. D., Ritchie, M. D., Rudan, I., Salomaa, V., Samani, N. J., Saramies, J., Sarzynski, M. A., Schunkert, H., Schwarz, P. E., Sever, P., Shuldiner, A. R., Sinisalo, J., Stolk, R. P., Strauch, K., Tonjes, A., Tregouet, D. A., Tremblay, A., Tremoli, E., Virtamo, J., Vohl, M. C., Volker, U., Waeber, G., Willemsen, G., Witteman, J. C., Zillikens, M. C., Adair, L. S., Amouyel, P., Asselbergs, F. W., Assimes, T. L., Bochud, M., Boehm, B. O., Boerwinkle, E., Bornstein, S. R., Bottinger, E. P., Bouchard, C., Cauchi, S., Chambers, J. C., Chanock, S. J., Cooper, R. S., de Bakker, P. I., Dedoussis, G., Ferrucci, L., Franks, P. W., Froguel, P., Groop, L. C., Haiman, C. A., Hamsten, A., Hui, J., Hunter, D. J., Hveem, K., Kaplan, R. C., Kivimaki, M., Kuh, D., Laakso, M., Liu, Y., Martin, N. G., Marz, W., Melbye, M., Metspalu, A., Moebus, S., Munroe, P. B., Njolstad, I., Oostra, B. A., Palmer, C. N., Pedersen, N. L.,

Perola, M., Perusse, L., Peters, U., Power, C., Quertermous, T., Rauramaa, R., Rivadeneira, F., Saaristo, T. E., Saleheen, D., Sattar, N., Schadt, E. E., Schlessinger, D., Slagboom, P. E., Snieder, H., Spector, T. D., Thorsteinsdottir, U., Stumvoll, M., Tuomilehto, J., Uitterlinden, A. G., Uusitupa, M., van der Harst, P., Walker, M., Wallaschofski, H., Wareham, N. J., Watkins, H., Weir, D. R., Wichmann, H. E., Wilson, J. F., Zanen, P., Borecki, I. B., Deloukas, P., Fox, C. S., Heid, I. M., O'Connell, J. R., Strachan, D. P., Stefansson, K., van Duijn, C. M., Abecasis, G. R., Franke, L., Frayling, T. M., McCarthy, M. I., Visscher, P. M., Scherag, A., Willer, C. J., Boehnke, M., Mohlke, K. L., Lindgren, C. M., Beckmann, J. S., Barroso, I., North, K. E., Ingelsson, E., Hirschhorn, J. N., Loos, R. J., & Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature, 518*(7538), 197-206. doi: 10.1038/nature14177

Loth, D. W., Artigas, M. S., Gharib, S. A., Wain, L. V., Franceschini, N., Koch, B., Pottinger, T. D., Smith, A. V., Duan, Q., Oldmeadow, C., Lee, M. K., Strachan, D. P., James, A. L., Huffman, J. E., Vitart, V., Ramasamy, A., Wareham, N. J., Kaprio, J., Wang, X. Q., Trochet, H., Kahonen, M., Flexeder, C., Albrecht, E., Lopez, L. M., de Jong, K., Thyagarajan, B., Alves, A. C., Enroth, S., Omenaas, E., Joshi, P. K., Fall, T., Vinuela, A., Launer, L. J., Loehr, L. R., Fornage, M., Li, G., Wilk, J. B., Tang, W., Manichaikul, A., Lahousse, L., Harris, T. B., North, K. E., Rudnicka, A. R., Hui, J., Gu, X., Lumley, T., Wright, A. F., Hastie, N. D., Campbell, S., Kumar, R., Pin, I., Scott, R. A., Pietilainen, K. H., Surakka, I., Liu, Y., Holliday, E. G., Schulz, H., Heinrich, J., Davies, G., Vonk, J. M., Wojczynski, M., Pouta, A., Johansson, A., Wild, S. H., Ingelsson, E., Rivadeneira, F., Volzke, H., Hysi, P. G., Eiriksdottir, G., Morrison, A. C., Rotter, J. I., Gao, W., Postma, D. S., White, W. B., Rich, S. S., Hofman, A., Aspelund, T., Couper, D., Smith, L. J., Psaty, B. M., Lohman, K., Burchard, E. G., Uitterlinden, A. G., Garcia, M., Joubert, B. R., McArdle, W. L., Musk, A. B., Hansel, N., Heckbert, S. R., Zgaga, L., van Meurs, J. B., Navarro, P., Rudan, I., Oh, Y. M., Redline, S., Jarvis, D. L., Zhao, J. H., Rantanen, T., O'Connor, G. T., Ripatti, S., Scott, R. J., Karrasch, S., Grallert, H., Gaddis, N. C., Starr, J. M., Wijmenga, C., Minster, R. L., Lederer, D. J., Pekkanen, J., Gyllensten, U., Campbell, H., Morris, A. P., Glaser, S., Hammond, C. J., Burkart, K. M., Beilby, J., Kritchevsky, S. B., Gudnason, V., Hancock, D. B., Williams, O. D., Polasek, O., Zemunik, T., Kolcic, I., Petrini, M. F., Wjst, M., Kim, W. J., Porteous, D. J., Scotland, G., Smith, B. H., Viljanen, A., Heliovaara, M., Attia, J. R., Sayers, I., Hampel, R., Gieger, C., Deary, I. J., Boezen, H. M., Newman, A., Jarvelin, M. R., Wilson, J. F., Lind, L., Stricker, B. H., Teumer, A., Spector, T. D., Melen, E., Peters, M. J., Lange, L. A., Barr, R. G., Bracke, K. R., Verhamme, F. M., Sung, J., Hiemstra, P. S., Cassano, P. A., Sood, A., Hayward, C., Dupuis, J., Hall, I. P., Brusselle, G. G., Tobin, M. D., & London, S. J. (2014). Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nat Genet, 46*(7), 669-677. doi: 10.1038/ng.3011

Mao, X., Li, Y., Liu, Y., Lange, L., & Li, M. (2013). Testing genetic association with rare variants in admixed populations. *Genet Epidemiol, 37*(1), 38-47. doi: 10.1002/gepi.21687

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet, 93*(2), 278-288. doi: 10.1016/j.ajhg.2013.06.020

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet, 11*(7), 499-511. doi: 10.1038/nrg2796

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet, 39*(7), 906-913. doi: 10.1038/ng2088

Marvelle, A. F., Lange, L. A., Qin, L., Wang, Y., Lange, E. M., Adair, L. S., & Mohlke, K. L. (2007). Comparison of ENCODE region SNPs between Cebu Filipino and Asian HapMap samples. *J Hum Genet, 52*(9), 729-737. doi: 10.1007/s10038-007-0175-9

McCarthy, M. I., & Hirschhorn, J. N. (2008). Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet, 17*(R2), R156-165. doi: 10.1093/hmg/ddn289

McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R., Hinds, D. A., Pennacchio, L. A., Tybjaerg-Hansen, A., Folsom, A. R., Boerwinkle, E., Hobbs, H. H., & Cohen, J. C. (2007). A common allele on chromosome 9 associated with coronary heart disease. *Science, 316*(5830), 1488-1491. doi: 10.1126/science.1142447

Mensah, G. A., Mokdad, A. H., Ford, E. S., Greenlund, K. J., & Croft, J. B. (2005). State of disparities in cardiovascular health in the United States. *Circulation, 111*(10), 1233-1241. doi: 10.1161/01.CIR.0000158136.76824.04

Morris, A. P. (2011). Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol, 35*(8), 809-822. doi: 10.1002/gepi.20630

Muntaner, C., Nieto, F. J., Cooper, L., Meyer, J., Szklo, M., & Tyroler, H. A. (1998). Work organization and atherosclerosis: findings from the ARIC study. Atherosclerosis Risk in Communities. *Am J Prev Med, 14*(1), 9-18.

Musunuru, K., Lettre, G., Young, T., Farlow, D. N., Pirruccello, J. P., Ejebe, K. G., Keating, B. J., Yang, Q., Chen, M. H., Lapchyk, N., Crenshaw, A., Ziaugra, L., Rachupka, A., Benjamin, E. J., Cupples, L. A., Fornage, M., Fox, E. R., Heckbert, S. R., Hirschhorn, J. N., Newton-Cheh, C., Nizzari, M. M., Paltoo, D. N., Papanicolaou, G. J., Patel, S. R., Psaty, B. M., Rader, D. J., Redline, S., Rich, S. S., Rotter, J. I., Taylor, H. A., Jr., Tracy,

R. P., Vasan, R. S., Wilson, J. G., Kathiresan, S., Fabsitz, R. R., Boerwinkle, E., Gabriel, S. B., & Resource, Nhlbi Candidate Gene Association. (2010). Candidate gene association resource (CARe): design, methods, and proof of concept. *Circ Cardiovasc Genet, 3*(3), 267-275. doi: 10.1161/CIRCGENETICS.109.882696

Myles, S., Davison, D., Barrett, J., Stoneking, M., & Timpson, N. (2008). Worldwide population differentiation at disease-associated SNPs. *BMC Med Genomics, 1*, 22. doi: 10.1186/1755-8794-1-22

Ng, M. C. (2015). Genetics of Type 2 Diabetes in African Americans. *Curr Diab Rep, 15*(10), 74. doi: 10.1007/s11892-015-0651-0

Norris, J. M., Langefeld, C. D., Talbert, M. E., Wing, M. R., Haritunians, T., Fingerlin, T. E., Hanley, A. J., Ziegler, J. T., Taylor, K. D., Haffner, S. M., Chen, Y. D., Bowden, D. W., & Wagenknecht, L. E. (2009). Genome-wide association study and follow-up analysis of adiposity traits in Hispanic Americans: the IRAS Family Study. *Obesity (Silver Spring), 17*(10), 1932-1941. doi: 10.1038/oby.2009.143

Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R., Forrester, T., Allison, D. B., Deka, R., Ferrell, R. E., & Shriver, M. D. (1998). Estimating African American admixture proportions by use of population-specific alleles. *Am J Hum Genet, 63*(6), 1839-1851. doi: 10.1086/302148

Pasaniuc, B., Sankararaman, S., Kimmel, G., & Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. *Bioinformatics, 25*(12), i213-221. doi: 10.1093/bioinformatics/btp197

Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Zaitlen, N., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J. G., Avila, P. C., Rodriguez-Santana, J., Chen, G. K., Le Marchand, L., Henderson, B., Reich, D., Haiman, C. A., Gonzalez Burchard, E., & Halperin, E. (2013). Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics, 29*(11), 1407-1415. doi: 10.1093/bioinformatics/btt166

Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G. K., Tandon, A., Kao, W. H., Ruczinski, I., Fornage, M., Siscovick, D. S., Zhu, X., Larkin, E., Lange, L. A., Cupples, L. A., Yang, Q., Akylbekova, E. L., Musani, S. K., Divers, J., Mychaleckyj, J., Li, M., Papanicolaou, G. J., Millikan, R. C., Ambrosone, C. B., John, E. M., Bernstein, L., Zheng, W., Hu, J. J., Ziegler, R. G., Nyante, S. J., Bandera, E. V., Ingles, S. A., Press, M. F., Chanock, S. J., Deming, S. L., Rodriguez-Gil, J. L., Palmer, C. D., Buxbaum, S., Ekunwe, L., Hirschhorn, J. N., Henderson, B. E., Myers, S., Haiman, C. A., Reich, D., Patterson, N.,

Wilson, J. G., & Price, A. L. (2011). Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genet, 7*(4), e1001371. doi: 10.1371/journal.pgen.1001371

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., O'Brien, S. J., Altshuler, D., Daly, M. J., & Reich, D. (2004). Methods for high-density admixture mapping of disease genes. *Am J Hum Genet, 74*(5), 979-1000. doi: 10.1086/420871

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet, 38*(8), 904-909. doi: 10.1038/ng1847

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., & Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet, 5*(6), e1000519. doi: 10.1371/journal.pgen.1000519

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat Rev Genet, 11*(7), 459-463. doi: 10.1038/nrg2813

Pritchard, J. K., & Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet, 69*(1), 1-14. doi: 10.1086/321275

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi: 10.1086/519795

Qin, H., Morris, N., Kang, S. J., Li, M., Tayo, B., Lyon, H., Hirschhorn, J., Cooper, R. S., & Zhu, X. (2010). Interrogating local population structure for fine mapping in genome-wide association studies. *Bioinformatics, 26*(23), 2961-2968. doi: 10.1093/bioinformatics/btq560

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., & Lander, E. S. (2001). Linkage

disequilibrium in the human genome. *Nature, 411*(6834), 199-204. doi: 10.1038/35075590

Reich, D., Patterson, N., Ramesh, V., De Jager, P. L., McDonald, G. J., Tandon, A., Choy, E., Hu, D., Tamraz, B., Pawlikowska, L., Wassel-Fyr, C., Huntsman, S., Waliszewska, A., Rossin, E., Li, R., Garcia, M., Reiner, A., Ferrell, R., Cummings, S., Kwok, P. Y., Harris, T., Zmuda, J. M., Ziv, E., Health, Aging, & Body Composition, Study. (2007). Admixture mapping of an allele affecting interleukin 6 soluble receptor and interleukin 6 levels. *Am J Hum Genet, 80*(4), 716-726. doi: 10.1086/513206

Reiner, A. P., Beleza, S., Franceschini, N., Auer, P. L., Robinson, J. G., Kooperberg, C., Peters, U., & Tang, H. (2012). Genome-wide association and population genetic analysis of C-reactive protein in African American and Hispanic American women. *Am J Hum Genet, 91*(3), 502-512. doi: 10.1016/j.ajhg.2012.07.023

Reiner, A. P., Lettre, G., Nalls, M. A., Ganesh, S. K., Mathias, R., Austin, M. A., Dean, E., Arepalli, S., Britton, A., Chen, Z., Couper, D., Curb, J. D., Eaton, C. B., Fornage, M., Grant, S. F., Harris, T. B., Hernandez, D., Kamatini, N., Keating, B. J., Kubo, M., LaCroix, A., Lange, L. A., Liu, S., Lohman, K., Meng, Y., Mohler, E. R., 3rd, Musani, S., Nakamura, Y., O'Donnell, C. J., Okada, Y., Palmer, C. D., Papanicolaou, G. J., Patel, K. V., Singleton, A. B., Takahashi, A., Tang, H., Taylor, H. A., Jr., Taylor, K., Thomson, C., Yanek, L. R., Yang, L., Ziv, E., Zonderman, A. B., Folsom, A. R., Evans, M. K., Liu, Y., Becker, D. M., Snively, B. M., & Wilson, J. G. (2011). Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). *PLoS Genet, 7*(6), e1002108. doi: 10.1371/journal.pgen.1002108

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., & Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nat Rev Genet, 11*(5), 356-366. doi: 10.1038/nrg2760

Rosenberg, N. A., & Nordborg, M. (2006). A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations. *Genetics, 173*(3), 1665-1678. doi: 10.1534/genetics.105.055335

Sampson, J. N., Jacobs, K., Wang, Z., Yeager, M., Chanock, S., & Chatterjee, N. (2012). A two-platform design for next generation genome-wide association studies. *Genet Epidemiol, 36*(4), 400-408. doi: 10.1002/gepi.21634

Sankararaman, S., Sridhar, S., Kimmel, G., & Halperin, E. (2008). Estimating local ancestry in admixed populations. *Am J Hum Genet, 82*(2), 290-303. doi: 10.1016/j.ajhg.2007.09.022

Sanna, S. (2012). Using low-pass whole genome sequencing to create a reference population for genome imputation in an isolated population: examples from the SardiNIA study. *Presented at the 62nd Annual Meeting of The American Society of Human Genetics. San Francisco, CA*.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., & Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res, 15*(11), 1576-1583. doi: 10.1101/gr.3709305

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., Prokunina-Olsson, L., Ding, C. J., Swift, A. J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X. Y., Conneely, K. N., Riebow, N. L., Sprau, A. G., Tong, M., White, P. P., Hetrick, K. N., Barnhart, M. W., Bark, C. W., Goldstein, J. L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T. A., Watanabe, R. M., Valle, T. T., Kinnunen, L., Abecasis, G. R., Pugh, E. W., Doheny, K. F., Bergman, R. N., Tuomilehto, J., Collins, F. S., & Boehnke, M. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science, 316*(5829), 1341-1345. doi: 10.1126/science.1142382

Seldin, M. F., Pasaniuc, B., & Price, A. L. (2011). New approaches to disease mapping in admixed populations. *Nat Rev Genet, 12*(8), 523-528. doi: 10.1038/nrg3002

Shriner, D. (2013). Overview of admixture mapping. *Curr Protoc Hum Genet, Chapter 1*, Unit 1 23. doi: 10.1002/0471142905.hg0123s76

Shriner, D., Adeyemo, A., & Rotimi, C. N. (2011). Joint ancestry and association testing in admixed individuals. *PLoS Comput Biol, 7*(12), e1002325. doi: 10.1371/journal.pcbi.1002325

Silva-Zolezzi, I., Hidalgo-Miranda, A., Estrada-Gil, J., Fernandez-Lopez, J. C., Uribe-Figueroa, L., Contreras, A., Balam-Ortiz, E., del Bosque-Plata, L., Velazquez-Fernandez, D., Lara, C., Goya, R., Hernandez-Lemus, E., Davila, C., Barrientos, E., March, S., & Jimenez-Sanchez, G. (2009). Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proc Natl Acad Sci U S A, 106*(21), 8611-8616. doi: 10.1073/pnas.0903045106

Smith, M. W., Patterson, N., Lautenberger, J. A., Truelove, A. L., McDonald, G. J., Waliszewska, A., Kessing, B. D., Malasky, M. J., Scafe, C., Le, E., De Jager, P. L.,

Mignault, A. A., Yi, Z., De The, G., Essex, M., Sankale, J. L., Moore, J. H., Poku, K., Phair, J. P., Goedert, J. J., Vlahov, D., Williams, S. M., Tishkoff, S. A., Winkler, C. A., De La Vega, F. M., Woodage, T., Sninsky, J. J., Hafler, D. A., Altshuler, D., Gilbert, D. A., O'Brien, S. J., & Reich, D. (2004). A high-density admixture map for disease gene discovery in african americans. *Am J Hum Genet, 74*(5), 1001-1013. doi: 10.1086/420856

Spencer, C. C., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet, 5*(5), e1000477. doi: 10.1371/journal.pgen.1000477

Stefflova, K., Dulik, M. C., Barnholtz-Sloan, J. S., Pai, A. A., Walker, A. H., & Rebbeck, T. R. (2011). Dissecting the within-Africa ancestry of populations of African descent in the Americas. *PLoS One, 6*(1), e14495. doi: 10.1371/journal.pone.0014495

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stutz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E. W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., Genomes Project, Consortium, Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E., & Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature, 526*(7571), 75-81. doi: 10.1038/nature15394

Sundquist, A., Fratkin, E., Do, C. B., & Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Res, 18*(4), 676-682. doi: 10.1101/gr.072850.107

Tandon, A., Chen, C. J., Penman, A., Hancock, H., James, M., Husain, D., Andreoli, C., Li, X., Kuo, J. Z., Idowu, O., Riche, D., Papavasilieou, E., Brauner, S., Smith, S. O., Hoadley, S., Richardson, C., Kieser, T., Vazquez, V., Chi, C., Fernandez, M., Harden, M., Cotch, M. F., Siscovick, D., Taylor, H. A., Wilson, J. G., Reich, D., Wong, T. Y., Klein, R., Klein, B. E., Rotter, J. I., Patterson, N., & Sobrin, L. (2015). African Ancestry Analysis and Admixture Genetic Mapping for Proliferative Diabetic Retinopathy in African Americans. *Invest Ophthalmol Vis Sci, 56*(6), 3999-4005. doi: 10.1167/iovs.15-16674

Tang, H., Coram, M., Wang, P., Zhu, X., & Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet, 79*(1), 1-12. doi: 10.1086/504302

Tang, H., Siegmund, D. O., Johnson, N. A., Romieu, I., & London, S. J. (2010). Joint testing of genotype and ancestry association in admixed families. *Genet Epidemiol, 34*(8), 783-791. doi: 10.1002/gepi.20520

Taylor, H. A., Jr., Wilson, J. G., Jones, D. W., Sarpong, D. F., Srinivasan, A., Garrison, R. J., Nelson, C., & Wyatt, S. B. (2005). Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. *Ethn Dis, 15*(4 Suppl 6), S6-4-17.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., Akey, J. M., Broad, G. O., Seattle, G. O., & Project, Nhlbi Exome Sequencing. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science, 337*(6090), 64-69. doi: 10.1126/science.1219240

Teo, Y. Y., Small, K. S., Fry, A. E., Wu, Y., Kwiatkowski, D. P., & Clark, T. G. (2009). Power consequences of linkage disequilibrium variation between populations. *Genet Epidemiol, 33*(2), 128-135. doi: 10.1002/gepi.20366

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., Johansen, C. T., Fouchier, S. W., Isaacs, A., Peloso, G. M., Barbalic, M., Ricketts, S. L., Bis, J. C., Aulchenko, Y. S., Thorleifsson, G., Feitosa, M. F., Chambers, J., Orho-Melander, M., Melander, O., Johnson, T., Li, X., Guo, X., Li, M., Shin Cho, Y., Jin Go, M., Jin Kim, Y., Lee, J. Y., Park, T., Kim, K., Sim, X., Twee-Hee Ong, R., Croteau-Chonka, D. C., Lange, L. A., Smith, J. D., Song, K., Hua Zhao, J., Yuan, X., Luan, J., Lamina, C., Ziegler, A., Zhang, W., Zee, R. Y., Wright, A. F., Witteman, J. C., Wilson, J. F., Willemsen, G., Wichmann, H. E., Whitfield, J. B., Waterworth, D. M., Wareham, N. J., Waeber, G., Vollenweider, P., Voight, B. F., Vitart, V., Uitterlinden, A. G., Uda, M., Tuomilehto, J., Thompson, J. R., Tanaka, T., Surakka, I., Stringham, H. M., Spector, T. D., Soranzo, N., Smit, J. H., Sinisalo, J., Silander, K., Sijbrands, E. J., Scuteri, A., Scott, J., Schlessinger, D., Sanna, S., Salomaa, V., Saharinen, J., Sabatti, C., Ruokonen, A., Rudan, I., Rose, L. M., Roberts, R., Rieder, M., Psaty, B. M., Pramstaller, P. P., Pichler, I., Perola, M., Penninx, B. W., Pedersen, N. L., Pattaro, C., Parker, A. N., Pare, G., Oostra, B. A., O'Donnell, C. J., Nieminen, M. S., Nickerson, D. A., Montgomery, G. W., Meitinger, T., McPherson, R., McCarthy, M. I., McArdle, W., Masson, D., Martin, N. G., Marroni, F., Mangino, M., Magnusson, P. K., Lucas, G., Luben, R., Loos, R. J., Lokki, M. L., Lettre, G., Langenberg, C., Launer, L. J., Lakatta, E. G., Laaksonen, R., Kyvik, K. O., Kronenberg, F., Konig, I. R., Khaw, K. T., Kaprio, J., Kaplan, L. M., Johansson, A., Jarvelin, M. R.,

Janssens, A. C., Ingelsson, E., Igl, W., Kees Hovingh, G., Hottenga, J. J., Hofman, A., Hicks, A. A., Hengstenberg, C., Heid, I. M., Hayward, C., Havulinna, A. S., Hastie, N. D., Harris, T. B., Haritunians, T., Hall, A. S., Gyllensten, U., Guiducci, C., Groop, L. C., Gonzalez, E., Gieger, C., Freimer, N. B., Ferrucci, L., Erdmann, J., Elliott, P., Ejebe, K. G., Doring, A., Dominiczak, A. F., Demissie, S., Deloukas, P., de Geus, E. J., de Faire, U., Crawford, G., Collins, F. S., Chen, Y. D., Caulfield, M. J., Campbell, H., Burtt, N. P., Bonnycastle, L. L., Boomsma, D. I., Boekholdt, S. M., Bergman, R. N., Barroso, I., Bandinelli, S., Ballantyne, C. M., Assimes, T. L., Quertermous, T., Altshuler, D., Seielstad, M., Wong, T. Y., Tai, E. S., Feranil, A. B., Kuzawa, C. W., Adair, L. S., Taylor, H. A., Jr., Borecki, I. B., Gabriel, S. B., Wilson, J. G., Holm, H., Thorsteinsdottir, U., Gudnason, V., Krauss, R. M., Mohlke, K. L., Ordovas, J. M., Munroe, P. B., Kooner, J. S., Tall, A. R., Hegele, R. A., Kastelein, J. J., Schadt, E. E., Rotter, J. I., Boerwinkle, E., Strachan, D. P., Mooser, V., Stefansson, K., Reilly, M. P., Samani, N. J., Schunkert, H., Cupples, L. A., Sandhu, M. S., Ridker, P. M., Rader, D. J., van Duijn, C. M., Peltonen, L., Abecasis, G. R., Boehnke, M., & Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature, 466*(7307), 707-713. doi: 10.1038/nature09270

The 1000 Genomes Project, Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073. doi: 10.1038/nature09534

The 1000 Genomes Project, Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., & McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65. doi: 10.1038/nature11632

The 1000 Genomes Project, Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi: 10.1038/nature15393

The Women's Health Initiative Study, Group. (1998). Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials, 19*(1), 61-109.

Tournamille, C., Colin, Y., Cartron, J. P., & Le Van Kim, C. (1995). Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet, 10*(2), 224-228. doi: 10.1038/ng0695-224

van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D. S., Elling, U., Allayee, H., Li, X., Radhakrishnan, A., Tan, S. T., Voss, K., Weichenberger, C. X., Albers, C. A., Al-Hussani, A., Asselbergs, F. W., Ciullo, M., Danjou, F., Dina, C., Esko, T., Evans, D. M., Franke, L., Gogele, M., Hartiala, J., Hersch, M., Holm, H., Hottenga, J. J., Kanoni, S., Kleber, M. E., Lagou, V., Langenberg, C., Lopez, L. M., Lyytikainen, L. P., Melander, O., Murgia, F., Nolte, I. M., O'Reilly, P. F., Padmanabhan, S., Parsa, A., Pirastu, N., Porcu, E., Portas, L., Prokopenko, I., Ried, J. S., Shin, S. Y., Tang, C. S., Teumer, A., Traglia, M., Ulivi, S., Westra, H. J., Yang, J., Zhao, J. H., Anni, F., Abdellaoui, A., Attwood, A., Balkau, B., Bandinelli, S., Bastardot, F., Benyamin, B., Boehm, B. O., Cookson, W. O., Das, D., de Bakker, P. I., de Boer, R. A., de Geus, E. J., de Moor, M. H., Dimitriou, M., Domingues, F. S., Doring, A., Engstrom, G., Eyjolfsson, G. I., Ferrucci, L., Fischer, K., Galanello, R., Garner, S. F., Genser, B., Gibson, Q. D., Girotto, G., Gudbjartsson, D. F., Harris, S. E., Hartikainen, A. L., Hastie, C. E., Hedblad, B., Illig, T., Jolley, J., Kahonen, M., Kema, I. P., Kemp, J. P., Liang, L., Lloyd-Jones, H., Loos, R. J., Meacham, S., Medland, S. E., Meisinger, C., Memari, Y., Mihailov, E., Miller, K., Moffatt, M. F., Nauck, M., Novatchkova, M., Nutile, T., Olafsson, I., Onundarson, P. T., Parracciani, D., Penninx, B. W., Perseu, L., Piga, A., Pistis, G., Pouta, A., Puc, U., Raitakari, O., Ring, S. M., Robino, A., Ruggiero, D., Ruokonen, A., Saint-Pierre, A., Sala, C., Salumets, A., Sambrook, J., Schepers, H., Schmidt, C. O., Sillje, H. H., Sladek, R., Smit, J. H., Starr, J. M., Stephens, J., Sulem, P., Tanaka, T., Thorsteinsdottir, U., Tragante, V., van Gilst, W. H., van Pelt, L. J., van Veldhuisen, D. J., Volker, U., Whitfield, J. B., Willemsen, G., Winkelmann, B. R., Wirnsberger, G., Algra, A., Cucca, F., d'Adamo, A. P., Danesh, J., Deary, I. J., Dominiczak, A. F., Elliott, P., Fortina, P., Froguel, P., Gasparini, P., Greinacher, A., Hazen, S. L., Jarvelin, M. R., Khaw, K. T., Lehtimaki, T., Maerz, W., Martin, N. G., Metspalu, A., Mitchell, B. D., Montgomery, G. W., Moore, C., Navis, G., Pirastu, M., Pramstaller, P. P., Ramirez-Solis, R., Schadt, E., Scott, J., Shuldiner, A. R., Smith, G. D., Smith, J. G., Snieder, H., Sorice, R., Spector, T. D., Stefansson, K., Stumvoll, M., Tang, W. H., Toniolo, D., Tonjes, A., Visscher, P. M., Vollenweider, P., Wareham, N. J., Wolffenbuttel, B. H., Boomsma, D. I., Beckmann, J. S., Dedoussis, G. V., Deloukas, P., Ferreira, M. A., Sanna, S., Uda, M., Hicks, A. A., Penninger, J. M., Gieger, C., Kooner, J. S., Ouwehand, W. H., Soranzo, N., & Chambers, J. C. (2012). Seventy-five genetic loci influencing the human red blood cell. *Nature, 492*(7429), 369-375. doi: 10.1038/nature11677

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *Am J Hum Genet, 90*(1), 7-24. doi: 10.1016/j.ajhg.2011.11.029

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., Burtt, N. P., Fuchsberger, C., Li, Y., Erdmann, J., Frayling, T. M., Heid, I. M., Jackson, A. U., Johnson, T., Kilpelainen, T. O., Lindgren, C. M., Morris, A. P., Prokopenko, I., Randall, J. C., Saxena, R., Soranzo, N., Speliotes, E. K., Teslovich, T. M., Wheeler, E., Maguire, J., Parkin, M., Potter, S., Rayner, N. W., Robertson, N., Stirrups, K., Winckler, W., Sanna, S., Mulas, A., Nagaraja, R., Cucca, F., Barroso, I., Deloukas, P., Loos, R. J., Kathiresan, S., Munroe, P. B., Newton-Cheh, C., Pfeufer, A., Samani, N. J., Schunkert, H., Hirschhorn, J. N., Altshuler, D., McCarthy, M. I., Abecasis, G. R., & Boehnke, M.

(2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet, 8*(8), e1002793. doi: 10.1371/journal.pgen.1002793

Wang, S., Ray, N., Rojas, W., Parra, M. V., Bedoya, G., Gallo, C., Poletti, G., Mazzotti, G., Hill, K., Hurtado, A. M., Camrena, B., Nicolini, H., Klitz, W., Barrantes, R., Molina, J. A., Freimer, N. B., Bortolini, M. C., Salzano, F. M., Petzl-Erler, M. L., Tsuneto, L. T., Dipierri, J. E., Alfaro, E. L., Bailliet, G., Bianchi, N. O., Llop, E., Rothhammer, F., Excoffier, L., & Ruiz-Linares, A. (2008). Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet, 4*(3), e1000037. doi: 10.1371/journal.pgen.1000037

Wang, X., Zhu, X., Qin, H., Cooper, R. S., Ewens, W. J., Li, C., & Li, M. (2011). Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics, 27*(5), 670-677. doi: 10.1093/bioinformatics/btq709

Wegmann, D., Kessner, D. E., Veeramah, K. R., Mathias, R. A., Nicolae, D. L., Yanek, L. R., Sun, Y. V., Torgerson, D. G., Rafaels, N., Mosley, T., Becker, L. C., Ruczinski, I., Beaty, T. H., Kardia, S. L., Meyers, D. A., Barnes, K. C., Becker, D. M., Freimer, N. B., & Novembre, J. (2011). Recombination rates in admixed individuals identified by ancestry-based inference. *Nat Genet, 43*(9), 847-853. doi: 10.1038/ng.894

Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet, 76*(5), 887-893. doi: 10.1086/429864

Winkler, C. A., Nelson, G. W., & Smith, M. W. (2010). Admixture mapping comes of age. *Annu Rev Genomics Hum Genet, 11*, 65-89. doi: 10.1146/annurev-genom-082509-141523

Zhang, J., & Stram, D. O. (2014). The role of local ancestry adjustment in association studies using admixed populations. *Genet Epidemiol, 38*(6), 502-515. doi: 10.1002/gepi.21835

Zheng, J., Li, Y., Abecasis, G. R., & Scheet, P. (2011). A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet Epidemiol, 35*(2), 102-110. doi: 10.1002/gepi.20552

Zhu, X., Luke, A., Cooper, R. S., Quertermous, T., Hanis, C., Mosley, T., Gu, C. C., Tang, H., Rao, D. C., Risch, N., & Weder, A. (2005). Admixture mapping for hypertension loci with genome-scan markers. *Nat Genet, 37*(2), 177-181. doi: 10.1038/ng1510