STATISTICAL METHODS AND ANALYSIS IN NEXT GENERATION SEQUENCING

by

Xiaowei Zhan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2014

Doctoral Committee:

Professor Gonçalo R. Abecasis, Chair
Professor Michael L. Boehnke
Professor Margit Burmeister
Assistant Professor Hyun Min Kang

Dedication

To my family for their love and support!

Acknowledgments

I am thankful to family, friends, committee and colleagues. Their help is indispensable to this thesis.

First, I would like to thank my advisor Dr. Gonçalo Abecasis. He is my academic advisor since I came to University of Michigan-Ann Arbor. He taught me statistical genetics and statistical computation in his unique way without reservation. He not only demonstrated how to think and research in the genetic field, but also patiently gave me time and financial support to enable me to grow my own research skills. I am very lucky to have you as you are an excellent mentor.

I would like to thank committee members for their support for my research: Dr. Michael Boehnke, Dr. Margit Burmeister, and Dr. Hyun Min Kang. They have dedicated their time and shared their comments and useful knowledge on my thesis. Besides that, Dr. Boehnke has always been extremely helpful for advising my graduate studies and career choices. Dr. Burmeister has introduced me wet lab skills in her bioinformatics boot camp course and this course arose my interests to combine statistical genetics methodology with biotechnology. Dr. Kang is supportive whenever I need his help, and he sets a good standard for me as an energetic, enthusiasm and efficient researcher (especially in developing algorithms).

I want to thank my friends and colleagues. They have helped me in numerous ways: they provided valuable suggestions on my presentations, maintained our high performance computation clusters, set up conference calls and reserved conference rooms, provided programming advices and communicated statistical modelling ideas. These made my graduate life not only easier but enjoyable. As my friends and colleagues are not mutually exclusive sets of people, I would express my appreciation in these rough categories: our Center for Statistical Genetics (CSG) staffs Laura J. Baker, Sean Caron, Dawn Keene, Mary Kate Wing, and Terry Gliedt; my friends in Bioinformatics program, especially Chaolong Wang, Peng Zhang, and Ryan Welch; CSG alumni, especially Bingshan Li, Liming Liang, Wei Chen, Jun Ding, Jennifer Bragg-Gresham, Yanming Li and Weihua Guan; people from Dr. Boehnke's lab, especially Xueling Sim, Tanya Telovish and Jeroen Huyghe; people from Dr. Abecasis' lab, especially Dajiang Liu, Lars Fritsche, Scott Vrieze, Goo Jun, Alex Tsoi Lam, Sara Rashkin and Matt Zawistowski; people from Biostatistics program, Jian Kang, Xi Xia, Xuejing Wang and Zihuai He; people who have helped my writing, Kirsten Herold, Simone Sessolo and Shawn Lacy; people at other institutions, David Larson, Daniel Koboldt, Kari Branham, Rinki Ratnapriya, Emily Chew and Anand Swaroop. I have learned a lot from all of them.

I would like to thank Dr. Kang James, Dr. Barry James and Dr. Zhuangyi Liu at the University of Minnesota-Duluth. From them, I built a solid knowledge base in mathematics and statistics.

At last, I want to thank my wife, my parents and all other family members for their continuous love and support for my doctoral study. Your encouragement and confidence are critical to my doctorate degree. For example, my wife, Yi Hou, prepares healthy and delicious food, which is essential for me to overcome research challenges. My parents, Zhenfei Deng and Peiqi Zhan, have devoted their best to help me grow up. I am grateful to have a sweet and supportive family.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

GWAS: Genome-wide association studies

NGS: Next generation sequencing

MAF: Minor allele frequency

SNP: Single nucleotide polymorphism

AMD: Age-related macular degeneration

PCA: Principal components analysis

HGDP: Human genome diversity project

POPRES: Population reference panel

Abstract

STATISTICAL METHODS AND ANALYSIS IN NEXT GENERATION SEQUENCING

by

Xiaowei Zhan

Chair: Gonçalo R. Abecasis

Next generation sequencing (NGS) is a technology that advances our knowledge of human medical genetics with unprecedented amount of data. This vast amount of data presents challenges to existing statistical methods. In this dissertation, I present three studies that demonstrate methods for efficiently analyzing NGS data using both simulated and real data.

In the first study, I develop a method to tackle ancestry inference using small amounts of sequence data. In comparison to microarray experiments, sequencing data produce more uneven coverage and genotypes with higher error rates than those traditionally used for principal components analysis (PCA) of genetic ancestry. I overcome some of these challenges using a new statistical method that models sequence data directly, using both on-target and off-target reads, without relying on intermediate genotype calls. My method achieves high accuracy in simulated data based on the Human Genome Diversity Panel as well as in a targeted sequencing study of age related macular degeneration. I show that inferred ancestry information helps address the population stratification problem. In our age-related macular degeneration study, our approach helps discover a high-risk rare variant in the Complement 3 gene.

In the second chapter, I develop a model-based ancestry inference method that improves upon previous the work described in the first study. It is based on a likelihood-based model of ancestral location, using sequencing data as input. Without losing accuracy, it increases computational efficiency. For each sample, a parallelizable optimization algorithm can infer ancestry using a fraction of the computational resources required for PCA-based methods such as the approach in the first study. Evaluation using in the Human Genome Diversity Panel and another age related macular degeneration data set demonstrates its accuracy and efficiency.

In the final study, I develop an improved genotype call method for low-coverage sequencing data. As high quality reference panels grow, it is helpful to incorporate these into genotype calling of new samples. Using a coalescent based simulation and real data from the 1000 Genomes Project, I evaluate the utility of my method (which uses a panel of previously sequenced samples) to improve analyses of samples sequenced at various depths. The improvement in accuracy and computation time will be measured as a function of reference panel size. This work will be useful to investigators undertaking sequencing and analysis of new human samples.

Chapter 1 Introduction

Next Generation Sequencing (NGS) has been revolutionizing genetic research. Compared to the previously popular automated Sanger sequencing, NGS provides accurate genetic information in a faster and inexpensive way (Metzker 2009). This fundamental advantage enables broad NGS applications including categorizing human genetic variants (The 1000 Genomes Project Consortium 2010; The 1000 Genomes Project Consortium et al. 2012), identifying functional elements in the genome (Birney et al. 2007), analyzing epigenetic markers (Simonis et al. 2006), studying genetic mutations responsible for cancer (McLendon et al. 2008) and classifying genetic materials gathered from the environment (Tringe et al. 2005).

Importantly, genetic association studies also benefit from NGS technology. A crucial advantage of NGS technology is that it enables discovery of rare variants, which are usually not accessible in array-based studies (Cirulli and Goldstein 2010). While most GWAS findings are common variants, and these variants explain a modest fraction of genetic heritability, it has been proposed that some of the "missing heritability" may be found among rare variants with large effect sizes (Manolio et al. 2009; Cirulli and Goldstein 2010). NGS studies, which do rely on lists of pre-existing variants, provide one means to systematically evaluate the contribution of rare variants to many human traits and diseases.

Although NGS technology is appealing, its cost is a barrier to many experiments. As of today, the rough cost of sequencing one genome is about $2,000 to $3,000 (Mardis 2011; Wetterstrand 2014). To study rare variants in population-based association studies, it is usually necessary to recruit thousands of samples to reach high statistical power (Li and Leal 2009), translating into multi-million dollar costs for each study. This dissertation explores statistical approaches to reduce sequencing costs.

One cost-effective sequence-based experimental design is to utilize public controls to reduce the total number of genomes to be sequenced. The idea of using public controls has been widely applied in epidemiology studies (Shu et al. 1988; Becerra et al. 1990; Wolk et al. 1998). In genetic association studies, a prominent example of using public controls is the Wellcome Trust Case Control Consoritum (WTCCC) (Burton et al. 2007). Their design includes 14,000 cases of seven diseases and 3,000 shared controls from across Britain, and the analysis reveals 24 association signals. In other association studies, it is also not rare to see the use of shared controls (Zaykin and Kozbur 2010; Berthoumieux et al. 2013; Chen et al. 2013).

In genetic association studies, rigorous analysts need to consider the population stratification to avoid spurious association signals. For example, in WTCCC studies, controlled the extent of population stratification problem by excluding non-European samples (Burton et al. 2007). Formally speaking, sample population is commonly treated as a confounding factor (Devlin and Roeder 1999; Pritchard and Rosenberg 1999; Price et al. 2006). To detect the true signals, which are associated with disease, it is important to model the ancestral locations of the samples (Price et al. 2010).

One approach to the population stratification problem is to infer ancestral locations of recruited samples using their genetic data and then adjust their locations in statistical models. Since the early days when microsatellite markers were popular, it is an interesting question to use genetic data to infer ancestry information (Menozzi et al. 1978; Sforza and Bodmer 1999; Rosenberg et al. 2002). Currently, single-nucleotide polymorphism (SNP) data is prevailing and there are a variety of ancestral inference and association analysis methods. In structured association analysis, ancestral locations are classified to subpopulation clusters and subsequent association analysis can be performed using stratified regression models (Pritchard et al. 2000b; Alexander et al. 2009). In Principal Components Analysis (PCA), principal components explaining much of the variation are used as covariates in analyses (Novembre et al. 2008; Novembre and Stephens 2008). Often, using just a few top principal components as covariates in a regression model can control the population stratification problem (Price et al. 2006).

Another approach to the population stratification problem is to model individual relationships. For example, in linear mixed model, the population is a random effect. This approach includes EMMA (Kang et al. 2008) and FASTA (fast association score test–based analysis)(Chen and Abecasis 2007). These methods are appealing as they usually not only adjust for population, but also handle cryptic family structure. As linear mixed model are computationally intensive, there are several computational improvements such as EMMAX (Kang et al. 2010), GEMMA(Zhou and Stephens 2012), Fast-LMM (Lippert et al. 2011) and GRAMMAR-Gamma(Svishcheva et al. 2012).

It is worth noting that the above methods rely on genome wide genetic data and often require demanding computational resources. For example, in target-sequence experiments, the majority of sequence reads cover a small fraction of the genome. As genome wide genetic markers are unavailable, these existing methods cannot be directly applied. .

To solve this problem, in Chapter 2, I present an ancestral inference algorithm (jointly developed by Chaolong Wang) and apply it to a targeted-sequencing study of the Age-related Macular Degeneration (AMD). Inferred ancestries are used to match AMD cases to additional controls. This method increases power, especially since the original design has highly unbalanced AMD case and control counts. Using this matched data set, I discover a rare coding variant (K155Q) that is associated with the AMD disease.

In Chapter 3, to improve computation efficiency while maintaining inference accuracy, I develop a likelihood-based ancestral inference method. Although the algorithm described in Chapter 2 provides adequate accuracy in ancestry-aware association analysis, it requires demanding computation resources, which can be prohibitive in large studies. Therefore, I develop an alternative likelihood model-based method. This method has higher computation efficiency as well as higher accuracy. I conduct simulations and use real targeted-sequencing data to demonstrate its advantages.

In Chapter 4, I tackle another challenging aspect of sequence analysis by evaluating the utility of imputation methods that integrate information from external reference panels in the analyses of sequenced samples.

Imputation can infer genotypes at untyped locations and can improve the statistical power in genome wide association scan (Li et al. 2009b; Marchini and Howie 2010). Typical imputation methods will use densely genotyped haplotypes from a reference panel to predict the unknown genotypes in less densely genotyped study samples. For example, as

GWAS chips usually include less than one million markers, researcher can use external reference panels, such as the HapMap Project (Gibbs et al. 2003) or the 1000 Genome Project (The 1000 Genomes Project Consortium 2010), to infer the genotypes that are not designed on GWAS chips. Using imputation to incorporate external information has become a common practice in GWAS studies (Cantor et al. 2010; Nalls et al. 2011).

However, in sequence studies, uses of external information have not been well studied. Unlike genotype array data, missingness in the sequence data is usually due to the lack of enough sequence depth (Li et al. 2011). As the cost of sequencing experiments is not negligible, the low-coverage sequencing design remains attractive, and it is therefore important to understand how to obtain accurate genotypes (Li et al. 2011). In current sequence studies, imputation methods are often applied without using reference panel (The 1000 Genomes Project Consortium 2010; Nielsen et al. 2011). Although Pasaniuc et al. 2012 demonstrated that the imputation methods can be applied on sequence data using external reference panels, but they did not provide details on variant discovery and genotyping accuracy.

Therefore, in Chapter 4, we conduct a quantitative approach to evaluate the relationship between imputation panel and genotype calls in a systematic way, with a focus on the potential benefits of very large reference panels. We expect the reference panel to imply variant sites and thus increase the chance to discover and genotype variants of sequenced individuals. To evaluate this method, we have used coalescent simulation and the 1000 Genome Project data. Our evaluation uses a fixed amount of test samples as a sequence genotyping baseline. Then we repeatedly genotype these samples together with reference panel of increasing size. Both simulation and real data support our findings that a large reference panel can improve genotyping of sequence samples.

Overall, these three methods hint at promising new directions and can be directly applied to genetic research using sequence technology. These methods are highly efficient and therefore can handle very large sequence data sets. Our AMD study discovered that a rare variant with high effect size can increase disease risk. Imputation-based methods utilizing large reference panels can produce more accurate genotype calls through well-designed experiments, saving both time and money.

Chapter 2 Ancestral inference and its application to age-related macular degeneration

2.1 Introduction

Age-related Macular Degeneration (AMD) is a leading cause of blindness among the elderly (Priya et al. 2012; Ratnapriya and Swaroop 2013). Over the past several years, genetic studies (Chen et al. 2010; Fritsche et al. 2013) of common variations have provided many clues about disease etiology. Due to assay limitations, these studies have typically ignored rare variants or examined them only in a small set of candidate regions. Here, we set out a targeted sequencing experiment to fine-map AMD susceptible loci to study the contribution of rare variants to disease risk.

We initially conducted standard association analyses (Purcell et al. 2007; Chen et al. 2010). Although several common variants can be replicated, we did not find new AMD variants with enough statistical evidence. Then we incorporated population controls from the NHGRI exome sequencing project (ESP). When augmenting this external data, we needed to avoid the population stratification problem which can happen if study samples come from heterogeneous populations. Therefore, we developed a statistical method that can efficiently estimate sample ancestries using sequence data such as targeted sequencing data. After that, we used a one-to-one ratio to match case (AMD cases) and control (AMD controls or ESP controls) samples that have the same population backgrounds. Using this matched data set, our analysis revealed a rare coding variant

K155Q in complement 3 gene that associated with the AMD disease and replicated previous disease-associated variants in the targeted regions.

This chapter is organized as follows: in Chapter 2.2, we describe a complete AMD genetic association study using target-sequencing data, which relied on the understanding of human ancestries; in Chapter 2.3, we detail the ancestral inference method and demonstrate the validity of the method using both simulation and AMD target-sequencing data; in Chapter 2.4, we summarize the implications of our ancestry inference method for future next-generation sequence studies.

2.2 Identification of a rare coding variant in complement 3 associated with age-related macular degeneration

TA slightly different version of his section has been published previously (Zhan et al. 2013).

Genetic and environmental factors contribute to age-related macular degeneration (AMD) (Swaroop et al. 2009; Priya et al. 2012), a major cause of vision loss in elderly individuals (Friedman et al. 2004). Pioneering discovery of association of AMD with complement factor H (*CFH* (Edwards et al. 2005; Haines et al. 2005; Klein et al. 2005)) was quickly followed by the identification of additional susceptibility loci including *ARMS2/HTRA1* (Jakobsdottir et al. 2005) and complement genes *C3*, *C2/CFB* and *CFI* (Gold et al. 2006; Maller et al. 2007; Yates et al. 2007; Fagerness et al. 2009). Genome-wide association studies (GWAS) of AMD cases and controls have now revealed common susceptibility variants at ~20 different loci (Arakawa et al. 2011; Fritsche et al. 2013) and begun to uncover specific cellular pathways involved in AMD biology.

While common variants tag the associated genomic region, rare variants coding can provide more specific clues about the underlying disease mechanism (Nejentsev et al. 2009). For example, the rare variant R1210C in the *CFH* gene was recently associated with a large increase in AMD risk using targeted sequencing of rare *CFH* risk haplotypes (Raychaudhuri et al. 2011). The resulting altered protein has decreased binding to C3b, C3d, heparin and endothelial cells (Manuelian et al. 2003; Jozsi et al. 2006; Ferreira et al. 2009). A reduction in *CFH*'s ability to inactivate *C3*, leading to increased cell killing activity by the complement pathway, could contribute to AMD – a much more specific and testable hypothesis about disease mechanism than provided by common *CFH* variants whose mechanistic consequences are unclear.

To systematically identify rare, large-effect variants, we carried out targeted sequencing of eight AMD risk loci identified in GWAS (Chen et al. 2010) (near *CFH*, *ARMS2*, *C3*, *C2*/*CFB*, *CFI*, *CETP*, *LIPC* and *TIMP3*/*SYN3*) and two candidate regions (*LPL* and *ABCA1*) (Table 2-2). We re-sequenced these regions in 3,124 individuals (2,335 cases and 789 controls) recruited in ophthalmology clinics at the University of Michigan and at the University of Pennsylvania and among Age-Related Eye Disease Study (AREDS) participants (Age-Related Eye Disease Study Research Group 2000; Chen et al. 2010). Genomic targets were enriched using a set of 150-bp probes designed by Agilent Technologies, and sequence data was generated on Illumina Genome Analyzer and HiSeq instruments. The ten loci comprised 115,596 nucleotides of protein coding sequence and totaled 2,757,914 nucleotides overall. We designed probes to capture 111,592 nucleotides (96.5% of coding sequence) and 966,607 nucleotides overall (35% of the locus sequence, Table 2-2), generating an average of 123,221,974 mapped bases of on-target sequence per individual (an 127.5x average depth counting bases with quality >20 in reads with mapping quality >30, after duplicate read removal); 98.49% of sites with designed probes were covered at >10x depth. We applied the same variant calling tools and quality control filters similar to those used to analyze NHLBI Exome Sequencing Project data (Tennessen et al. 2012; Fu et al. 2013) (Table 2-6). We identified an average of 1,714 non-reference sites in each sequenced individual. In total, this resulted in 31,527 single nucleotide variants of which 18,956 were not in dbSNP 135. Discovered sites included 834 synonymous variants, 1,380 nonsynonymous variants and 43 nonsense variants. Among 13 samples sequenced in duplicate, genotype concordance was 99.82% (when depth >10x). Among 908 samples previously examined with GWAS arrays (Chen et al. 2010), sequence-based genotypes were 98.99% genotype concordant with array-based calls (again, when depth >10x).

In an initial comparison of AMD cases and controls, no rare coding variants with frequency <1% reached experiment wide significance ($p < 0.05 / 31{,}527 = 1.6\text{x}10^{-6}$, including all discovered variants, or $p < 0.05 / 1{,}423 = 3.5\text{x}10^{-5}$ considering only protein altering variants), although several showed encouraging patterns. For example, rare variant R1210C in the *CFH* gene was observed in 23 of the 2,335 sequenced cases, but in none of the 789 sequenced controls (exact test p=0.0025). Common variants in several loci exhibited strong evidence of association, including in *CFH* (peak variant rs9427642 with case frequency f_{case} = 12%, control frequency f_{control} = 27%, P-value = $2.52\text{x}10^{-48}$), *ARMS2* (rs10490924, f_{case} = 33%, f_{control} = 18%, P-value = $5.48\text{x}10^{-27}$), *C3* (rs2230199, f_{case} = 25%, f_{control} = 17%, P-value = $3.94\text{x}10^{-9}$), *C2/CFB* (rs556679, f_{case} = 7%, f_{control} = 12%, P-value = $1.32\text{x}10^{-10}$).

A key requirement for establishing significance of rare disease associated variants is the availability of sufficient numbers of control samples. To increase power, we sought to identify additional controls and focused on samples from the NHLBI Exome Sequencing Project (ESP) (Fu et al. 2013), which sequenced 15,336 genes across 6,515 individuals. Sequence data for our samples and the NHLBI Exome Sequencing Project samples were analyzed with the same analysis pipeline, which minimized potential differences due to heterogeneity in analysis tools and parameters. To further avoid sequencing and variant calling artifacts, we restricted our analysis to sites within regions targeted in both sequencing experiments, genotyped and covered with >10 reads in >90% of the samples examined in each project, and >5-bp away from insertion/deletion polymorphisms catalogued by the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2012). Since careful matching of genetic ancestry is critical for rare variant association studies (Mathieson and McVean 2012; The 1000 Genomes Project Consortium et al. 2012), we selected an ancestry-matched subset of our samples and of samples from the NHLBI Exome Sequencing Project. We used principal components analysis to construct a genetic ancestry map of the world with samples from the Human Genome Diversity Project, each genotyped at 632,958 SNPs (Novembre et al. 2008; Wang et al. 2014). If GWAS array genotypes were available for our samples and for the NHLBI Exome Sequencing Project samples, it would be straightforward to place them directly in this genetic ancestry map. Using targeted sequence data, however, the analysis is more challenging: targeted regions include too few variants to accurately represent global ancestry and off-target regions are covered too poorly, precluding estimation of the accurate genotypes needed for standard principal components analysis. Thus, we developed a new algorithm (described in **Companion Manuscript**(Wang et al. 2014)) to place each sequenced sample in a pre-defined genetic ancestry map of the world. The method can accurately place individuals on this worldwide ancestry map with <0.05X average coverage of the genome and is thus ideal for targeted sequence data, such as ours and the NHLBI Exome Sequence data, which have average off-target coverage of ~0.23x and ~0.90x, respectively (see Figure 2-5**,** which shows that PCA coordinates inferred using 0.10x genome coverage or using GWAS array genotypes are highly similar). We focused on samples where PCA coordinates could be estimated confidently (Procrustes similarity larger than 0.95; see Appendix) and used a greedy algorithm to match cases and controls based on estimated genetic ancestry. As shown in the Appendix, alternative matching algorithms do not alter our conclusions. After matching, we focused on a set of 2,268 AMD cases and 2,268 ancestry-matched controls, matched one-to-one. Since AMD phenotype information was not available for most controls, we expect that a small proportion may eventually develop disease; however, this should not impact power substantially (Wellcome Trust Case Control Consortium 2007). After matching case-control samples, we excluded 1 variant with Hardy-Weinberg Equilibrium test p-value $<10^{-6}$ and focused our analysis on 430 remaining nonsynonymous variants.

In this expanded analysis (see Table 2-1), common variant signals at all loci increased in significance (in comparison to Table 2-4). In addition, two rare coding variants exhibited association with $p < 0.01$. The first was R1210C in the *CFH* gene (observed in one control and 23 cases, OR = 23.11, $p_{exact} = 2.9\times10^{-6}$), providing strong support for the original report (Raychaudhuri et al. 2011). The second variant was K155Q in the *C3* gene (18 controls, 48 cases, OR = 2.68, $p_{exact} = 2.7\times10^{-4}$). When controlling for a previously described common variant signal nearby, rs2230199 (f_{control} = 20.63%, f_{case} = 25.26%, marginal $p_{\text{exact}} = 1.8\times10^{-7}$, OR = 1.31), the evidence for association with K155Q increased slightly (conditional OR = 2.91, $p_{\text{exact}} = 2.8\times10^{-5}$). Inspection of the raw read data shows the variant is well supported and is unlikely to be a sequencing or alignment artifact (see Figure 2-6 and Figure 2-7).

Further strong evidence for association of this variant with macular degeneration is provided in independent work by deCODE Genetics, examining 1,143 Icelandic macular degeneration cases and 51,435 Icelandic controls (control frequency 0.55%, OR = 3.45, $p = 1.1\text{x}10^{-7}$; companion manuscript by Drs. Unnur Thorsteindottir and Kari Stefansson submitted simultaneously(Helgason et al. 2013)). In 1,606 directly genotyped cases of macular degeneration from the Age-Related Eye Disease Study II (The AREDS2 Research Group et al. 2012) the variant has frequency 1.77%, similar to our sequenced AMD cases (frequency 1.10%) and the deCODE AMD cases and is notably higher than in our sequenced controls (0.30%), in NHLBI Exome Sequencing Project participants with primarily European Ancestry (0.40%) and in deCODE controls (0.55%).

We next investigated the potential functional consequences of the K155Q variant. **Figure 1** shows that *CFH* variant R1210C (OR=23.11), *C3* variant K155Q (OR=2.91) and *C3* variant R102G (OR=1.31) all map near the surface where *CFH* and *C3b* interact and can potentially affect binding of complement factor H to *C3b*. Factor H inhibits *C3b* and limits immune responses mediated by the alternate component pathway. The analysis of crystal structures summarized in Figure 1 suggests that K155Q and R102G can affect binding of the first macro-globular domain of *C3* to *CFH* and thus potentially interfere with inactivation of the alternative component pathway (Heurich et al. 2011). Interestingly, the three variants (R102G and K155Q in *C3* and R1210C in *CFH*) all are associated with replacement of a positively charged residue.

In summary, our work and the companion paper identify K155Q as a rare *C3* variant associated with a ~2.91-fold increased risk of macular degeneration. Together with rare *CFH* variant R1210C and previously described common *C3* variant R102G, K155Q may reduce binding of *CFH* to *C3b*, inhibiting the ability of Factor H to inactivate the alternative complement pathway. Our work relied on targeted sequencing of GWAS loci, genetic ancestry matching of our sequenced samples to additional sequenced controls analyzed with the same variant calling and filtering tools, focused analysis of regions deeply sequenced in both our project and previously sequenced controls, and avoidance of common calling artifacts near insertion/deletion polymorphisms. The use of publicly available samples to augment control sets may be useful to many targeted sequencing studies, but the strictness of matching and variant filtering required for preventing false-positive findings due to population stratification and/or sequence analysis artifacts are areas deserving of further study. As the number of sequenced human genome and exomes grows, we expect that the utility of the approach will grow – making it possible to match multiple controls to each case and making it possible to focus on progressively finer ancestry matches. Our results also emphasize that large sample sizes will be required for rare variant studies of complex human traits, and illustrate the promise of these studies to highlight disease biology, as illustrated by the interaction between Factor H and *C3b* suggested as a key factor in AMD biology here.

2.3 Ancestral inference using next generation sequence data

This section describes a joint work with Chaolong Wang and the texts of this section based on (Wang et al. 2014).

2.3.1 Ancestral inference method

We develop the LASER (Locating Ancestry using Sequence Reads) method to infer ancestry of sequence samples. It uses genotype data of external reference panel, which usually includes genome wide markers and includes diverse populations. We use the HGDP data of 938 individuals of 53 worldwide populations in this paper. Our method can utilize targeted sequencing data or exome sequencing data. For example, we can use the AMD case-control targeted sequencing data or the 1000 Genome Project exome sequencing data.

The LASER method consists of four steps (Figure 2-2), including (1) PCA on genotype data, (2) simulation of sequence reads, (3) PCA on sequence data, and (4) Procrustes analysis. Step 1 is only performed once for the reference panel, and steps 2-4 are repeated independently for each study sample until all samples are mapped in the reference PCA space. If PCA coordinates of the reference panel are available, we can directly use the coordinates and skip the first step. Details of each step are described below.

PCA on genotype data. We code reference data consisting of autosomal SNP genotypes for N individuals across L loci as $N \times L$ matrices $\boldsymbol{G}$, in which $\boldsymbol{G}_{ij} = 0, 1, 2$ or missing, representing the number of copies of the reference allele at locus j for individual i. To perform PCA on the genotype data, we first standardize the genotypic matrix $\boldsymbol{G}$ as

$\boldsymbol{Q}_{ij} = (\boldsymbol{G}_{ij} - \mu_j)/\sigma_j$, for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, L$. In this equation, μ_j and σ_j represent the jth column mean and standard deviation of $\boldsymbol{G}$**,** and were calculated using non-missing entries. If locus j was monomorphic (*i.e.*, $\sigma_j = 0$), we set all entries in column j of matrix $\boldsymbol{Q}$ to be 0. Missing entries in $\boldsymbol{G}$ were also set to 0 in $\boldsymbol{Q}$. We then performed eigen decomposition on the $N \times N$ matrix $\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{Q}^T$. The kth PC is given by $\lambda_k^{1/2}\vec{v}_k$, in which λ_k is the kth eigen value of matrix $\boldsymbol{M}$, sorted in decreasing order, and $\vec{v}_k$ is the corresponding eigen vector. We recorded coordinates of the top K PCs for n reference individuals in $N \times K$ matrix $\boldsymbol{Y}$. The PCA method described here is similar to that in Wang *et al.* (2012) (Wang et al. 2012) except for a slight difference in the standardization procedure.

Simulating sequence data for the reference individuals. To analyze sequence data of one study sample together with the genotype data of N reference individuals, we simulated sequence data for the reference individuals to match the coverage pattern of the study sample. This allowed us to estimate eigen vectors appropriate for analysis of low coverage data. An alternative approach is calling genotypes across the whole genome, but it is challenging to obtain accurate genotypes at most low-coverage sites. We denote the simulated sequence data for N reference individuals as two $N \times L$ matrices, $\boldsymbol{C}'$ and $\boldsymbol{S}'$, representing the simulated coverage and the counts of reference alleles at L loci. Given one study sample h, we simulated sequence data for the reference individuals by fixing the simulated coverage as $\boldsymbol{C}'_{ij} = \boldsymbol{C}_{hj}$ (for $i = 1,2, \ldots, N$ and $j = 1, 2, \ldots, L$; thus exactly matching the coverage at the sample being analyzed), and drawing the count of reference alleles at each locus conditional on SNP genotypes and the simulated coverage:

$$\boldsymbol{S}'_{ij}|\boldsymbol{G}_{ij},\boldsymbol{C}'_{ij} \sim \begin{cases} \text{Binomial}\left(\boldsymbol{C}'_{ij},\epsilon\right), \text{if } \boldsymbol{G}_{ij}=0 \\ \text{Binomial}\left(\boldsymbol{C}'_{ij},0.5\right), \text{if } \boldsymbol{G}_{ij}=1 \\ \text{Binomial}\left(\boldsymbol{C}'_{ij},1-\epsilon\right), \text{if } \boldsymbol{G}_{ij}=2 \end{cases} \quad (1)$$

in which, ϵ was the estimated sequencing error rate per base. We set ϵ to 0.01 across all loci unless otherwise noted. If $\boldsymbol{G}_{ij}$ was missing, we accordingly set $\boldsymbol{S}'_{ij}$ to missing.

PCA on combined sequence data. When analyzing sample h, sequence data for each of the N reference individuals is simulated to match $\boldsymbol{C}_{h\cdot}$ (row h in matrix $\boldsymbol{C}$ which is also the coverage profile of sample h) and thus genetic variation in the reference panel is reflected only in matrix $\boldsymbol{S}'$. To perform PCA on the reference individuals together with the study sample, we first stacked matrix $\boldsymbol{S}'$ and the row vector $\boldsymbol{S}_{h\cdot}$ (corresponding to row h in matrix $\boldsymbol{S}$) to produce an $(N+1)\times L$ matrix $\boldsymbol{S}$". When examining off-target regions in a targeted sequencing experiment, most loci have zero reads (*i.e.*, $\boldsymbol{C}_{hj}=0$) and many columns in matrix $\boldsymbol{S}$" are uninformative because all elements equal 0. To reduce computational complexity, we excluded these uninformative columns from $\boldsymbol{S}$" and obtained a $(N+1)\times\tilde{L}$ matrix $\tilde{\boldsymbol{S}}$, in which $\tilde{L}\ll L$. We then performed PCA on matrix $\tilde{\boldsymbol{S}}$ following the procedure described in the section *PCA on genotype data*. We recorded coordinates of the top K PCs for the reference individuals as an $N\times K$ matrix $\boldsymbol{X}$, and for the study sample as a $1\times K$ row vector $\boldsymbol{Z}_{h\cdot}$. This set of PCs from sequence data can be compared with PCs from genotype data as described next.

Procrustes analysis. To place the study sample into the PCA space generated by genotypes of the reference individuals, we applied Procrustes analysis(Schönemann and Carroll 1970; Wang et al. 2010) on $\boldsymbol{X}$ and $\boldsymbol{Y}$ to find an optimal transformation f on $\boldsymbol{X}$ that maximizes the similarity between $f(\boldsymbol{X})$ and $\boldsymbol{Y}$ while preserving the relative pairwise distances among points within $\boldsymbol{X}$. The optimal transformation f can be analytically expressed by equations 2-4 in Wang *et al.* (Wang et al. 2010), including translation, scaling, rotation, and reflection. The similarity between $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be quantified by a Procrustes similarity statistic $t(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{1 - D}$, in which D is the minimum sum of squared Euclidean distances between $f(\boldsymbol{X})$ and $\boldsymbol{Y}$ across all possible transformations. D is given by equation 6 in Wang *et al.* (Wang et al. 2010), which has been scaled to range from 0 to 1. After identifying f based on $\boldsymbol{X}$ and $\boldsymbol{Y}$, we applied the transformation to $\boldsymbol{Z}_{h\cdot}$ such that $\boldsymbol{Z}_{h\cdot}^{*} = f(\boldsymbol{Z}_{h\cdot})$ represents coordinates of the study sample in the reference PCA space $\boldsymbol{Y}$. The Procrustes similarity statistic $t(\boldsymbol{X}, \boldsymbol{Y})$ reflects the accuracy of $\boldsymbol{Z}_{h\cdot}^{*}$; lower similarity between $\boldsymbol{X}$ and $\boldsymbol{Y}$ corresponds to greater uncertainty in $\boldsymbol{X}$ and therefore a less reliable $\boldsymbol{Z}_{h\cdot}^{*}$. LASER reports the Procrustes similarity statistic $t(\boldsymbol{X}, \boldsymbol{Y})$ for each sample so that users can filter out samples expected to have less accurate ancestry estimation.

2.3.2 Simulation

Simulation of Sequence Data for Method Validation

We first simulated the coverage independently across L loci for n_t test individuals from a Poisson distribution: $\boldsymbol{C}_{ij} \sim \text{Poission}(\lambda)$ for $i = 1, 2, \dots, n_t$ and $j = 1, 2, \dots, L$. For the HGDP, $n_t = 238$ and $L = 632{,}958$, and for the POPRES, $n_t = 385$ and $L = 318{,}682$. Then we draw copies of the reference alleles $\boldsymbol{S}_{ij} | \boldsymbol{G}_{ij}, \boldsymbol{C}_{ij}$ using the binomial distributions given by **Equation 1** with a uniform sequencing error rate $\epsilon = 0.01$. For both the test sets of HGDP and POPRES, we simulated multiple datasets of different mean coverage by varying λ from 0.001 to 0.40. The expected number of loci with ≥ 1 reads is calculated as $L_1 = L \times (1 - e^{-\lambda})$ for each simulated dataset (Table 2-5**,** Table 2-6).

We also performed additional simulations to explore how our method can help exome sequencing studies. We used genotypes for 385 POPRES individuals (the same set of samples that were used in previous simulations) as templates to simulate sequence reads across 318,682 loci. Instead of simulating the coverage, we used the empirical coverage patterns in 385 exomes from the NHLBI Exome Sequencing Project (Fu et al. 2013). Among the 318,682 loci, 2,547 loci were targeted and had mean coverage ~88.9X. The remaining 316,135 off-target loci had mean coverage ~1.0X.

Estimating Coordinates based on SNP Genotypes

To evaluate the performance of LASER, we compared the sequence-based coordinates estimated by LASER with coordinates estimated using SNP genotypes. The method to place a sample into an existing PCA space of N reference individuals based on SNP genotypes is similar to our LASER method, except that we do not need to simulate sequence reads for the reference panel. We first combined the genotype data of the study sample and N reference individuals, and then perform PCA on the shared set of SNPs for these $N + 1$ individuals. Finally, we used the same Procrustes analysis approach as described in the LASER method to project the study sample into the reference PCA space. We repeated this procedure for all study samples whose SNP genotypes were available. A very similar approach was previously applied to study ancient DNA samples. Those samples have large amount of missing genotypes but do not have sequence data (Skoglund et al. 2012).

For a set of test samples, we used the squared Pearson correlation r^2 to measure the concordance between the sequence-based coordinates and the SNP-based coordinates along each PC. We also reported an overall similarity measurement between two sets of coordinates using the Procrustes similarity statistic t_0, which was obtained by applying Procrustes analysis to sequence-based coordinates and SNP-based coordinates of the test samples.

Evaluation based on simulation

For the worldwide sample set, we randomly selected 238 individuals from the HGDP (Novembre et al. 2008), and used their array genotypes at 632,958 loci as templates to simulate sequence data. We simulated multiple sequence datasets with mean coverage ranging from 0.001X to 0.25X. The remaining 700 HGDP samples were used to construct the reference PCA space. We examined the first four principal components. These can be used to separate major continental groups in the HGDP (Figure 2-3): PC1 and PC2 separate major continental groups in the Old World (Figure 2-3 **A**), while PC3 and PC4 further separate Native American and Oceanian populations, respectively (Figure 2-3 **B**). We applied LASER to each simulated sequence dataset to estimate the ancestry coordinates of the test individuals in the reference PCA space. We assessed the accuracy by comparing ancestry estimates derived from LASER to PCA coordinates of the test individuals based on their original SNP genotypes using the squared Pearson correlation r^2 along each PC and the Procrustes similarity t_0. Our results show consistently high accuracy across all simulated datasets (Figure 2-3, Table 2-5). When the simulated coverage is 0.001X (corresponding to ~630 loci covered with ≥ 1 reads), r^2 ranges from 0.7396 for PC4 to 0.9506 for PC1 and the Procrustes similarity is $t_0 = 0.9508$. Panels C and D in Figure 2-3 show that although the patterns are a bit fuzzy, major continental groups are well separated at 0.001X coverage. Accuracy increases with coverage; when the coverage is 0.10X, the estimated coordinates are almost identical to coordinates estimated using a GWAS SNP panel with $t_0 = 0.9993$ (Figure 2-3**,** Table 2-5). Thus, our method should be able to reconstruct worldwide ancestry with even very modest amounts of sequence data.

Similarly, for estimates of fine-scale ancestry within Europe, we used genotypes at 318,682 loci and 385 randomly selected POPRES individuals (Novembre et al. 2008) as templates to simulate low coverage sequence data (from 0.01X to 0.40X). The remaining 1,000 POPRES European ancestry samples were used to construct the reference PCA space. We focused on the top two PCs of the POPRES reference panel, which mirror the geographic map of Europe (Novembre et al. 2008) (**Figure 3A**). Compared to the estimates of worldwide continental ancestry, much higher coverage is required to reveal the more subtle differences in population structure within Europe (Figure 2-4, Table 2-6). With an average coverage of 0.01X, samples clump in the center of the reference PCA space (Figure 2-4 **B**, $r^2 = 0.5687$ for PC1 and 0.0108 for PC2, $t_0 = 0.4786$). As coverage increases to 0.05X (Figure 2-4 **C**), we become able to observe population structure along PC1 ($r^2 = 0.8851$), which separates Northern and Southern Europeans, but still no structure along PC2 ($r^2 = 0.2516$). Clear population structure within Europe is revealed when coverage is >0.10X (Figure 2-4 **D-F**), with t_0 increasing from 0.9126 (0.10X coverage) to 0.9764 (0.40X coverage) (Table 2-6). Thus, reconstructing ancestry within Europe requires substantially more data than reconstructing continental ancestry in a worldwide sample.

2.3.3 Infer ancestry of samples in AMD study

We next applied LASER to 3,159 samples sequenced around 8 macular degeneration susceptibility loci and two candidate regions (Zhan et al. 2013). The samples include 2,362 macular degeneration cases, 789 controls, two samples with unknown phenotype, and one European (CEU) and one Yoruba (YRI) nuclear family selected among the HapMap Project samples (each nuclear family included mother, father and a child). Macular degeneration cases and controls were recruited in Ophthalmology clinics across the United States. In these samples, off-target coverage was 0.224X across the 633K loci in HGDP, and 0.241X across the 319K loci in POPRES. When using the HGDP as the reference panel, the two trios were placed to the correct positions: the CEU trio clustered with the HGDP Europeans, and the YRI trio clustered with the HGDP Africans. Diverse ancestral background was observed among the 3,153 case-control samples: 3,069 clustered with Europeans/Middle Eastern ancestry individuals; 73 aligned between Africans and Europeans (likely corresponding to African American samples); 5 aligned between Europeans and Native Americans; 3 clustered with Central/South Asians; and 3 clustered with East Asians (Figure 2-7 **A-B**). We then used the POPRES reference panel to dissect the population structure among samples in the cluster with European/Middle Eastern ancestry. Our results show that although most of these samples had northern European ancestry, many other samples formed a small cluster around southern Europe (Figure 2-8 **C-D**). For 931 of the sequenced AMD cases and controls, GWAS array genotype data are also available (Chen et al. 2010). For these samples, results based on the off-target reads match well with the coordinates estimated using SNP genotypes, in both the HGDP PCA space ($t_0 = 0.9068$, Figure 2-9) and the POPRES PCA space ($t_0 = 0.9209$, Figure 2-10). The accuracy increased for samples with higher off-target coverage (Table 2-7).

2.4 Conclusion and discussion

This chapter presents a complete AMD target-sequence study. Our work has augmented with publicly available control data, replicated a previously reported rare variant association in the CFH gene and identified a new rare variant signal in the C3 gene. In both instances, these rare variants are associated with substantially larger odds ratios than common variants in the same regions. We hope the finding can improve our understanding of the AMD disease etiology.

The methodology contribution is that we propose a new method that allows the genetic ancestry of an individual to be accurately estimated in targeted sequencing experiments. This method skips genotype calling and directly analyzes sequence reads to place samples into a principal component ancestry map constructed using genotypes for a reference set of individuals. After careful validation, we show that the method can accurately infer worldwide continental ancestry with whole genome shotgun coverage as low as 0.001X. For estimates of fine-scale ancestry within Europe, the method performs well with coverage of 0.1X, as illustrated by the analysis of samples from a targeted sequencing study of age-related macular degeneration.

2.5 Appendix

Study samples

Macular degeneration cases and controls were recruited at Ophthalmology clinics at the University of Michigan and the University of Pennsylvania and through the Age Related Eye Diseases Study, as previously described. All participants provided informed consent.

Sequence Production and Quality Control

Illumina multiplexed libraries were constructed with 5µg of whole genome amplified material or 1µg native genomic DNA according to the manufacturer's protocol (Illumina Inc, San Diego, CA) with the following modifications: 1) DNA was fragmented using a Covaris E220 DNA Sonicator (Covaris, Inc. Woburn, MA) to range in size between 100 and 400bp. 2) Illumina adapter-ligated library fragments were amplified in four 50µL PCR reactions for eighteen cycles. 3) Solid Phase Reversible Immobilization (SPRI) bead cleanup was used for enzymatic purification throughout the library process, as well as final library size selection targeting 300-500bp fragments. Ninety percent of the samples for this project had four multiplexed libraries pooled together while the remaining ten percent had twenty-four multiplexed libraries pooled prior to hybridization. A custom targeted probe set of 150bp probes was designed (Agilent Technologies, Santa Clara, CA) and captured 0.97 Mb of sequence from 10 loci (spanning 2.76Mb and including 57 genes). The concentration of each captured library pool was accurately determined through qPCR according to the manufacturer's protocol

(Kapa Biosystems, Inc, Woburn, MA) to produce cluster counts appropriate for the Illumina GAIIx and HiSeq 2000 platforms. Pools of four samples were loaded on the GAIIx and pools of twenty four samples were loaded on the HiSeq. With the throughput of both instruments, one lane of 2 x 101bp paired end sequence data generated approximately 1.7 GB per sample. This data quantity covered 80% of the targeted space at a depth of 20x or higher. Reads from each sample were aligned to the NCBI37/hg19 reference sequence using BWA. For approximately forty percent of the sample set where pre-existing genotype information was available, sample identity was confirmed by comparing sequence data (SAMtools consensus calls) with pre-existing genotype array data and only samples with >90% concordance were carried forward to analysis.

Quality control and variant calling

Sequence data was generated for 3159 samples. Quality control steps for all BAM files included: removal of duplicated reads using MarkDuplicates in Picard 1.42; recalibration of base qualities using GATK(McKenna et al. 2010) 1.0.5974; diagnostic graphs and evaluation of sequencing quality using QPLOT (Li et al. 2013); check for contamination using verifyBamId (Jun et al. 2012). After removing samples with high contamination, unexpected relatedness or with high discordance rate, we analyzed 3124 samples retaining 2,335 cases and 789 controls for an initial round of analysis.

We calculated the sequencing depth using high quality reads (BWA mapping quality >30) and bases (recalibrated base quality >20). Across the 966,607 base pair target region, we retained an average 123,221,974 bases per individual. That equates to

127.5x average coverage. Within targeted regions, 98.49% of the protein coding exons had coverage >10x.

We performed variant calling step using UMAKE, which was also used for variant calling in the NHLBI Exome Sequencing Project (Fu et al. 2013). Genotype calling and polymorphism discovery was attempted across an "extended target region" (original target +/- 50 base pairs). To remove low quality variants, we excluded: 1) sites with average depth <0.5 or >500; 2) sites with evidence of strand bias or cycle bias; 3) sites within 5 base pairs of a 1000 Genomes Project indel; 4) sites with excess heterozygosity. These filters excluded 15,219 low quality variants. The transition-transversion ratio (Ts/Tv) for the remaining 31,527 site was 2.10. In addition, we examined concordance rates between sequenced-based genotypes for 13 duplicated samples, which was 99.82% when depth >10x. The concordance rate between array-based genotypes and sequence-based genotypes for 908 samples overlapping with our GWAS experiment was 98.99% when depth >10x.

In this sequencing experience, 59.8% of discovered variants are novel (not previously reported in dbSNP 135 or 1000 Genomes Project). The variant call set included 834 synonymous variants, 1380 nonsynonymous variants and 43 nonsense variants overall. On average, each sample carried 40 synonymous variants, 34 nonsynonymous variants and 1 nonsense variants.

Initial analyses

Using 2,335 cases and 789 controls, we first performed single variant tests on 41,202 variants using Fisher's exact test. This analysis confirmed strong association for common variants near *CFH*, *C2*, *ARMS2* and *C3* genes (as expected). An initial examination of rare variants suggested some effects were shadows of common variants with larger effects, so we focused on the rare variant with MAF less than 1% where association remained significant after accounting for nearby common variants using sequential logistic regression (Chen et al. 2010). The marginal association for each variant was evaluated using Fisher's exact test and conditional signals were evaluated by exact logistic regression (Cox and Shell 1970; Hirji et al. 1987). This analysis identified three coding variants with conditional exact P-value less than 0.01 (all also had marginal p-values < 0.01).

Augmenting population controls

To increase power, we identify appropriate ancestry matched controls for our samples among samples sequenced in the ESP project. Major challenges in this analysis are (a) comparing our data set and ESP data set, which were sequenced using different protocols, and (b) ensuring that the genetic background of cases and controls was appropriately matched.

To avoid the potential effects of population stratification, we matched case and control samples based on genetic ancestry. First, we use genome-wide reads to infer sample ancestries on a worldwide population map [summarized in four principal components of ancestry]. The process for placing samples on this worldwide ancestry map is described in detail in a companion manuscript (Wang et al. 2014). Briefly, we first generate a genetic ancestry PCA space using genotyped reference samples (such as those from the Human Genome Diversity Panel). Then, we generate a series of sample specific genetic ancestry PCA that are calibrated to the exact sequencing depth and coverage pattern of each sample and include the reference samples together with a single sequenced sample. Finally, we transform these sample-specific PCA coordinates to the original map using Procrustes analysis. This procedure generates a metric (the Procrustes similarity) that summarizes similarity of reference sample placements using array genotypes to placements using sequence data and we only considered samples where this metric was >0.95 as candidates for matching. Second, we used a procedure inspired on propensity score matching to pair cases and controls (Rosenbaum and Rubin 1983). Briefly, this procedure uses logistic regression to predict the probability that an individual is a case using the four principal components of ancestry as predictors and disease status as the outcome. This estimated probability of being a case for each sample is a propensity score and can be used to match cases and controls. For matching, we used a greedy algorithm to match cases and controls; allowing matches when the respective propensity scores differed by <.0001. An alternative matching algorithm that matched cases and controls mapping close together in principal component space according to the Euclidean distance between them gave similar results (association at K155Q had OR=2.68, exact p-value $4.5x10^{-5}$ using Fisher's exact test).

To avoid variant calling artifacts, we applied very stringent filters to both the AMD study and ESP study call sets. For both studies, we examined only sites with call rates >90%, Phred-scaled variant quality scores >30, passing all study specific quality control filters, with depth >10x for >90% of the samples in the AMD or ESP callsets, and >5-bp from a 1000 Genomes Project indel. In the combined variant call set, we examined 1,148 variants including 280 synonymous variants, 416 nonsynonymous variants and 10 nonsense variants.

Analyses using the combined AMD and ESP data set

After matching, 2,268 cases were matched to 2,268 controls and 1,148 variants [of which 708 were coding variants] were available for analysis. Similar to our initial analysis, we first applied Fisher's exact test for association to all variants. With this larger control set, common variants in LPL and CETP genes reached significance in addition to the common variant signals in our initial analysis. Next, we examined variants with frequency <1% for which signal remained significant after adjusting for the K155Q signal. This analysis highlighted R1210C in *CFH* and K155Q in *C3* (**Figure 1**).

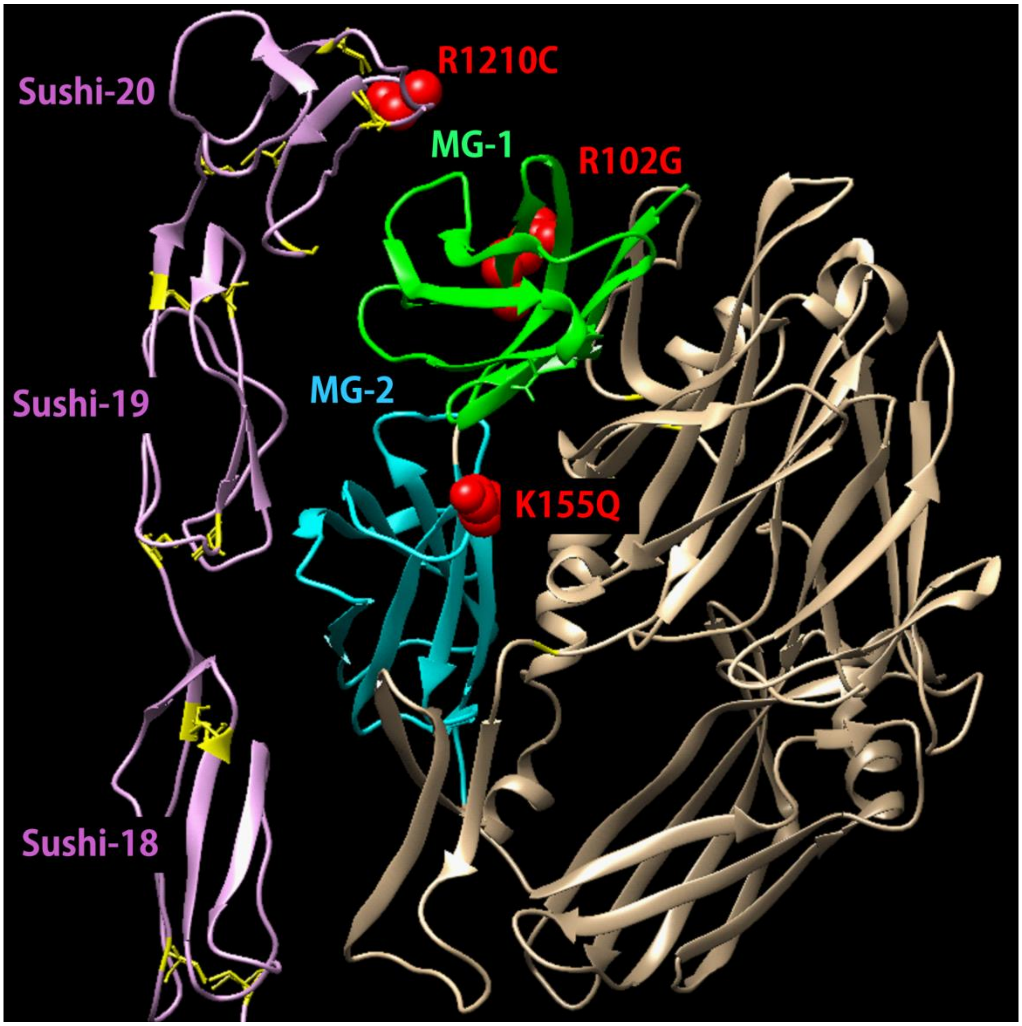
Additional Variant Quality Checks for K155Q

As in Figure 2-6, we examined read depth at the K155Q variant, in comparison to other sites and across individuals. Average depth of coverage at K155Q in the ESP and AMD studies was 90.5x and 63.7x respectively (Figure 2-6 **A, B**). The average depth at samples with heterozygous genotypes at K155Q was 87.9x and 64.1x respectively (Figure 2-6 **C, D**).

To avoid possible alignment artifacts, we inspected raw read alignments in the region for a subset of samples using hapviz (Garrison). As shown in Figure 2-7, there are no patterns indicative of alignment artifacts (excess of reads with mismatches or hard to align insertion deletion polymorphisms) and reads representing the reference and alternative allele are roughly balanced at the variant site.

2.6 Figures and tables

Figure 2-1 R102G and K155Q variants



C3 variants R102G and K155Q and CFH variant R1210C are in the interaction domains the first alpha-macro-globular domain of C3b and CFH, respectively. The fragment of the crystal structure of the four Sushi domains (purple in figure, one not shown for clarity) of CFH in a complex with complement fragment C3b (PDB file: 2wii) was used to explore the effect of disease associated nonsynonymous changes. The CFH residues 987-1230 were used to generate the structure using the first four Sushi domains from 2wii as a structural template (shown in pink, with cysteine residue side chains in yellow). The C-terminal Sushi domains were docked to the binding site in C3b. The first two alpha-macro-globulin domains of C3b, MG-1 and MG-2, are shown in green and cyan, respectively. The location of mutations R102G, K155Q, and R1210C is marked in red.

Figure 2-2 Graphical illustration of the LASER method.

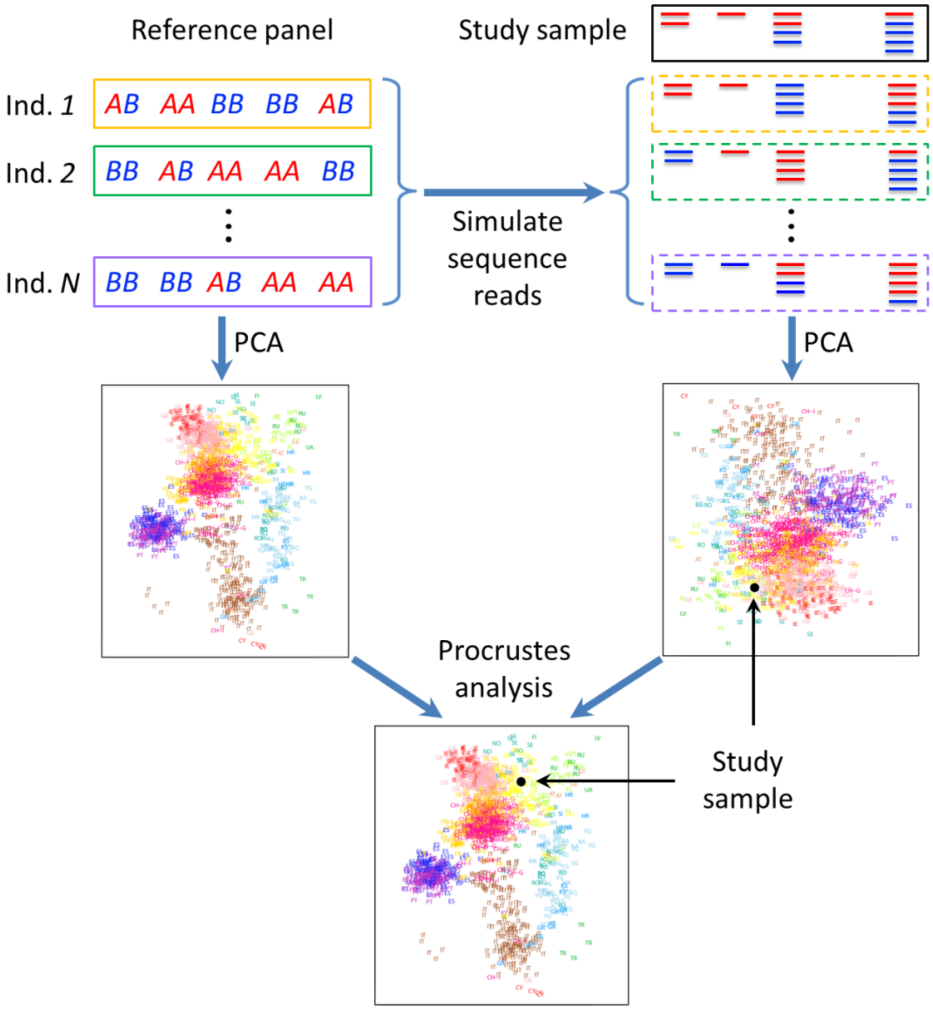


Figure 2-3 Estimation of worldwide continental ancestry

238 individuals were randomly selected from the HGDP as the testing set (colored symbols), and the remaining 700 HGDP individuals were used as the reference panel (gray symbols). (A-B) Results based on SNP genotypes. (C-D) Results based on simulated sequence data at 0.001X coverage. The Procrustes similarity to the SNP-based coordinates is $t_0 = 0.9508$. (E-F) Results based on simulated sequence data at 0.01X coverage ($t_0 = 0.9949$). (G-H) Results based on simulated sequence data at 0.10X coverage ($t_0 = 0.9993$).

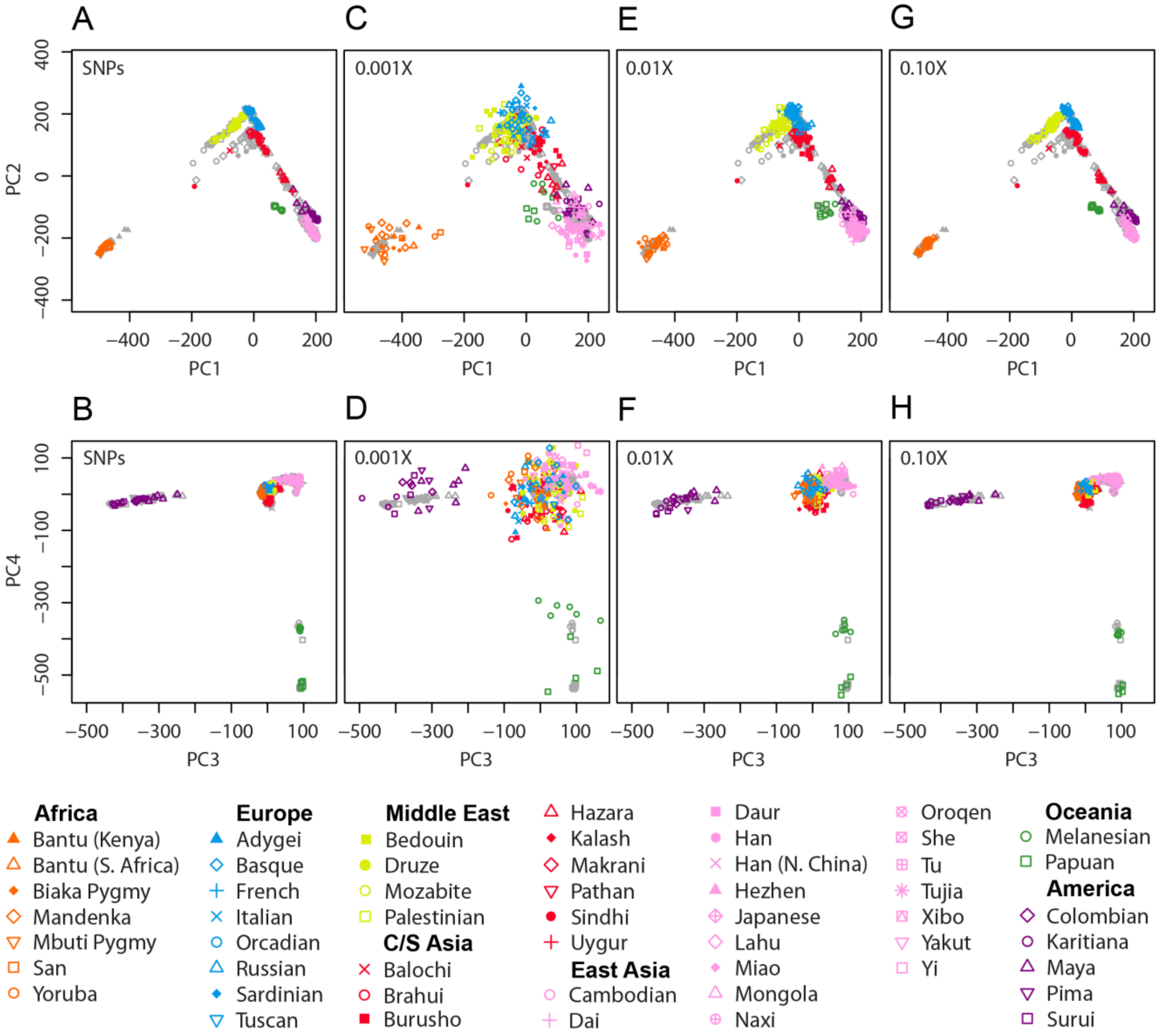


Figure 2-4 Estimation of ancestry within Europe

385 individuals were randomly selected from the POPRES as the testing set (colored symbols), and the remaining 1,000 POPRES individuals were used as the reference panel (gray symbols). (A) Results based on SNP genotypes. (B) Results based on simulated sequence data at 0.01X coverage. The Procrustes similarity to the SNP-based coordinates is $\mathrm{t}_0 = 0.4786$. (C) Results based on simulated sequence data at 0.05X coverage ($\mathrm{t}_0 = 0.7720$). (D) Results based on simulated sequence data at 0.10X coverage ($\mathrm{t}_0 = 0.9137$). (E) Results based on simulated sequence data at 0.20X coverage ($\mathrm{t}_0 = 0.9495$). (F) Results based on simulated sequence data at 0.40X coverage ($\mathrm{t}_0 = 0.9764$). Population labels follow the color scheme of Novembre et al. (2008) (Novembre et al. 2008). Abbreviations are as follows: AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH-F, Swiss-French; CH-G, Swiss-German; CH-I, Swiss-Italian; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NL, Netherlands; NO, Norway; PL, Poland; PT, Portugal; RO, Romania; RU, Russia; Sct, Scotland; SE, Sweden; SI, Slovenia; TR, Turkey; UA, Ukraine; YG, Serbia and Montenegro.

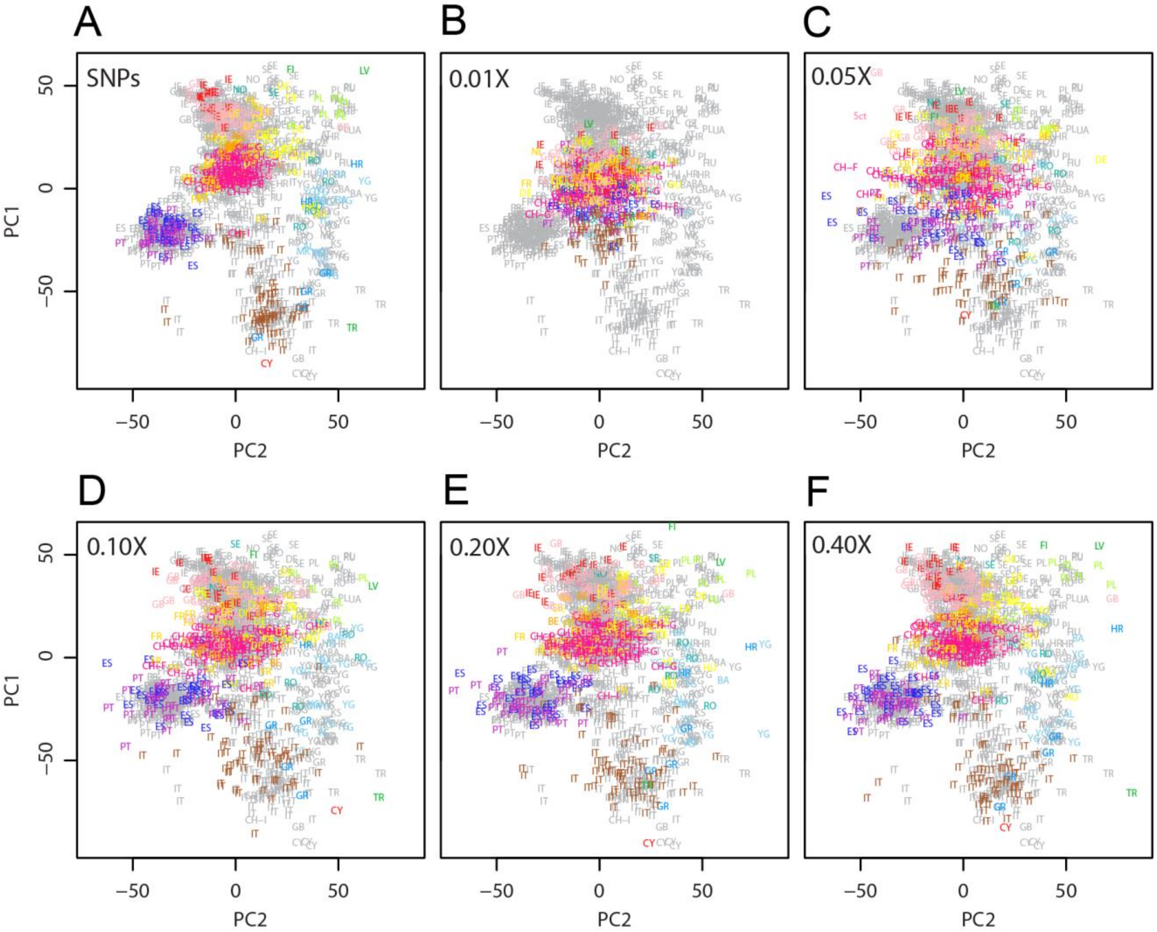


Table 2-1 Summary association ressults for 2,268 sequenced AMD cases and 2,268 sequenced controls

						Frequency (alt allele)				
SNP	**Chromosome**	**Position(bp)**	**Nearest Gene**	**Consequence**	**Alleles (ref/alt)**	**Cases**	**Controls**	**OR**	**P-value**	**Conditional P-value**
Common variant hits										
rs200244837	1	196,884,290	*CFH*	Intron:*CFHR4*	T/A	0.020	0.109	0.17	6.7 x10^{-73}	
rs6467	6	32,006,858	*C2*	Intron:*CYP21A2*	C/A	0.760	0.637	1.81	1.2 x10^{-37}	
rs255	8	19,811,901	LPL	Intron:LPL	T/C	0.151	0.088	1.83	3.6 x10^{-20}	
rs45519541	10	124,183,691	ARMS2	Intron:PLEKHA1	T/C	0.145	0.037	4.41	1.5 x10^{-75}	
rs11076176	16	57,007,446	CETP	Intron:CETP	T/G	0.145	0.177	0.79	4.4 x10^{-5}	
rs2230199	19	6,718,387	C3	R102G	G/C	0.253	0.206	1.30	1.8x10^{-7}	
Rare variant hits MAF < 1% Conditional P <.01 Sort by P										
rs121913059	1	196,716,375	*CFH*	R1210C	C/T	0.005	0.000	23.11	2.9x10^{-6}	6.0x10^{-4}
rs147859257	19	6,718,146	*C3*	K155Q	T/G	0.011	0.004	2.68	2.7x10^{-4}	2.8 x10^{-5}

All p-values were calculated using exact logistic regression. For rare variants, we re-evaluated statistical significance after adjusting for the top common variant in the locus, to avoid shadow signals driven by linkage disequilibrium.

2.7 Supplementary figures and tables

Figure 2-5 Ancestry based matching using the HGDP reference Panel

We label red color as cases and blue color as controls. We showed AMD samples' ancestry using their genotype data in panel A (PC1 and PC2) and in panel E (PC3 and PC4). Comparatively, we showed the same set of samples' ancestries using off-target sequencing reads in panel B (PC1 and PC2) and in panel E (PC3 and PC4). These four panels illustrated the similar ancestries can be inferred from either genotype data or target sequencing data. After we matched 2,268 cases and 2,268 controls from AMD study and ESP study, we showed their ancestries in Panel C (PC1 and PC2) and Panel G (PC3 and PC4). Cases (red) and controls (blue) are well matched in each graph. Further, we displayed the K155Q variant carriers' ancestries in Panel D (PC1 and PC2) and Panel H (PC3 and PC4). Although the numbers of cases and controls are different, their ancestries are similarly gathered.

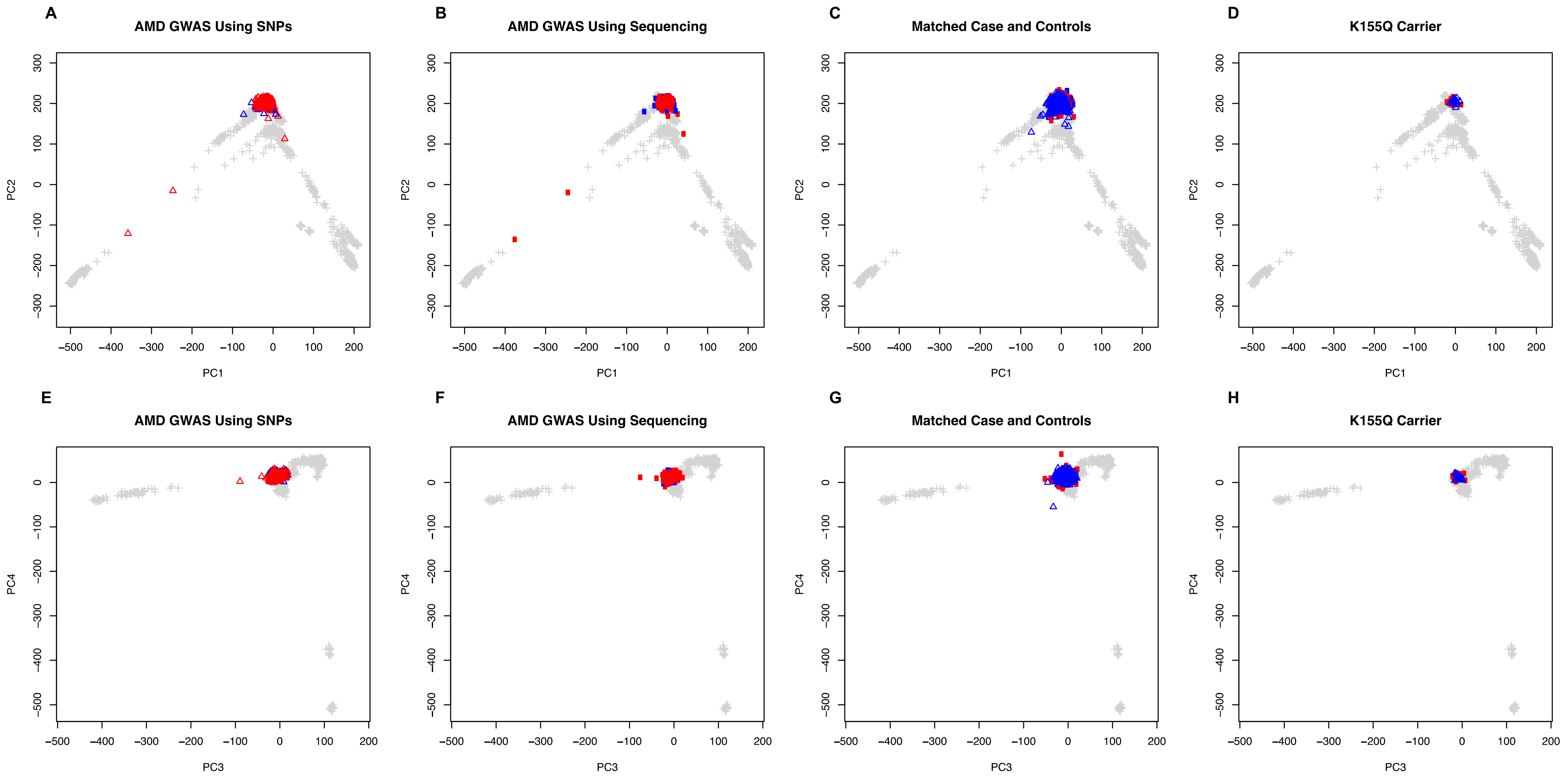

A
AMD GWAS Using SNPs
B
AMD GWAS Using Sequencing
C
Matched Case and Controls
D
K155Q Carrier
PC1
PC2
−500
−400
−300
−200
−100
0
100
200
300
E
AMD GWAS Using SNPs
F
AMD GWAS Using Sequencing
G
Matched Case and Controls
H
K155Q Carrier
PC3
PC4

Figure 2-6 Depth distribution and sequencing alignment diagnostics.

Panel A: Density plot comparing average sequencing depth at K155Q in our targeted sequence data (red line) to that in the 1,148 sites examined in the comparison to ESP (histogram). The average sequencing depth at K155Q was 63.73.

Panel B: Density plot comparing total sequencing depth at K155Q in ESP (red line) to that in the additional 1,148 ESP sites examined in the comparison to ESP (histogram). The average sequencing depth at K155Q was 90.53.

Panel C: Density plot examining depth at K155Q in heterozygote carriers in our targeted sequencing sample. The red line marks average depth (64.11) at K155Q for carriers, the histogram summarizes depth distribution across all genotyped sites.

Panel D: Density plot examining depth at K155Q in heterozygote carriers in the ESP sample. The red line marks average depth (87.93) at K155Q for carriers, the histogram summarizes depth distribution across all genotyped sites.

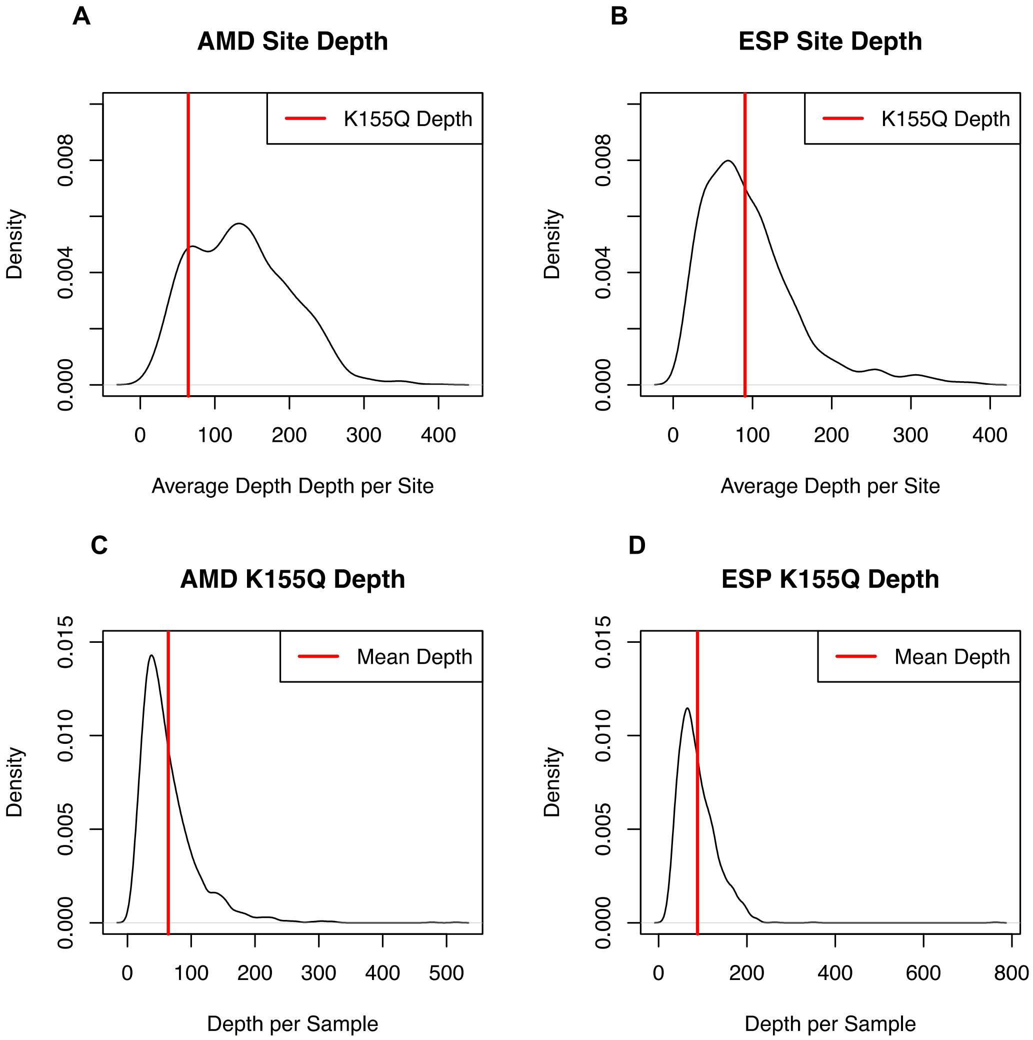

A
AMD Site Depth
K155Q Depth
Density
0.000
0.004
0.008
0
100
200
300
400
Average Depth Depth per Site
B
ESP Site Depth
K155Q Depth
Density
0.000
0.004
0.008
0
100
200
300
400
Average Depth per Site
C
AMD K155Q Depth
Mean Depth
Density
0.000
0.005
0.010
0.015
0
100
200
300
400
500
Depth per Sample
D
ESP K155Q Depth
Mean Depth
Density
0.000
0.005
0.010
0.015
0
200
400
600
800
Depth per Sample

Figure 2-7 Sequencing reads aligned to K155Q variant.

We examined the reads covers the K155Q variants and show that the alignment pattern is clear and that 37 reads overlap this variant. Below the reference genome, all the reads overlapping the site in a predicted K155Q heterozygote are listed. For ease of visualization, we use a '-' (dash) to represent bases that match the reference genome.

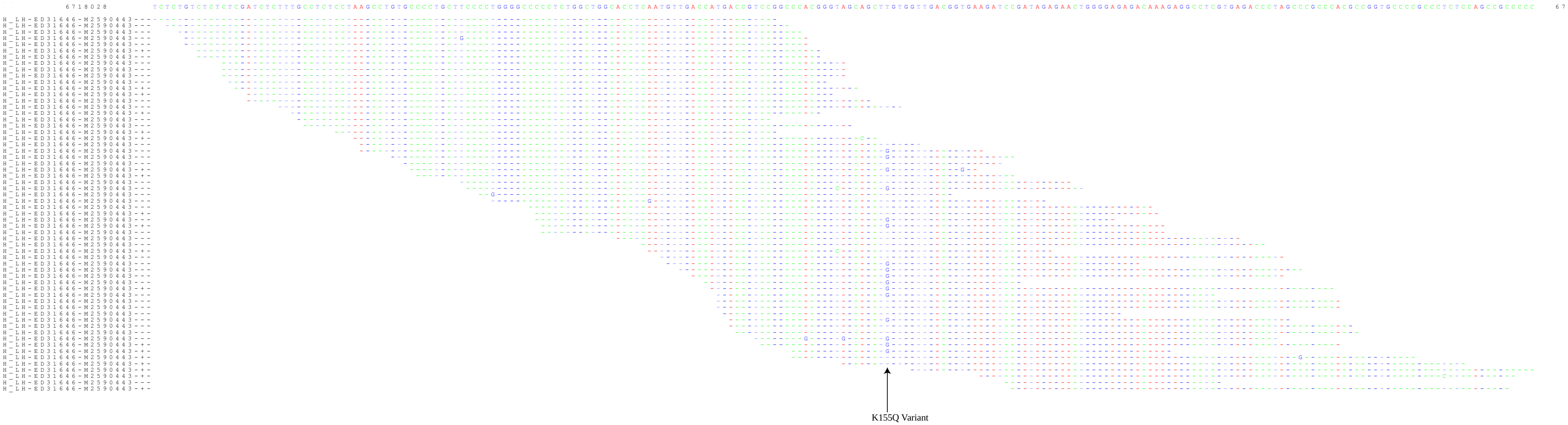


Figure 2-8 Estimation of ancestry for 3,159 samples in the AMD targeted sequencing dataset.

(A-B) Results based on the HGDP reference panel, whose colors and symbols follow Figure 2-3. AMD samples are displayed in black, with different symbols representing possible ancestries based on their estimated PC coordinates. Two HapMap trios are labeled in gray. (C-D) Results based on the POPRES reference panel. Panel C displays PC1 and PC2 of the POPRES; panel D displays 3,072 AMD samples on top of the POPRES. These samples are possibly Europeans or Middle Eastern as indicated in panels A and B. Population labels for the POPRES samples are as follows: AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH-F, Swiss-French; CH-G, Swiss-German; CH-I, Swiss-Italian; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NL, Netherlands; NO, Norway; PL, Poland; PT, Portugal; RO, Romania; RU, Russia; Sct, Scotland; SE, Sweden; SI, Slovenia; TR, Turkey; UA, Ukraine; YG, Serbia and Montenegro.

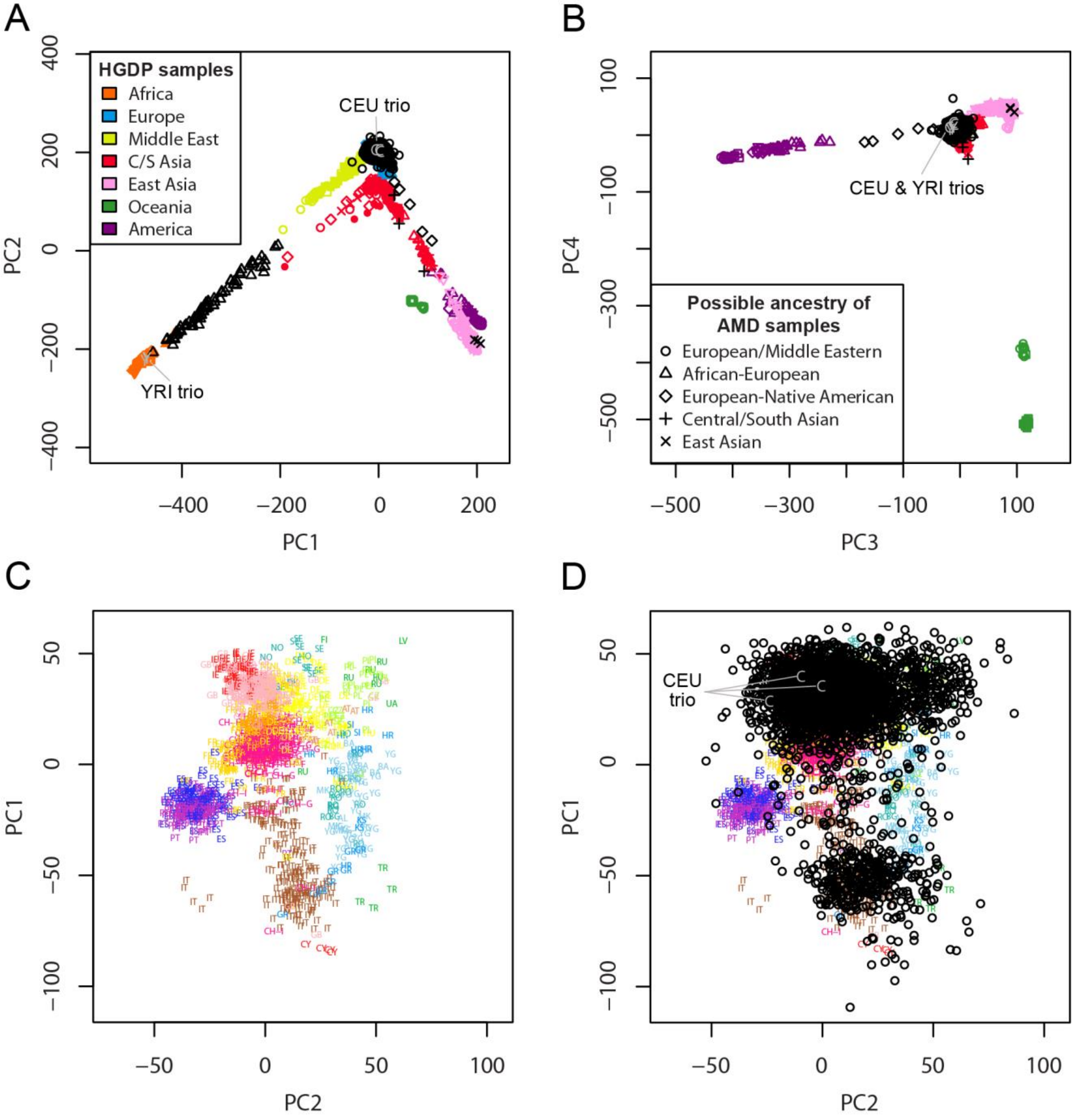

A
HGDP samples
🟧 Africa
🟦 Europe
🟨 Middle East
🟥 C/S Asia
🟪 East Asia
🟩 Oceania
🟪 America
CEU trio
YRI trio
PC2
PC1
400
200
0
−200
−400
−400
−200
0
200
B
Possible ancestry of AMD samples
○ European/Middle Eastern
△ African-European
◇ European-Native American
+ Central/South Asian
× East Asian
CEU & YRI trios
PC4
PC3
100
−100
−300
−500
−500
−300
−100
100
C
PC1
PC2
50
0
−50
−100
−50
0
50
100
LV
FI
SE
NO
RU
UA
HR
YG
BA
SI
AT
CH-I
ES
PT
IT
GR
TR
CY
D
CEU trio
PC1
PC2
50
0
−50
−100
−50
0
50
100

Figure 2-9 Sequence-based coordinates and SNP-based coordinates for 931 AMD samples when using the HGDP reference panel.

Colors and symbols for HGDP and AMD samples follow Figure 2-3, respectively. (A-B) Results based on 45,700 SNPs that are shared by the HGDP, the POPRES, and the AMD SNP datasets. (C-D) Results based on off-target sequence data. The Procrustes similarity between results in panels A-B and in panels C-D is $t_0 = 0.9068$. $r^2 = 0.9104$, 0.8881, 0.6031, and 0.1828 for PC1, PC2, PC3, and PC4, respectively.

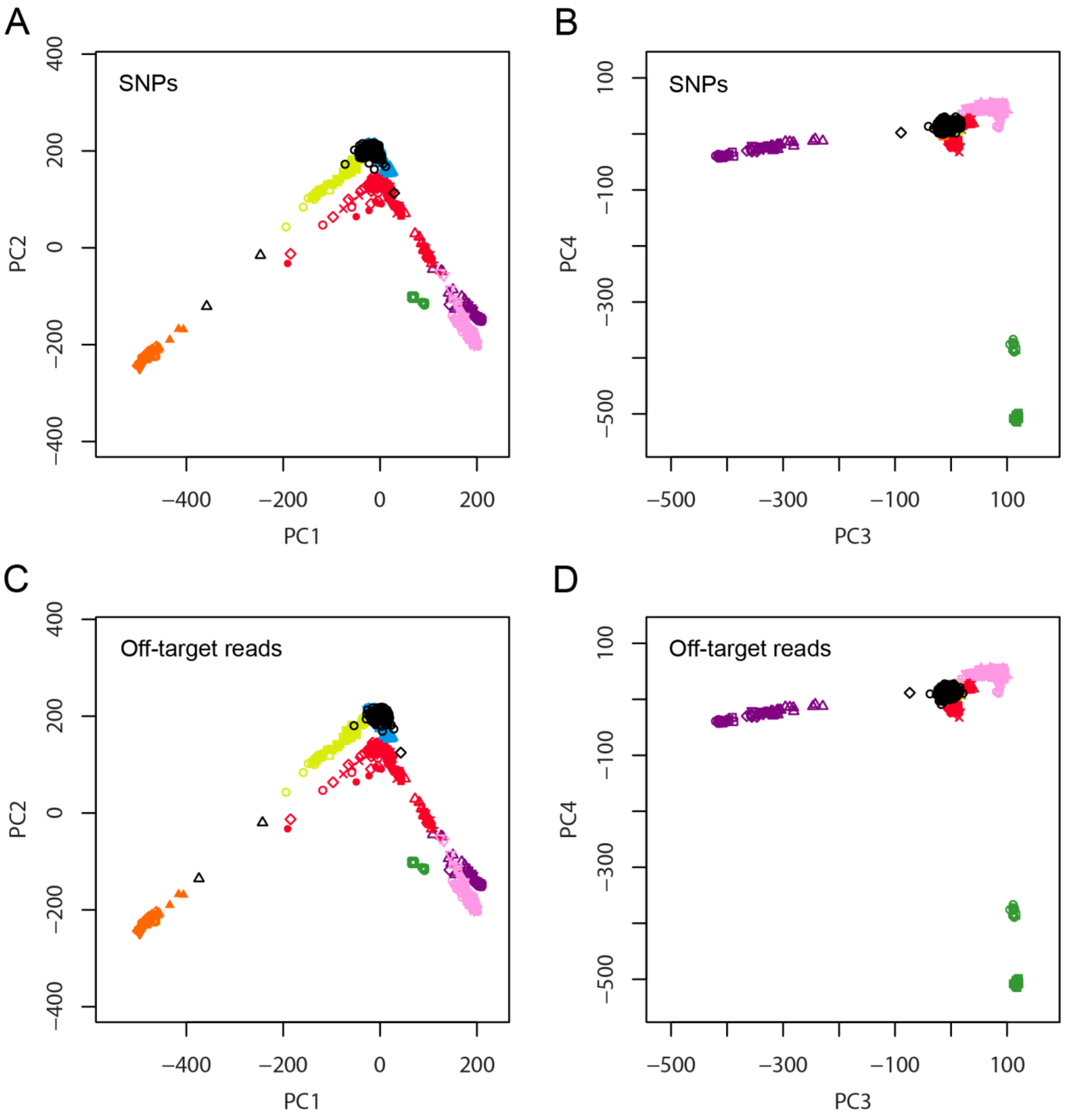

A
SNPs
PC2
400
200
0
−200
−400
PC1
−400
−200
0
200
B
SNPs
PC4
100
−100
−300
−500
PC3
−500
−300
−100
100
C
Off-target reads
PC2
400
200
0
−200
−400
PC1
−400
−200
0
200
D
Off-target reads
PC4
100
−100
−300
−500
PC3
−500
−300
−100
100

Figure 2-10 Sequence-based coordinates and SNP-based coordinates for AMD samples when using the POPRES reference panel.

We only included 928 AMD samples whose genotype data are available and who might be Europeans or Middle Eastern according to results in Figure 2-9. (A) Results based on 45,700 SNPs that are shared by the HGDP, the POPRES, and the AMD SNP datasets. (B) Results based on off-target sequence data. The Procrustes similarity between results in panel A and in panel B is $t_0 = 0.9209$. $r^2 = 0.9557$ and 0.6389 for PC1 and PC2, respectively.

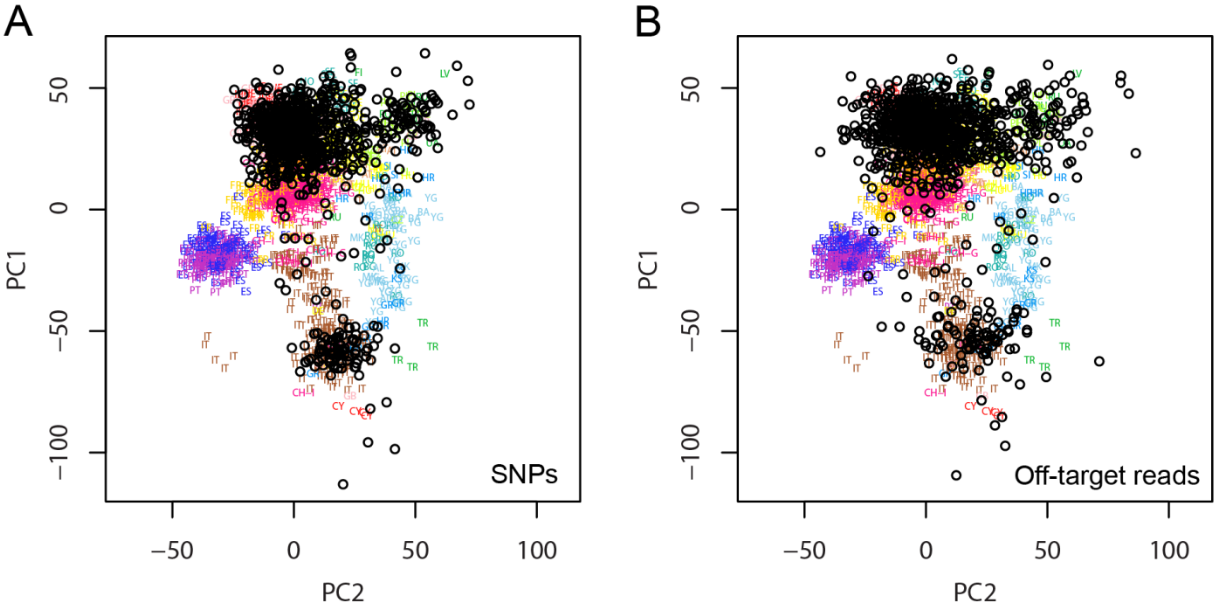


Figure 2-11 Manhattan plot of K155Q association signal for 2,268 sequenced AMD cases and 2,268 sequenced controls.

This plot is produced by LocusZoom (Pruim et al. 2010). The top signal is R102G variant and the second top signal is K155Q variant. The shapes of the legends represents annotations (triangle: nonsynonymous variant; rectangle: synonymous/UTR variant; circle: nothing; filled shapes: variants within LocusZoom database; hollow shapes: variants not in LocusZoom database; complete list can be found from LocusZoom website: http://statgen.sph.umich.edu/locuszoom/).

The data is selected from AMD dataset and ESP dataset using one-to-one propensity-score based matching. Then the association analysis is performed by Fisher's exact test (**Appendix**).

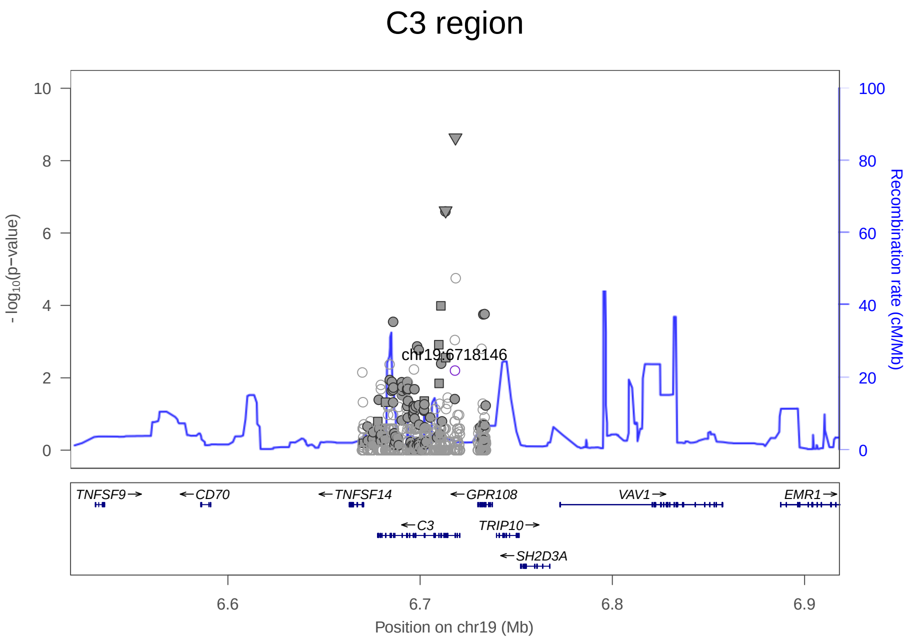


Table 2-2 Target information

Interval					Target Information						
Chr	Start Position	End Position	Length	Protein Coding Bases	# Probes	# Bases	% Interval	Protein Coding Bases	% Protein Coding Bases	Locus Name	# Genes in region
1	196,341,101	196,994,612	653,511	11,359	1,520	226,684	34.69	11,007	96.90	*CFH*	7
4	110,547,457	110,733,347	185,890	4,116	891	132,950	71.52	4,087	99.30	*CFI*	4
6	31,720,915	32,087,186	366,271	66,023	1,393	207,700	56.71	63,090	95.56	*C2/CFB*	29
8	19,786,532	19,938,633	152,101	1,428	737	109,963	72.30	1,418	99.30	*LPL*	1
9	107,533,234	107,700,286	167,052	10,408	860	128,141	76.71	10,341	99.36	*ABCA1*	3
10	124,113,939	124,412,943	299,004	10,432	388	57,812	19.33	10,146	97.26	*ARMS2*	5
15	58,555,986	58,870,773	314,787	1,500	197	29,453	9.36	1,488	99.20	*LIPC*	1
16	56,980,401	57,026,900	46,499	1,482	61	9,089	19.55	1,451	97.91	*CETP*	2
19	6,669,795	6,734,343	64,548	6,469	122	18,178	28.16	6,204	95.90	*C3*	3
22	32,904,490	33,412,741	508,251	2,379	313	46,637	9.18	2,360	99.20	*SYN3/TIMP3*	2
Total			2,757,914	115,596	6,482	966,607	35.05	111,592	96.54		57

Table 2-3 Summary of analyzed variants

Variants were called using UMAKE with standard filters (See supplementary material for details). Comparisons to ESP were restricted to regions targeted in both ESP and our experiment, where depth of coverage >10X for 90% of samples, and >5-bp away from an insertion-deletion polymorphism (as noted in text).

# Fraction of variant sites covered. We showed average values and quartile ranges are shown within parentheses.

* Fraction novel denotes the fractions of variants that not reported in 1000 Genomes Project Phase 1 or dbSNP 135.

	Initial Call Set	Protein Coding Regions	Sites Compared To ESP
Target Summary			
Targeted nucleotides	2,757,914	115,596	-
Examined nucleotides	966,607	111,592	97,196
Mean coverage	106.8	128.6	133.0
Fraction >10x[#]	.95 (.92-.99)	.98 (.98-1.00)	.98 (.98-1.00)
Overall			
SNP			
No. sites	31,527	2,368	1,148
No. in 1000 Genomes Phase I	11,721	750	707
No. in dbSNP 135	12,571	1,017	797
Fraction Novel[*]	59.82%	55.03%	25.78%
No. synonymous	834	834	280
No. nonsynonymous	1,380	1,380	416
No. nonsense	43	43	10
Ts/Tv ratio	2.09	2.88	2.73
Variation Per Sample			
SNP			
No. sites	1,714	78	89
No. in 1000 Genomes Phase I	1,650	75	88
No. in dbSNP 135	1,691	76	87
Fraction Novel[*]	1%	0%	0%
No. synonymous	40	40	24
No. nonsynonymous	34	34	19
No. nonsense	1	1	1

Table 2-4 Initial statistical association analysis of AMD 2,335 cases and 789 controls.

All p-values were calculated using exact logistic regression. For rare variants, we evaluated statistical significance after adjusting for the top common variant at the locus to avoid shadow signals driven by linkage disequilibrium.

						Frequency (alt allele)				
SNP	**Chromosome**	**Position(bp)**	**Nearest Gene**	**Consequence**	**Alleles (ref/alt)**	**Cases**	**Controls**	**OR**	**P-value**	**Conditional P-value**
Common variant hits										
rs1061170	1	196,659,237	CFH	H402Y	C/T	0.481	0.662	0.47	4.48 $\times 10^{-36}$	
rs641153	6	31,914,180	C2	R32Q	G/A	0.060	0.105	0.55	1.26 $\times 10^{-8}$	
rs10490924	10	124,214,448	ARMS2	A69S	G/T	0.326	0.184	2.15	1.85 $\times 10^{-28}$	
rs2230199	19	6,718,387	C3	R102G	G/C	0.247	0.175	1.55	2.31 $\times 10^{-9}$	
Rare variant hits with MAF < 1% and P <.01 (after conditioning on nearby common variants).										
rs121913059	1	196,716,375	CFH	R1210C	C/T	0.005	0.000	∞	2.57 $\times 10^{-3}$	2.00 $\times 10^{-4}$
rs143667999	6	31,922,453	RDBP	D208E	G/C	0.001	0.005	0.21	5.99 $\times 10^{-3}$	6.70 $\times 10^{-3}$
rs147859257	19	6,718,146	C3	K155Q	T/G	0.010	0.003	3.27	6.30 $\times 10^{-3}$	2.50 $\times 10^{-3}$

Table 2-5 Results on simulated worldwide samples with different sequencing coverage.

Sequence data were simulated for 238 individuals randomly selected from the HGDP dataset and the remaining 700 individuals in the HGDP dataset were used as the reference panel. For each simulated dataset, we compared the estimated ancestry coordinates of the 238 testing individuals to their SNP-based coordinates in Figure 2-3.

Simulated mean coverage λ	Expected number of loci with $\geq$ 1 reads	Sequence-based coordinates vs. SNP-based coordinates				
		Squared correlation of PC1	Squared correlation of PC2	Squared correlation of PC3	Squared correlation of PC4	Procrustes similarity t_0
0.25	140,010	0.9996	0.9996	0.9992	0.9988	0.9997
0.20	114,736	0.9996	0.9996	0.9992	0.9986	0.9996
0.15	88,166	0.9994	0.9996	0.9988	0.9978	0.9995
0.10	60,234	0.9992	0.9992	0.9982	0.9974	0.9993
0.05	30,870	0.9988	0.9986	0.9964	0.9946	0.9989
0.01	6,298	0.9948	0.9932	0.9819	0.9716	0.9949
0.008	5,043	0.9940	0.9920	0.9783	0.9663	0.9940
0.006	3,786	0.9896	0.9882	0.9671	0.9586	0.9911
0.004	2,527	0.9894	0.9882	0.9536	0.9347	0.9887

0.002	1,265	0.9756	0.9706	0.8964	0.8356	0.9729
0.001	633	0.9506	0.9388	0.8350	0.7396	0.9508

Table 2-6 Results on simulated European samples with different sequencing coverage.

Sequence data were simulated for 385 individuals randomly selected from the POPRES dataset and the remaining 1000 individuals in the POPRES dataset were used as the reference panel. For each simulated dataset, we compared the estimated ancestry coordinates of the 385 testing individuals to their SNP-based coordinates in Figure 2-4.

Simulated mean coverage λ	Expected number of loci with ≥ 1 reads	Sequence-based coordinates vs. SNP-based coordinates		
		Squared correlation of PC1	Squared correlation of PC2	Procrustes similarity t_0
0.40	105,063	0.9855	0.9078	0.9764
0.35	94,111	0.9866	0.8945	0.9737
0.30	82,597	0.9813	0.8725	0.9671
0.25	70,492	0.9797	0.8540	0.9636
0.20	57,767	0.9738	0.7973	0.9495
0.15	44,390	0.9653	0.7763	0.9428
0.10	30,327	0.9510	0.6647	0.9126
0.05	15,542	0.8851	0.2516	0.7720
0.01	3,171	0.5687	0.0108	0.4786

Table 2-7 Comparison between sequence-based and SNP-based coordinates for a subset of the AMD samples.

This includes 928 samples analyzed with the POPRES reference panel.

Range of coverage per sample	Number of samples	Mean coverage per sample	Average number of loci with ≥ 1 reads	Sequence-based coordinates vs. SNP-based coordinates		
				Squared correlation of PC1	Squared correlation of PC2	Procrustes similarity t_0
$[0.05, 0.20)$	232	0.16	34,114	0.9299	0.5460	0.8770
$[0.20, 0.25)$	232	0.22	45,603	0.9588	0.6655	0.9285
$[0.25, 0.30)$	232	0.27	54,837	0.9616	0.6821	0.9254
$[0.30, 0.79]$	232	0.37	71,102	0.9690	0.6783	0.9480

Chapter 3 More efficient ancestry inference using low-coverage sequencing data

3.1 Introduction

Genetic studies using next generation sequence (NGS) technology are increasingly popular (Mardis 2011). In medical genetics studies, sequencing technology enables the discovery of disease associated variants of very low-frequency and can broaden understanding of disease etiology.

To explore rare disease associated variants, genetic association studies need to sequence large numbers of samples (Li and Leal 2009). Cost constraints mean there is usually a trade-off between the number of samples to be sequenced, sequencing depth and the proportion of the genome examined in each sample (Cirulli and Goldstein 2010). For example, deep whole-genome sequencing studies provide the most complete information per individual but remain an expensive approach (Wetterstrand 2014) and thus typically limit the number of samples that can be examined. A popular alternative, targeted sequencing studies focus sequencing effort on a pre-designated fraction of the genome. These studies can deploy sequencing across 1,000s or 10,000s of samples (O’Roak et al. 2012). Successful examples of targeted sequencing studies include the association of a rare complement factor H variant with age-related macular degeneration (AMD) (Raychaudhuri et al. 2011; Zhan et al. 2013) and of several rare IFIH1 (interferon induced with helicase C domain 1) variants with type-1 diabetes (Nejentsev et al. 2009).

A key question in these studies is the potential to recruit controls from other sequencing studies (increasing power and sample size), while controlling genetic and experimental heterogeneity as well as avoiding potential population stratification. The consequences of population stratification problem have been widely discussed in the context of genome wide association studies (Pritchard and Rosenberg 1999; Cardon and Palmer 2003; Freedman et al. 2004; Campbell et al. 2005; Price et al. 2006; Price et al. 2010). The prevailing strategies for controlling stratification in GWAS include assigning samples to subpopulation clusters using a model based analysis (Pritchard et al. 2000a; Alexander et al. 2009), PCA analysis (Price et al. 2006; Purcell et al. 2007) and analysis using linear mixed models (Yu et al. 2006; Kang et al. 2008; Kang et al. 2010). These strategies might not work well in sequence based studies of rare variants (Mathieson and McVean 2012). Wang et al (2014) suggest an alternative strategy. Their approach, implemented in the software LASER, requires an external reference panel that is used to generate an ancestry map and coordinate space. LASER then uses sequence data to reconstruct ancestry coordinates for each sample and further uses these coordinates to guide one-to-one matching of prospective cases and controls. The strategy appears to work well.

A limitation of the LASER method is that it is computationally demanding and may require multiple stochastic analyses of each sample to estimate the most accurate ancestry coordinates. Here, we introduce a new model-based approach for ancestral inference from sequence data. Extending previous work (Yang et al. 2012), our likelihood-based model of ancestral origin can accommodate low pass or deep sequencing data and model sequencing errors. Our method is based on maximum likelihood and we show it is able to estimate the ancestry of each study sample quickly and accurately. Through simulation and analysis of real sequence data, our method shows superior computational efficiency and accuracy compared to standard PCA analyses and our previous LASER method.

3.2 Material and methods

3.2.1 Workflow

Our ancestry inference model utilizes a three-step workflow (Figure 3-1 Workflow). In the first step, using genotype data on the external reference panel, we calculate reference coordinates. In the second step, we estimate SNP allele frequency gradients using pre-calculated coordinates of reference samples. In the third step, we use the allele frequency gradients and sequence data for each study sample to identify the most likely coordinates for each sample. We define these as the coordinates that result in a set of allele frequencies that maximizes the likelihood of observed sequence data for the sample. The inferred ancestry of each sample can be visualized in the coordinate space constructed using the reference samples.

Calculation of reference coordinates

We calculate reference coordinates from an external reference panel. This panel should genome wide genotypes for diverse populations. We apply Principal Component Analysis (PCA) on the reference genotypes to obtain coordinates for each reference sample (Price et al. 2006). These coordinates form a reference coordinate space where study samples will be in placed in later steps. The calculation of reference coordinates only needs to be performed once.

Given N' reference samples genotyped at M markers, we store genotypes of all reference samples in $N'x\ M$ matrix G' where element g_{ij} denotes the number of reference alleles observed at marker j in sample i. We store the estimated PCA for this matrix, which represent the ancestral origins for each sample, in matrix Z.

Estimation of SNP gradients

SNP allele frequency gradients (Figure 3-2) were suggested by Yang et al (Yang et al. 2012). The gradients represent allele frequencies as a function of ancestral locations and allow us to estimate allele frequencies for any population given its ancestral coordinates. Specifically, the predicted allele frequency of the j^{th} marker as a function of ancestral location x_i is:

$$f_j(x_i) = \frac{1}{\exp\left(-a_j^T\ x_i - b_j\right) + 1}$$

In this equation, a_j is a vector that describes changes in allele frequency as a function of ancestry x_i, and b_j is a number that describes allele frequency of the j^{th} variant at the origin of the current coordinate space. For simplicity, we let A represent all gradients $a_1, a_2, \dots, a_M$ and B represent all intercepts $b_1, b_2, \dots, b_M$.

To estimate SNP gradients, we follow the work by Yang et al. (Yang et al. 2012), and estimate A and B in the following likelihood model:

$$L(A,B;G',Z) \propto \sum_i \sum_j g'_{ij} \ln f_{ij}(z_i) + \left(2 - g'_{ij}\right) \ln\left(1 - f_{ij}(z_i)\right)$$

The above likelihood has been shown to be convex, and therefore SNP gradients A and B can be inferred through an efficient estimation procedure. The implementation is provided in the *spa* program (Yang et al. 2012).

These estimated SNP gradients will be in later analysis steps. Like reference ancestry coordinates, these only need to be calculated once. They can be re-used in the likelihood models for each sample as described in the next step.

<u>The likelihood model for sequence data</u>

We extend the model for low-coverage sequencing data where genotypes are not directly observed. Instead of observing accurate genotype calls (G), in sequencing data we observe sequence reads. We denote R_{ij} as the observed sequence data, and $R_{ij} = (r_{ij}, o_{ij})$ where r_{ij} and o_{ij} represent the number of reference bases and the number of non-reference bases at the jth marker of the ith individual.

The joint likelihood of all samples on all markers can be written as:

$$L(R;X,A,B) \ = \sum_i L_i(R;x_i) = \sum_i \sum_j log\, P\left(R_{ij}\right) = \sum_i \sum_j log \sum_{G_{ij}} P\left(R_{ij} \middle| G_{ij}\right) P\left(G_{ij}\right)$$

(Equation 1)

In this model, $P\left(G_{ij}\right)$ is a function of $f_j(x_i)$. Specifically, assuming Hardy-Weinberg equilibrium:

$$P\left(G_{ij} = (\text{ref}, \text{ref})\right) = f_j^2(x_i)$$

$$P\left(G_{ij} = (\text{ref}, \text{alt})\right) = 2f_j(x_i)\left(1 - f_j(x_i)\right)$$

$$P\left(G_{ij} = (\text{alt}, \text{alt})\right) = \left(1 - f_j(x_i)\right)^2$$

The other term, $P(R_{ij}|G_{ij})$, is called the genotype likelihood (Li et al. 2008; Li et al. 2010) and represents the probability of observed sequence bases as a function of the true genotype. The probability can be approximated as:

$$P(R_{ij} = (r_{ij}, o_{ij})|G_{ij} = (\text{ref}, \text{ref})) = (1 - e)^{r_{ij}} \cdot e^{o_{ij}}$$

$$P(R_{ij} = (r_{ij}, o_{ij})|G_{ij} = (\text{ref}, \text{alt})) = 0.5^{r_{ij}} \cdot 0.5^{o_{ij}} = 0.5^{r_{ij}+o_{ij}}$$

$$P(R_{ij} = (r_{ij}, o_{ij})|G_{ij} = (\text{alt}, \text{alt})) = e^{r_{ij}} \cdot (1 - e)^{o_{ij}}$$

In the above formulae, *e* represents the estimated per-nucleotide base sequencing error. Using current technologies, this error rate is usually between 0.1% and 1% (Bentley et al. 2008; Mardis 2011). Unless noted, we assume this error rate to be 1%. Our simulations show that ancestral inference results are not very sensitive to misspecification of this parameter.

Calculation of maximum likelihood estimator (MLE)

To find the most likely set of ancestral coordinates *X* for study samples, we maximize *L* with respect to ancestry coordinates for each sample x_i (note that our definition of *L* treats each individual independently). We use the simplex method (Figure 3-3) for its simplicity (Nelder and Mead 1965). Since the likelihood surface may have more than one mode, we attempt maximization using 4 different random starts to avoid potential local maxima.

3.2.2 HGDP genotype data set

The Human Genome Diversity Project (Novembre et al. 2008) contains 938 individuals from worldwide populations and has genome wide genotypes. Following previously described quality control steps (Wang et al. 2014), we analyze 632,907 autosomal SNPs with allele frequency >0.

3.2.3 Simulation of sequence reads using the HGDP data

We simulate sequence reads for HGDP individuals. The simulation has two steps: first, we simulate the sequence depth D_{ij} for sample i at jth marker from a Poisson distribution with mean λ. Second, given the sequence depth D_{ij} and true genotypes G_{ij}, we simulate r_{ij}, which is the count of reference bases, from binomial distribution as follows:

$$r_{ij}|D_{ij}, G_{ij} \sim \begin{cases} \text{Binomial}\left(D_{ij}, e\right), \text{if } G_{ij} = (\text{ref}, \text{ref}) \\ \text{Binomial}\left(D_{ij}, 0.5\right), \text{if } G_{ij} = (\text{ref}, \text{alt}) \\ \text{Binomial}\left(D_{ij}, 1 - e\right), \text{if } G_{ij} = (\text{alt}, \text{alt}) \end{cases}$$

The simulation parameters evaluated here include sequencing error rates from 0.1% to 10% (0.1%, 0.2%, 0.5%, 1%, 2%, 5% and 10%); 15 sequencing depths from 0.001X to 50X (0.001X, 0.002X, 0.005X, 0.01X, 0.02X, 0.05X, 0.1X, 0.2X, 0.5X, 1X, 2X, 5X, 10X, 20X and 50X). These combinations cover various types of typical sequence experiments such as targeted sequencing and whole genome sequencing.

3.2.4 AMD targeted-sequencing data set

The AMD target-sequence data set (Zhan et al.) consists of 3,159 samples sequenced around 8 susceptibility AMD loci and two candidate regions. The samples include 2,362 AMD cases and 789 controls. In addition, this data set includes one CEU trio and one

YRI trio. The on-target regions span 0.97 Mbp with >85X depth. Coverage of the 632,907 loci genotyped on HGDP samples averages 0.2X.

To compare the accuracy of our likelihood method with existing methods, we select 920 samples among all target-sequenced samples, as they have been previously enrolled in GWAS studies (Chen et al. 2010). We use the PCA based ancestry coordinates estimated using GWAS genotypes as a gold standard (The coordinates of AMD samples are transform to the HGDP reference coordinate space using Procrustes analysis, see Wang et al. 2012 for details). Then we apply the LASER method (Wang et al. 2014) and our proposed likelihood based method to infer ancestries of each sample and compare the three sets of coordinates.

3.2.5 Evaluation criteria

To evaluate accuracy, we compared ancestry coordinates estimated using sequence data to those estimated using GWAS genotypes. We chose the mean Pearson correlation coefficient r^2 between coordinates calculated by applying PCA analysis to GWAS genotypes and coordinates calculated using our likelihood-based method as accuracy measure of similarity between the two approaches. Suppose the GWAS based PCA coordinates for the ith sample are $y_i = (y_{i,1}, y_{i,2})$ and the coordinates estimated by our likelihood-based method using sequence data are $x_i = (x_{i,1}, x_{i,2})$. We define r_1^2 as the Pearson correlation coefficient between $x_{i,1}$ and $y_{i,2}$, and r_2^2 as the Pearson correlation coefficient between $x_{i,1}$ and $y_{i,2}$,. Then we define the average squared correlation as $r^2 = \frac{r_1^2+r_2^2}{2}$. Higher r^2 thus indicates higher similarity between sequence-based and GWAS-based coordinates.

3.3 Results

3.3.1 Evaluation using simulated sequence data

First, we evaluate the relationship between sequencing depth and the accuracy of estimated coordinates (Table 3-1) using simulated sequence data from HGDP data set. With genome wide depth of 0.001X, the likelihood model already works well (e.g. $r^2 = 0.9573$ when per base sequencing error = 0.001). As the sequence depth increases from 0.001X to 0.2X, the accuracy of the ancestral inference model keeps increasing (to $r^2 = 0.9981$ at 0.1X when error = 0.001). When the sequence depth increases further, this accuracy increases slightly (e.g. $r^2 = 0.9982$ at 0.2X when error = 0.001). This trend shows that the accuracy of our method is stable at a wide range of sequence depths. When the sequence depths are low or moderate, our method performs especially well.

Second, we evaluated the impact of base sequencing errors on our model (Table 3-1). We have simulated data with varying sequencing error, ranging from a high error rate (0.1) to a low error rate (0.001). During analysis, we fixed the estimated error rate at 0.01, regardless of the true simulated error rate. These settings cover the range of sequence error rates that might occur in real data. When the simulated sequencing error rate was highest (0.1), our likelihood based method retains high accuracy (e.g. $r^2 > 0.9982$ under all simulated sequence depths). When the sequence error rate is low (0.001), the accuracy improves. In general, our likelihood method is not sensitive to the true sequence error.

Third, we visually compared the ancestral inference results under a variety of simulated settings (Figure 3-4). We reviewed 9 ancestry maps for 938 HGDP samples, each estimated under different sequence depths (0.1X, 1X or 5X) and/or base sequencing errors (0.001, 0.01 or 0.1). In all cases, continental population groupings can be distinguished. This indicates our method can separate worldwide populations using sequence data of different quality profiles. Still, we note that the higher depths and lower sequence error rates are preferred, as ancestry maps under these settings have more compact population clusters.

Finally, we show the sequence base error parameter specified in the likelihood model can produce robust ancestral inference results (Table 3-1, Table 3-2).We compare the two error rates, 0.01 and 0.001. The accuracy (r^2) is similar under these different error rates. For example, the maximum difference of accuracy is 0.0055 when sequence depth is 0.1X and error rate is 0.1. In other settings, the difference in terms of r^2 is smaller.

3.3.2 Evaluation of AMD target-sequence data

In this evaluation, we use 920 AMD samples to evaluate our likelihood based method compared to the PCA method (gold standard) and the LASER method (Wang et al. 2014). Our method and the LASER method are both designed to support low-coverage sequence data without requiring explicit genotyping of genetic markers. In Figure 3-5, we compare the three methods on the first two principal components. When we compared coordinates to those estimated directly from GWAS genotypes, our method has slightly higher accuracy ($r^2 = 0.8878$) than LASER ($r^2 = 0.8666$). The two methods provide very similar ancestry coordinates (r^2 = 0.9796). In visual comparison, all three methods can identify the same set of samples whose ancestral coordinates are clustered to the top (which is the cluster of the European HGDP samples). This information is necessary to avoid the population stratification problem during the subsequent genetic association studies (Zhan et al. 2013). In addition to quantitative measurement, we notice the inferred origins of some outlier samples (in the bottom right of Figure 3-5) are between the cluster of European samples and African samples, which indicates these sequence samples are likely to have admixture ancestral background. In all, this likelihood based method shows higher concordance compared to existing methods.

3.3.3 Computational cost

In addition to the accuracy, our method has excellent computational advantages. When analyzing sequence data, the likelihood based method implemented in C++ takes 3 seconds to estimate coordinates for each sample and consumes 0.5 GB in memory. In comparison, the LASER C++ implementation requires 126 seconds and 4.1 GB in memory per sample. In addition to sequence data, our program can infer ancestries using genotype data. In our evaluation, the program can estimate the first two principal components of ancestry using 3 seconds per sample and 0.09 GB in memory. A similar analysis using SNPRelate (Zheng et al. 2012), which heavily optimizes matrix algebra procedures, takes about 27 seconds for sample. Our method can perform ancestral inference of 32,266 samples in one day while the SNPRelate software will take about 10 days.

3.4 Conclusion and discussion

This chapter describes an efficient and accurate ancestral inference model using sequence data. The model belongs in the likelihood framework, and it extends the previous work (Yang et al.). Unlike the popular PCA method or the LASER method that are computationally demanding, our method can explicitly incorporate observed sequence data in a joint likelihood model, and the maximum likelihood estimation step has the computational complexity of O(NM). This greatly reduces the computational cost (e.g. PCA method has O(N^3) computational complexity). In addition to its efficiency, our method has good accuracy. Through extensive simulations and real data from a large-scale genetic study, we have demonstrated the inferred ancestral origins are precise.

Our method calculates ancestral locations in the likelihood maximization step. To obtain numerically reliable results, we applied the simplex method with random start points. In practice, we found this approach is insensitive to the start points and thus generates robust results. It is worth noting that there are other optimization procedures such as the BGFS method and L-BGFS method (Broyden 1970; Fletcher 1970; Kelley 1999). They use derivatives of the likelihood function to speed up convergence speed during optimizations. These methods may be faster. However, since the likelihood is not guaranteed to be convex, similar to simplex method, these approaches can be trapped in local maxima (Shanno and Phua 1980; Sekhon and Mebane 1998).

Our inference method has wide applicability in sequencing studies. Because it only requires a genome-wide coverage higher than 0.001X (as shown in the simulation), it can provide global ancestral information using a wide range of sequencing experiment settings, such as targeted sequencing, exome sequencing and whole genome sequencing.

The inferred ancestries are important in genetic association analysis, since they can be used to control population stratification caused by heterogeneous populations. For example, the ancestral locations have been shown to help discover rare disease-susceptible variants for the AMD disease (Zhan et al. 2013). Other studies have also shown the inferred ancestries of case and control samples can be matched to reduce spurious association.

We describe how to obtain reference coordinates from an external reference panel. The advantage of this step is to ensure the inferred ancestral locations of sequence samples are placed in the same coordinates. When the external reference panel is unavailable, our method can be slightly modified from *Equation 1* so that it can jointly infer the SNP gradients, SNP intercepts and ancestral locations of sequence samples. As $P(G_{ij})$ is a function of $f_j(x_i)$ and the $f_j(x_i)$ has unknown parameters a_j and b_j, the parameters' space now has high dimensionality. The optimization procedure will need to iteratively optimize between gradients/intercepts (A and B) and ancestral locations (X), which is similar to (Yang et al. 2012). However, the inferred ancestral locations will not likely be in the same reference coordinates space where HGDP samples are utilized.

As genetic studies are growing fast, we envision that researchers will aggregate huge amount of sequencing data and will face the so-called "big data" challenge. To tackle the data at this scale, we expect our method and its implementation can rapidly and accurately infer sample ancestries, which can be helpful in a wide range of genetic studies such as association studies. Our method described here has been implemented and is publicly available. We expect that it can facilitate genetic researches in the future.

3.5 Figures and tables

Figure 3-1 Workflow of the likelihood based method

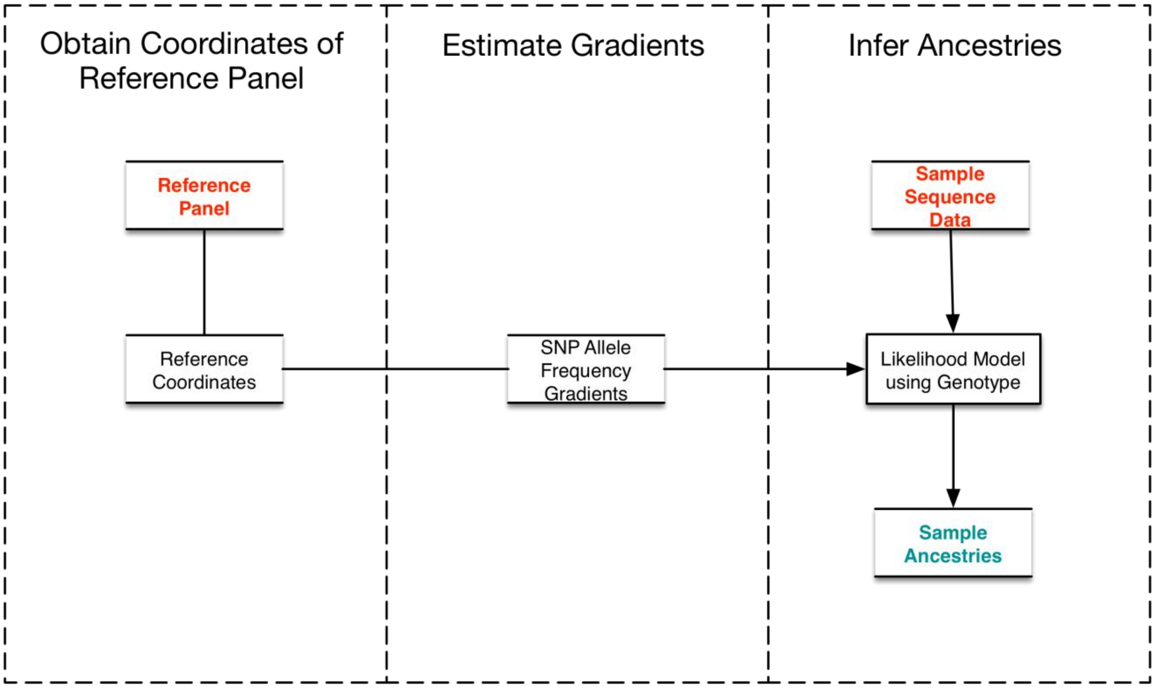


Figure 3-2 Gradients of three markers and an example of joint likelihood surface

(a-c) allele frequency for three markers, their gradients are:

$$a_1 = (1, 0); a_2 = \left(-\frac{1}{2}, \frac{\sqrt{3}}{2}\right);\ a_3 = \left(-\frac{1}{2}, -\frac{\sqrt{3}}{2}\right); b_1 = b_2 = b_3 = 0$$

(d) The joint likelihood surface for a sample with genotypes $g_1 = g_2 = g_3 = (ref, alt)$

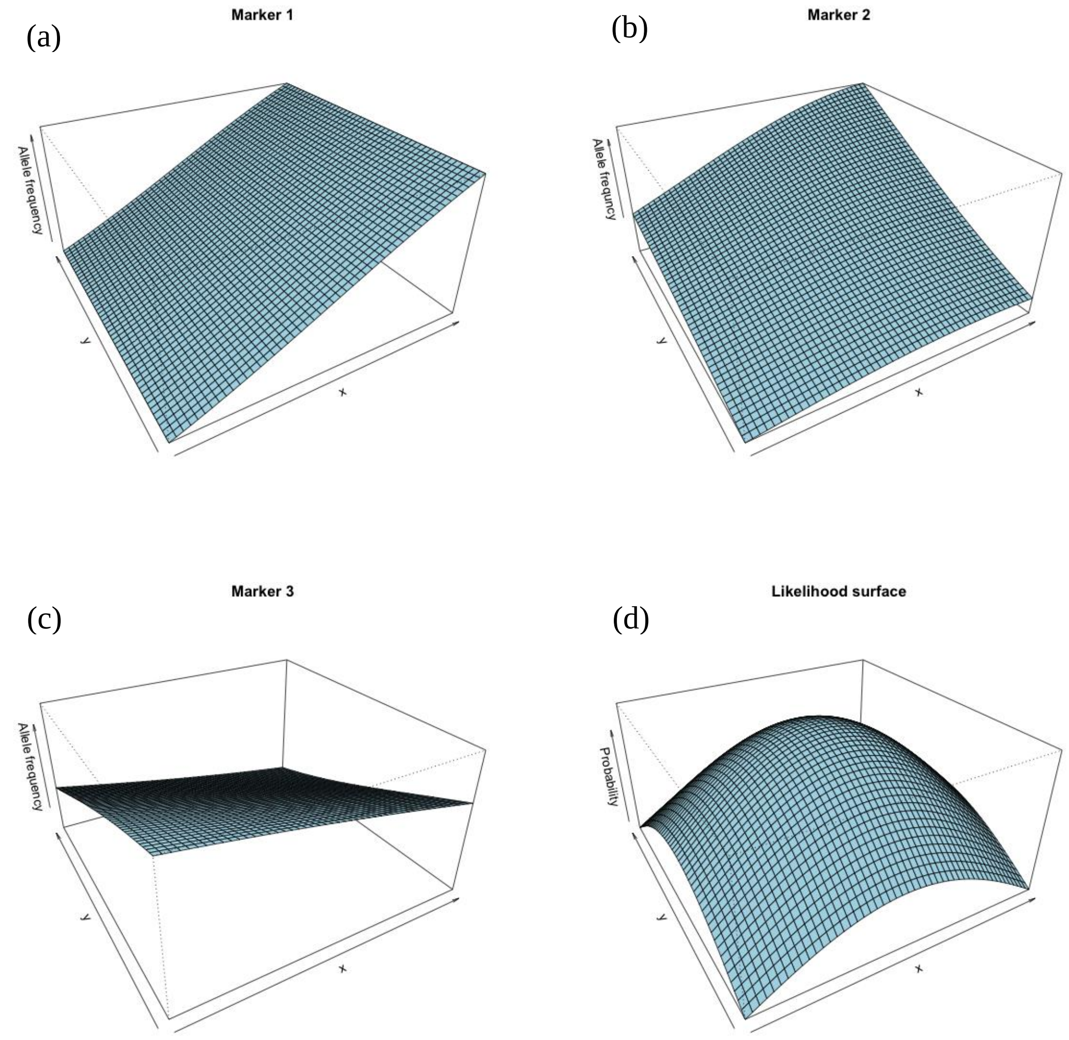


Figure 3-3 Optimization using simplex method

This figure illustrates how simplex method searches the maxima on the likelihood surface. The search path at the beginning is along the gradient direction, then it wiggles the search path but it finally identifies the maxima point in the flatter parts of the likelihood surface.

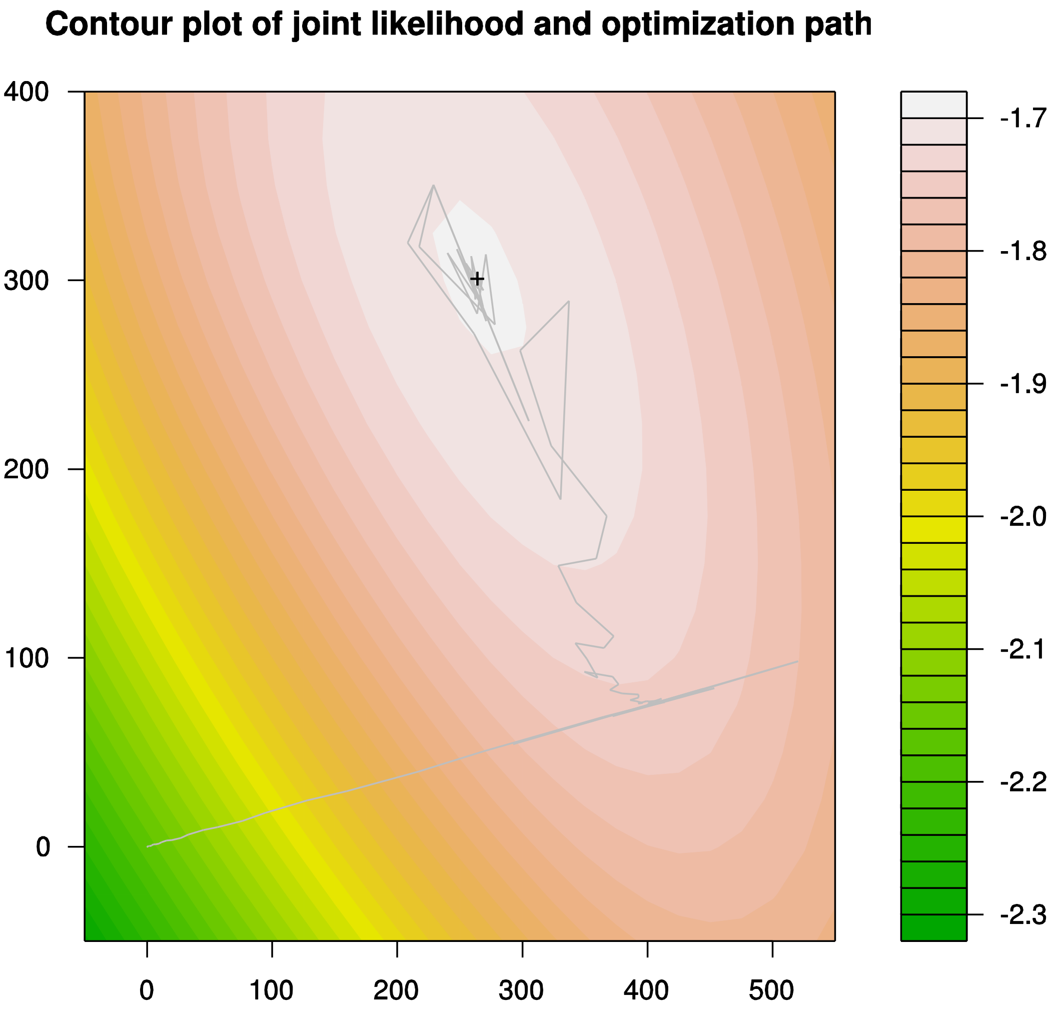
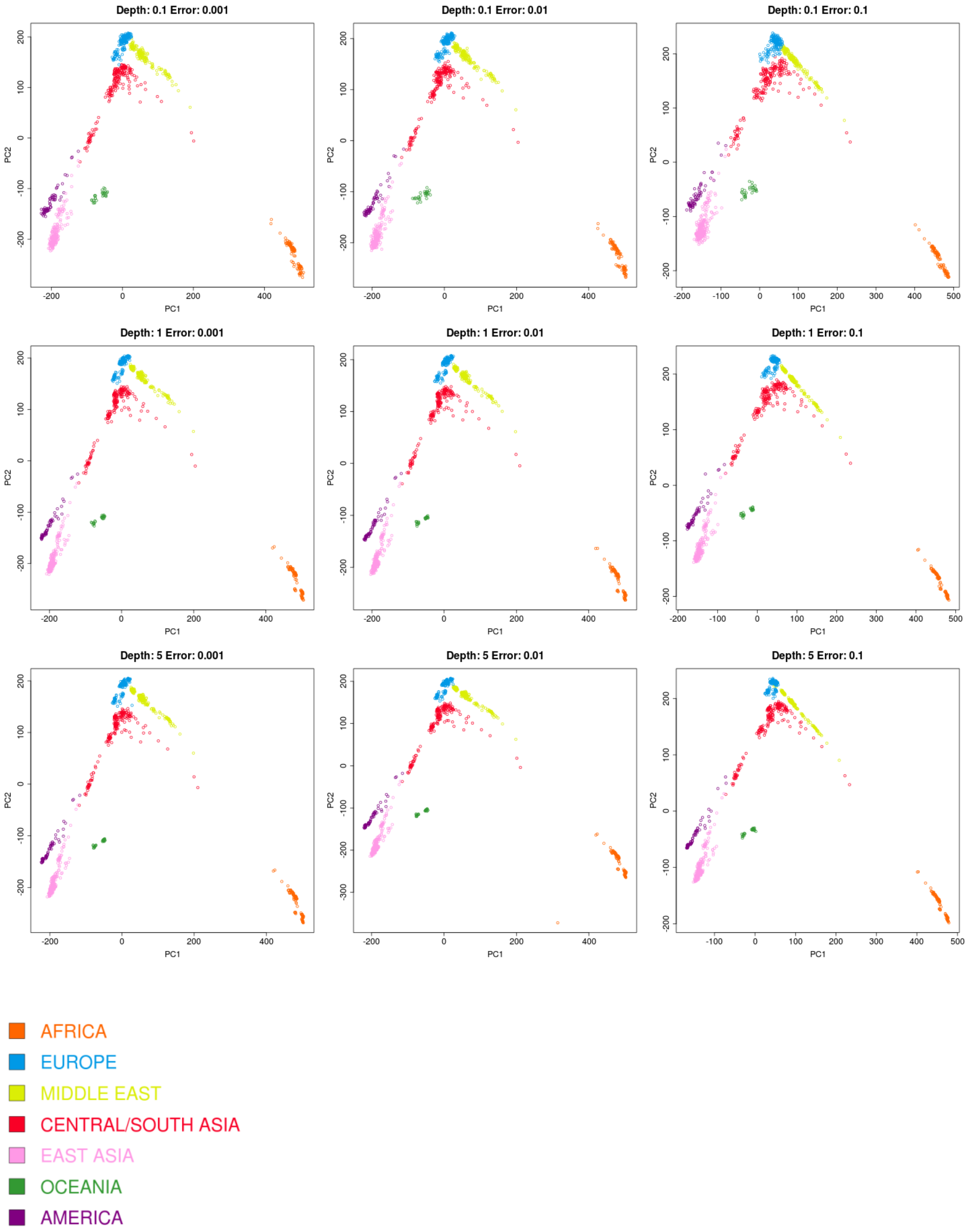


Figure 3-4 Ancestral inference of sequenced data using different depth and error rate

632,907 single nucleotide variants of 938 Human Genome Diversity Project samples are used to simulate sequence data. Sequence error rates per base per read are set at 0.1%, 1% and 10%. The sequence depths are set at 0.1X, 1X and 5X. The seven color legends represent various continental populations.



Depth: 0.1 Error: 0.001
Depth: 0.1 Error: 0.01
Depth: 0.1 Error: 0.1
Depth: 1 Error: 0.001
Depth: 1 Error: 0.01
Depth: 1 Error: 0.1
Depth: 5 Error: 0.001
Depth: 5 Error: 0.01
Depth: 5 Error: 0.1
PC1
PC2
-200
-100
0
100
200
300
400
500
AFRICA
EUROPE
MIDDLE EAST
CENTRAL/SOUTH ASIA
EAST ASIA
OCEANIA
AMERICA

Figure 3-5 Comparison between PCA method, LASER method and the likelihood method

920 samples that are used in both AMD targeted sequencing project and previous International AMD Gene Consortium. We infer their ancestries using different methods.

(Top) The Principal Components Analysis (PCA) method is applied on genome wide SNP data of the 920 samples.

(Middle) The LASER method is applied on sequence data of these samples from the AMD targeted sequencing project.

(Bottom) Our likelihood based method is applied on sequence data of these samples from the AMD targeted sequencing project.

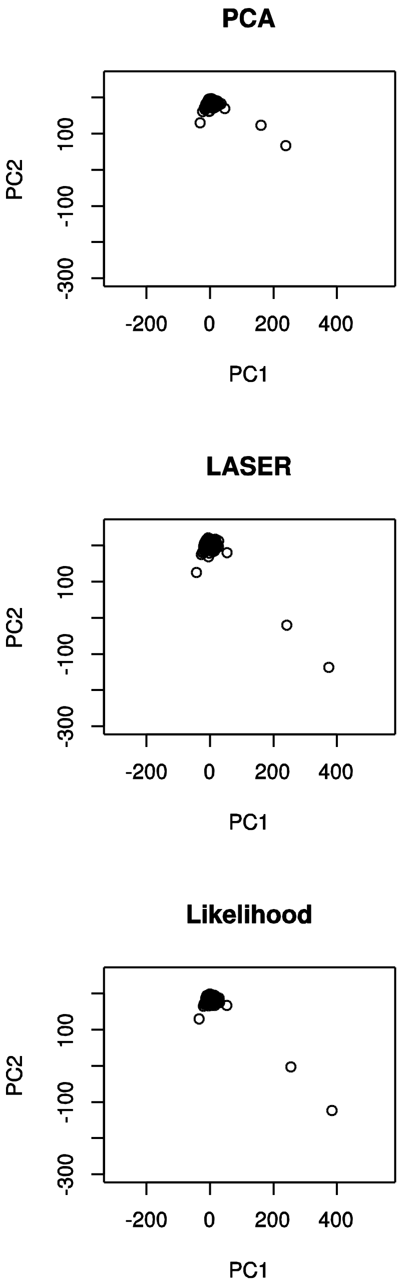

PCA
LASER
Likelihood
PC2
PC1
-300
-100
100
-200
0
200
400

Table 3-1 Accuracy using simulated sequence reads assuming base error of 0.01.

632,907 single nucleotide variants of 938 Human Genome Diversity Project samples are used to simulate sequence data. Sequence error rate per base per read is set from 0.001 to 0.1 and sequence depth is set from 0.001X to 50X. Pearson correlation between likelihood model-based ancestral coordinates and known principal component coordinates are used in evaluation.

Depth	Error						
	0.001	0.002	0.005	0.01	0.02	0.05	0.1
0.001	0.9573	0.9563	0.9571	0.9564	0.9502	0.9395	0.9136
0.002	0.9781	0.9784	0.9775	0.9773	0.9774	0.9689	0.9567
0.005	0.9904	0.9906	0.9896	0.9893	0.9885	0.9859	0.9788
0.01	0.9942	0.9944	0.9945	0.9942	0.9937	0.9923	0.9879
0.02	0.9962	0.9964	0.9962	0.9961	0.9960	0.9953	0.9925
0.05	0.9975	0.9975	0.9976	0.9977	0.9975	0.9972	0.9956
0.1	0.9981	0.9980	0.9979	0.9980	0.9979	0.9979	0.9968
0.2	0.9982	0.9982	0.9982	0.9983	0.9982	0.9981	0.9971
0.5	0.9983	0.9983	0.9983	0.9983	0.9982	0.9983	0.9973
1	0.9984	0.9984	0.9983	0.9983	0.9983	0.9984	0.9972
2	0.9984	0.9984	0.9984	0.9984	0.9984	0.9983	0.9970
5	0.9983	0.9983	0.9983	0.9983	0.9983	0.9983	0.9961
10	0.9982	0.9983	0.9982	0.9982	0.9982	0.9982	0.9956
20	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9961
50	0.9982	0.9982	0.9982	0.9982	0.9982	0.9982	0.9967

Table 3-2 Accuracy using simulated sequence reads assuming base error of 0.001.

632,907 single nucleotide variants of 938 Human Genome Diversity Project samples are used to simulate sequence data. Sequence error rate per base per read is set from 0.001 to 0.1 and sequence depth is set from 0.001X to 50X. Pearson correlation between likelihood model-based ancestral coordinates and known principal component coordinates are used in evaluation.

Depth	Error						
	0.001	0.002	0.005	0.01	0.02	0.05	0.1
0.001	0.9582	0.9562	0.9579	0.9542	0.9519	0.9351	0.8877
0.002	0.9773	0.9777	0.9785	0.9750	0.9722	0.9641	0.9414
0.005	0.9907	0.9903	0.9894	0.9897	0.9881	0.9839	0.9718
0.01	0.9943	0.9941	0.9938	0.9939	0.9929	0.9904	0.9817
0.02	0.9963	0.9964	0.9962	0.9959	0.9957	0.9943	0.9872
0.05	0.9975	0.9974	0.9972	0.9973	0.9973	0.9963	0.9901
0.1	0.9978	0.9978	0.9978	0.9978	0.9978	0.9968	0.9912
0.2	0.9980	0.9981	0.9980	0.9981	0.9980	0.9973	0.9914
0.5	0.9982	0.9981	0.9982	0.9982	0.9982	0.9973	0.9912
1	0.9982	0.9982	0.9982	0.9982	0.9982	0.9973	0.9904
2	0.9982	0.9982	0.9982	0.9983	0.9983	0.9972	0.9884
5	0.9983	0.9982	0.9982	0.9982	0.9982	0.9968	0.9770
10	0.9982	0.9982	0.9982	0.9982	0.9982	0.9968	0.9596
20	0.9982	0.9982	0.9982	0.9982	0.9982	0.9973	0.9678
30	0.9982	0.9982	0.9982	0.9982	0.9982	0.9977	0.9838
50	0.9982	0.9982	0.9982	0.9982	0.9982	0.9975	0.9904
100	0.9982	0.9982	0.9982	0.9982	0.9981	0.9976	0.9905

Chapter 4 Improve genotype calling using external reference panel

4.1 Abstract

Next generation sequencing (NGS) enables more comprehensive researches of genetic variations at a fraction of costs of the traditional capillary sequencing. A key step in the NGS is to overcome sequence errors and generate a large amount of accurate genotype calls. A majority of genotype callers process sequence reads overlapping certain genomic locations, and estimate genotypes using this site-wise information. However, this method can be improved by incorporating linkage-disequilibrium (LD) information using external reference panels. Here, we propose an integrated workflow that can benefits from the ever growing high quality reference panels. In our simulation, this procedure is shown to have high variant discovery rates and improved accuracy. For example, the concordance can be improved from 86.1% to 91.0% in a sequence of 0.5X coverage, using a population-based genotype caller and our proposed LD-aware genotype caller (using a reference panel of 100 samples) respectively. Using the 1000 Genome Project data, our approach is shown to improve the genotype calls with realistic computational costs (genotype concordance improved to 93% using 100 samples). In all, we expect this imputation-based workflow can be applied in population-based sequencing experiments in the future.

4.2 Introduction

Genotype calling algorithms are improving over the time. The earliest approach to call genotypes used one sample at a time (Levy et al. 2007; Wheeler et al. 2008; Lupski et al. 2010), which works relatively well for a small number of deeply sequenced genomes. Methods that examine single samples typically require very deep coverage of each genome to estimate accurate genomes (Bentley et al. 2008; Kim et al. 2009). Later genotype callers examine multiple samples together (Li et al. 2008; Li et al. 2009a; McKenna et al. 2010; Garrison and Marth 2012). This approach uses population information can improve genotype accuracy (e.g. for low coverage samples, accuracy improved from 80% to 87%, (Nielsen et al. 2011)). The most recent callers use linkage disequilibrium (LD)-information by examining haplotypes shared between study samples (Browning and Yu 2009; Li et al. 2009b; Le and Durbin 2011; Wang et al. 2013). It is now clear that methods that use haplotype sharing information can greatly reduce error rates, particularly as the number of sequenced samples increases. With these methods, highly accurate genotypes can be obtained even at relatively low depths, reducing the cost per sequenced sample. For example, in the 1000 Genomes Project pilot, with 179 sequenced samples sequenced at 2X-6X depth, the estimated error rate at heterozygous sites was >70% (The 1000 Genomes Project Consortium 2010). In the 1000 Genomes Phase 1 paper, with 1092 sequenced samples sequenced at 2X-6X depth, the estimated error rate at heterozygous sites decreased to >95% (The 1000 Genomes Project Consortium et al. 2012).

Models that use haplotype sharing describe each sample as an imperfect mosaic of the haplotypes seen in other samples (Li and Stephens 2003). A similar idea provides the foundation to imputation algorithms (Li et al. 2009b; Marchini and Howie 2010). This imperfect mosaic can be modeled using Hidden Markov Models (HMMs) (Browning 2006; Marchini et al. 2007; Li et al. 2010; Hu et al. 2013).

It is clear that, in large studies, combining multi-sample calling to discover variant sites with haplotype sharing analysis to estimate accurate genotypes is now the approach of choice. However, for small studies, the benefits of haplotype sharing approaches are more limited and obtaining good results often still requires deep sequencing of each sample. In principle, many of these small studies could benefit from analyses that combine information on sequenced samples with information on previously estimated haplotypes from other studies. In this chapter, we evaluate the benefits of external reference panels in the analysis of low-coverage sequence data from relatively small studies. We evaluate the genotype accuracy for reference panels of various sizes using both simulated sequence data and data from the 1000 Genome Project. Our results demonstrate that utilizing a large reference panel in the analysis of low pass sequence data for a single sample aids in both variant discovery and genotype accuracy. Our work suggests cost-effective sequencing strategies that leverage the growing number of publicly available reference panels.

4.3 Materials and methods

4.3.1 Models

We illustrate a HMM model and its application in Figure 4-1. In Figure 4-1 (a), following previous work (Li et al. 2011), we denote j as a genomic position along a chromosome and the state S_j as the use of the jth haplotype in the reference panel as template (Figure

4-1 (b)). The transition probability $P(S_{j+1}|S_j)$ corresponds to the probability of recombination (or switching) from state S_j to state S_{j+1}. The emission probability $P(G_j|S_j)$ represents the probability of observing a genotype G_j given the underlying template specified by state S_j. In GWAS studies, G_j often remains unobserved, for example, as a result of genotyping array design.

In practice we use the software beagle to fit the HMM model (Browning 2006). The variable-length Markov Chain model implemented in beagle is an extension of the HMM model described above and is appealing in terms of high computation efficiency and high imputation accuracy (Browning 2006). Here the results of beagle represent the performance of the LD-aware genotype callers.

The goal of applying a HMM model is to combine haplotype information across samples (Figure 4-1 (b)) to infer the untyped genotypes or to provide a prior distribution for low pass genotypes (Figure 4-1 (c)). Utilizing external reference panels of high quality haplotypes, we expect the improved interpretations of sequenced data for most studies.

4.3.2 Simulation design

Simulating true genotypes using coalescent theory

We simulated 15,000 haplotypes using a coalescent model (Figure 4-3) and the program ms (Hudson 2002), each one million base pairs long. The demographic model was consistent with European demographic history (Adams and Hudson 2004; Novembre et al. 2008), which includes an ancestral bottleneck followed by more recent population differentiation and exponential growth.

Our model assumes an ancestral population with an effective population size of $N_1 = 10{,}000$, where an instantaneous bottleneck event 3,000 generations in the past reduced population size to $N_{bottle\,neck} = 75$. Then, our simulations assume that this population simultaneously split into present day populations 500 generations before the present. Following the divergence from the ancestral population, the present-day populations underwent recent exponential growth, each growing to a present day effective population size of $N_0 = 1 \times 10^6$ over 400 generations. We assume equal, symmetric migration rates between the sub-populations with a per-haplotype, per-generation migration rate of 5×10^{-4}. We also assume a per-basepair, per-generation mutation rate of 2.5×10^{-8} and a recombination rate equivalent to 1cM/Mb. We sampled 5,000 haplotypes from each of the present day subpopulations.

Simulating sequence data and reference panels

After coalescent simulation, we randomly paired haplotypes to assemble 2,500 individuals within each population (Figure 4-4). We selected 100 of these as our test data set. To mimic the next generation sequencing technology, we simulated sequence reads with length 100 base pairs and 0.1% base error rate. We set the number of sequence reads such that the average sequence depths equal to 0.5X, 1X, 2X, 4X, 6X, 12X and 30X. Following the simulation workflow (Figure 4-4), we align sequence reads using BWA (Li and Durbin 2009). We used the haplotypes of the remaining individuals as a reference panel. As the coalescent model provides their true haplotypes, we did not model additional genotyping or haplotyping errors.

Simulation scenarios

We set out to explore the utility of the reference. After excluding the 100 test samples, the remaining 7,400 samples were randomized and organized into reference panels of different sizes including between 100 and 2400 individuals. First, we considered reference panels including samples from one population. Then the reference panels are selected from the same population as the test data. Second, we considered reference panels including samples from different populations. We chose 50% reference samples from the same population as the test samples, and 25% reference samples from each of the other two populations.

These simulations mimic real sequencing experiments in which a limited number of samples is sequenced but where large sets of reference samples might be available by aggregating information across prior sequencing studies. To explore optimal strategies for obtaining high quality genotype calls for the test samples, we benchmark various sizes of the reference panels, matching between the reference panels and samples being studied, the sequencing depths and the choice of genotype callers.

4.3.3 The 1000 Genome Project data

In real data examples, we use European samples from the 1000 Genome Project to evaluate sequence experiment design. In the Phase 1 release, there are 330 European ancestry samples drawn from five populations: 184 Utah residents (CEPH) with northern or western European ancestry (CEU), 105 Finnish in Finland (FIN), 107 British in England and Scotland (GBR), 162 Iberian in Spain (IBS) and 112 Toscani in Italia (TSI). We randomly select 10 unrelated samples from CEU population as test samples and use their sequence data on chromosome 20. Their sequencing depths range from 3x to 7x with mean depth of 5x calculated by QPLOT (Li et al. 2013). We assemble the reference panels in two ways. In the first reference panel, the reference panel included 50 CEU samples; in the second reference panel, it included the same 50 CEU samples as well as 25 FIN samples and 25 GBR samples. There were no overlapping samples between test data sets and reference panels.

4.3.4 Genotype calls

We evaluate three types of algorithms to call genotypes. The first genotype caller, implemented in glfSingle, infers genotypes one sample at a time. The second genotype caller, implemented in glfMultiples, infers genotypes examining all sequenced samples jointly (100 samples in the simulations, 10 samples in the 1,000 genomes project analysis). The third genotype caller, implemented in beagle (Browning and Yu 2009), is a LD-aware genotyping algorithm using both sequence reads of all individuals and an external reference panel. Under this setting, we prepare genotype likelihoods for all sites in the external reference panel, and then estimate genotypes using both reference panel haplotypes and sequence information for each sample. The version of glfSingle and glfMultiples is a compiled executable dated in June 2010, and beagle is version 3.3.1 (26 Dec 2010).

4.3.5 Evaluation

We compare these genotype callers in discovery rate and concordance. We define the discovery rate as the fraction of detected variant sites (those called by various genotype callers) among all true variants sites of the test samples (Figure 4-2). The discovery rate is the contrary of false negative rate. Due to the limitation of sequencing technology, it is often the case that rare variants cannot be detected (e.g. low sequence depths). For simulations, the coalescent model provides true genotypes. For real data, we use OMNI genotype array data as the gold standard (The 1000 Genomes Project Consortium 2010).

Similarly, we define the concordance rate as the fraction of correct genotype calls of all variants (Figure 4-2). In simulations, as reference panels have different sizes, we select variants that are located on all reference panels. Then the concordance for each sample is the fraction of correct estimated genotypes at these variant sites. In the real data, we use the genotypes typed on the OMNI chips as gold standard. The concordance can be close to one at rare variant sites even when their genotypes are wrongly called as reference genotypes. However, we compare this statistic across different simulation settings, so it is still informative to identify a good genotype caller.

Ideally, a good genotype caller should simultaneously achieve high discovery rate and high concordance rate, meaning accurate genotype calls at abundant polymorphic sites. To evaluate the three genotype callers, we consistently calculate these two measurements in a wide range of settings.

4.4 Results

4.4.1 Evaluation using simulated data

Coalescent simulation

Using a coalescent model, we obtained 88,550 variant sites among 7,500 simulated samples. Stratified by minor allele frequency, we observed an excessive number of rare variants (81,1164 SNPs with minor allele frequency less than 0.1%, 1,143 SNPs between

0.1% and 0.2%, 1,145 SNPs between 0.2% and 0.5%, 734 SNPs between 0.5% and 1%, 610 SNPs between 1% and 2%, 791 SNPs between 2% and 5%, 2,963 SNPs between 5% and 50%). To measure the difference between populations, we used vcftools to calculate Fst (Weir and Cockerham 1984; Danecek et al. 2011) statistics. In our simulation, we found the Fst value of 0.007, which is comparable to Fst between different European populations (Nelis et al. 2009; Tian et al. 2009).

Test samples are randomly drawn from one population

We randomly selected 100 samples from one of the three populations as test samples. Among all samples, there are 7,180 polymorphic sites (1,808 SNPs with minor allele frequency less than 0.1%, 222 SNPs between 0.1% and 0.2%, 465 SNPs between 0.2% and 0.5%, 449 SNPs between 0.5% and 1%, 504 SNPs between 1% and 2%, 769 SNPs between 2% and 5%, 2,963 SNPs between 5% and 50%). Depending on the sequence depths set from 0.5X to 30X, we obtained from 10,000 to 600,000 paired reads per sample. Using a single sample based genotype caller, glfSingle, we obtained 2,407 (0.5X) to 7,173 (30X) variant per sample. Using a population based genotype caller, glfMultiples, we obtained 3,099 (0.5X) to 7,172 (30X) variants among the same 100 samples. As the above genotype callers do not utilize any external reference panel, we will then evaluate how to incorporate this information.

When reference samples are drawn from the same population

We first explored how the size of the reference panel affects the genotype calling quality in one population. In Figure 4-5, glfSingle and glfMultiples do not use reference information so their variant discovery rates and concordance rates are independent of reference panel size. Noticeably, the single sample based method provides the worst in discovery rate. In contrast, imputation-based methods use the reference panels to help identify variant sites. The discovery rate for the imputation method improved upon that for methods that used the test samples alone (mean discovery rate for beagle: 98.84%) compared to other methods. Figure 4-5 also shows that the imputation-based method performs best when the largest reference panel is used. When the reference panel increased from 100 samples to 2,400 samples, the reference polymorphic markers increased from 7,006 SNPs to 34,771 SNPs. The reference panels with denser SNP markers can help discover more variants of the test samples. In terms of overall concordance on a common set of 7,005 SNPs, imputation-based software also had advantages comparing to other methods and it performed better when reference panel size increased as well (e.g. at 0.5x depth, glfSingle: 84.4%; glfMultiples 86.1%; beagle using 100 reference samples: 90.9%; beagle using 800 reference samples = 94.0%).

Imputation-based genotype analyses also outperformed other methods when we performed comparisons stratified by allele frequencies. In Figure 4-6, we separated variants into 6 minor allele frequency bins and evaluated performance in each bin. When examining discovery rates, glfSingle had poor power to detect variants with minor allele frequency between 0.1% and 0.2% (discovery rate: 50.1%) while imputation methods (beagle) outperformed population-based analysis of test samples when the reference panel size was larger than 200 (discovery rate: 77.5% using 200 reference samples and 100.0% using 400 or more reference samples versus 46.4% using the test samples alone and a population based caller like glfMultiples). When evaluating concordance, we also noticed that larger reference panels helped imputation-based genotype callers to outperform other genotype callers when MAF was larger than 1% (Figure 4-6). For example, at sequence depth of 4x, the concordance of glfSingle and glfMultiples is 98.01% and 99.02% for the low frequency variants (MAF between 1% and 2%). Using the LD-aware genotyping method, beagle can achieve 99.4% accuracy using only 100 reference samples. When the reference panel increases, the accuracy will increase as well. We also notice imputation-base method has more prominent heterozygous accuracy. Evaluated using 1% to 2% variants again, the accuracy for glfSingle, glfMultiples and beagle (100 reference sample) in the heterozygous sites are 44.0%, 74.6% and 79.8% respectively.

When imputation panel consists of the multiple populations

We then set out to investigate reference panels with multiple populations. In Table 4-1, we compared a reference panel of the same population to a reference panel of three populations with 2:1:1 ratio. Under both settings, imputation-based genotype callers performed better with larger reference panels. For example, when 100 reference samples of the same population served as reference and 100 tests samples were sequenced at 0.5X, imputation-based produced 90.73% concordance. When reference samples increased to 800, the concordance rate also increased to 93.67%. We observed that, when the reference panel size is kept constant, the imputation-based methods performed better when the reference panel contained more individuals from the same population as test samples.

4.4.2 Evaluation using the 1000 Genome Project samples

We then evaluated genotype callers using two reference panels from 1000 Genome Project data, and the results were similar to our simulations (Figure 4-4). The population based caller, glfMultiples, only discovered 92.7% of the variants and had concordance of 82.8% for all variants. Our imputation-based methods improved both the discovery rate and concordance. Using 50 CEU reference samples and 100 European reference samples (50 CEU, 25 FIN and 25 GBR), the overall concordance was 92.1% and 92.3% respectively. We also measure the computational times, 55 minutes and 114 minutes, for the two scenarios respectively.

4.5 Discussion

Here, we evaluate imputation-based genotype callers using external reference panels. The reference panel summarizes population level haplotype information and helps improve genotype calls. Using simulations and 1000 Genome Project data, we have shown that imputation-based methods can outperform non-LD aware callers in terms of the concordance and the discovery rate.

Our results directly inform the analysis of future genetic studies. As we expect large-scale sequencing experiments to become more popular, large numbers of high quality haplotypes will soon become available. Our work shows that these can serve as a resource to improve analysis of newly sequenced samples, reducing sequencing costs and thus speeding up sequencing based genetic researches. Incorporating reference panels can improve the efficiency of lower coverage sequencing. In our simulation, with 4x average sequencing coverage, the genotype concordance can be improved from 96.8% to 99.6% using a reference panel of 2400 samples. Specifically, the overall concordance at the heterozygous sites can be improved from 81.1% to 97.2%.

The advantage of the reference panel is to provide high quality population-level haplotype information. At variant sties, the information can improve genotype calls due to precise allele frequency prior of the variant or the haplotype stretch around the variant sites. We have shown that by modeling LD near the variant sites, the genotype calling can be improved across variants of all allele frequencies. However, we also notice that some variants are unique in the sequence samples but not in the reference panel. These variants cannot be improved by our method. Another limitation emerges when the reference panel size increase. The computational cost in a standard HMM model has quadratic computational complexity. When scaling up to a whole genome low coverage sequencing experiments, the computational time can consume 50-100 hours for 100 low coverage sequence samples using 50-100 reference samples. When thousands of reference samples are used as a reference panel, the computational time may not be neglected.

Large panel of haplotypes is a useful resource, and there have been growing interests in building and applying it. Previously, the major histone complex (MHC) haplotype consortium deepen the understanding our understanding of the common variation in the MHC region (Horton et al. 2008). In the recent years, there is another haplotype consortium (O'Connell et al.) in which thousands of sequence based haplotypes can be utilized in improving genotype calls, the same goal as we described here. Therefore, we envision the workflow described here can be further integrated with this type of resources.

4.6 Figures and tables

Figure 4-1 Illustration of HMM model using external reference panel

(a) Hidden Markov Model where S are the hidden states and G represent observed genotypes; (b) Reference haplotypes. Each haplotype represents a possible state, and can be selected to update sample genotypes; (c) Sample with uncertain genotypes and uncertain phase. Each pair of the genotype can be copied from the reference haplotypes.

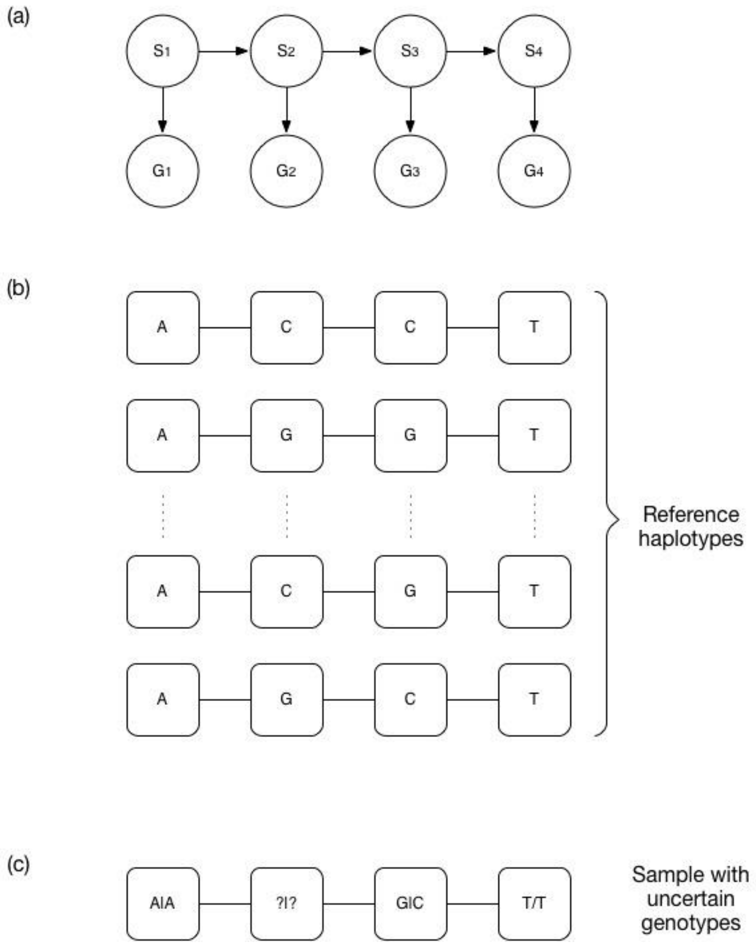
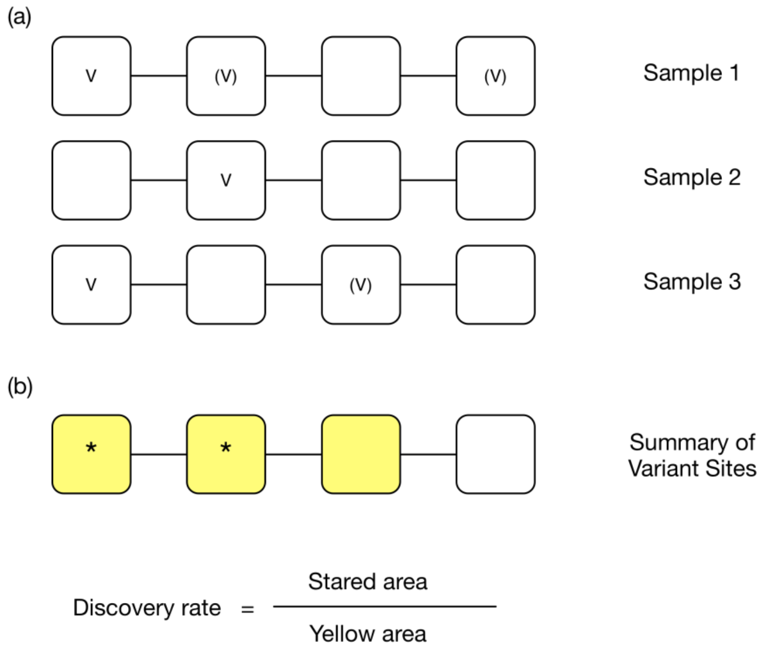


Figure 4-2 Definition of discovery rate and concordance

Discovery rate is a measurement of sensitivity, which is defined as the fraction of detected variant sites among all variant sites. In (a), we illustrate four genomic locations of three samples. The letter “V” without brackets denotes a sample carries a variant detected by a genotype caller, and the letter “V” with bracket denotes the variant that is not detected. In (b), we illustrate a summary of variant sites of the three samples. The star symbols indicate detected variant sites, and the yellow boxes denote all true variant sits among three samples. We can define discovery rate as 2 (counts of stared area) divided by 3 (counts of yellow area).



Genotype concordance is a measurement of accuracy, which is defined as the fraction of correctly estimated genotypes among all estimated genotypes. HomRef: homozygous reference alleles; Het: heterozygous alleles; HomAlt, homozygous alternative alleles.

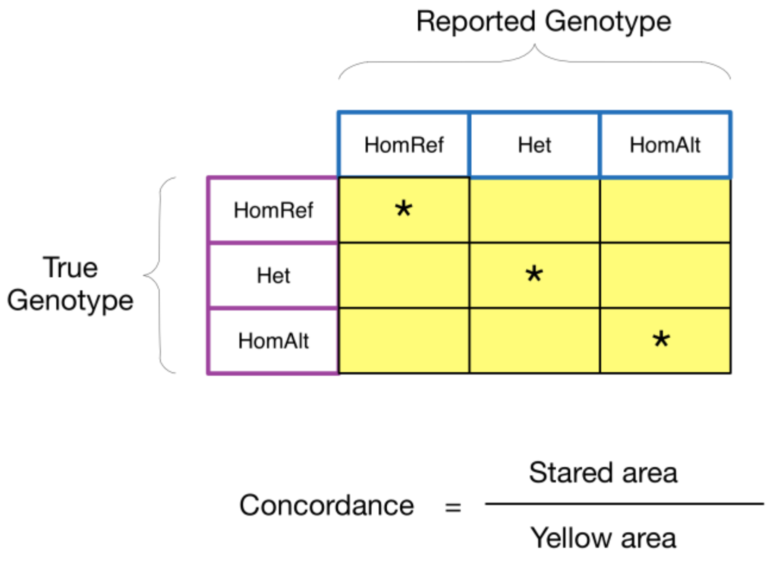


Figure 4-3 Coalescent simulations of European haplotypes

Demographic model for simulated European populations. The demographic model includes an ancient population bottleneck, recent exponential growth, differentiation and migration. The model parameters were calibrated to mimic populations sampled in continental Europe. The model parameters are described in the method part “coalescent simulation”.

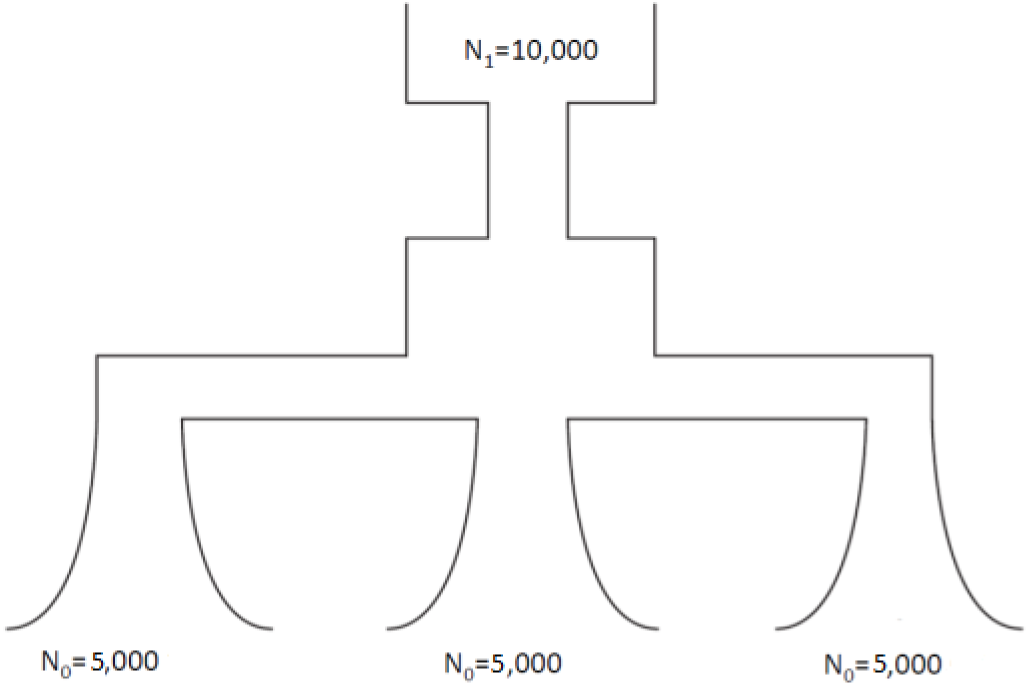


Figure 4-4 Simulation workflow

In simulation, we first simulation genotypes for test samples and reference samples. For test samples, we perform standard sequence simulations (details described in 4.3.2). For reference samples, we choose all or part of them as reference panels of various sizes. Finally, we compare three methods: glfSingle, glfMultiples and beagle.

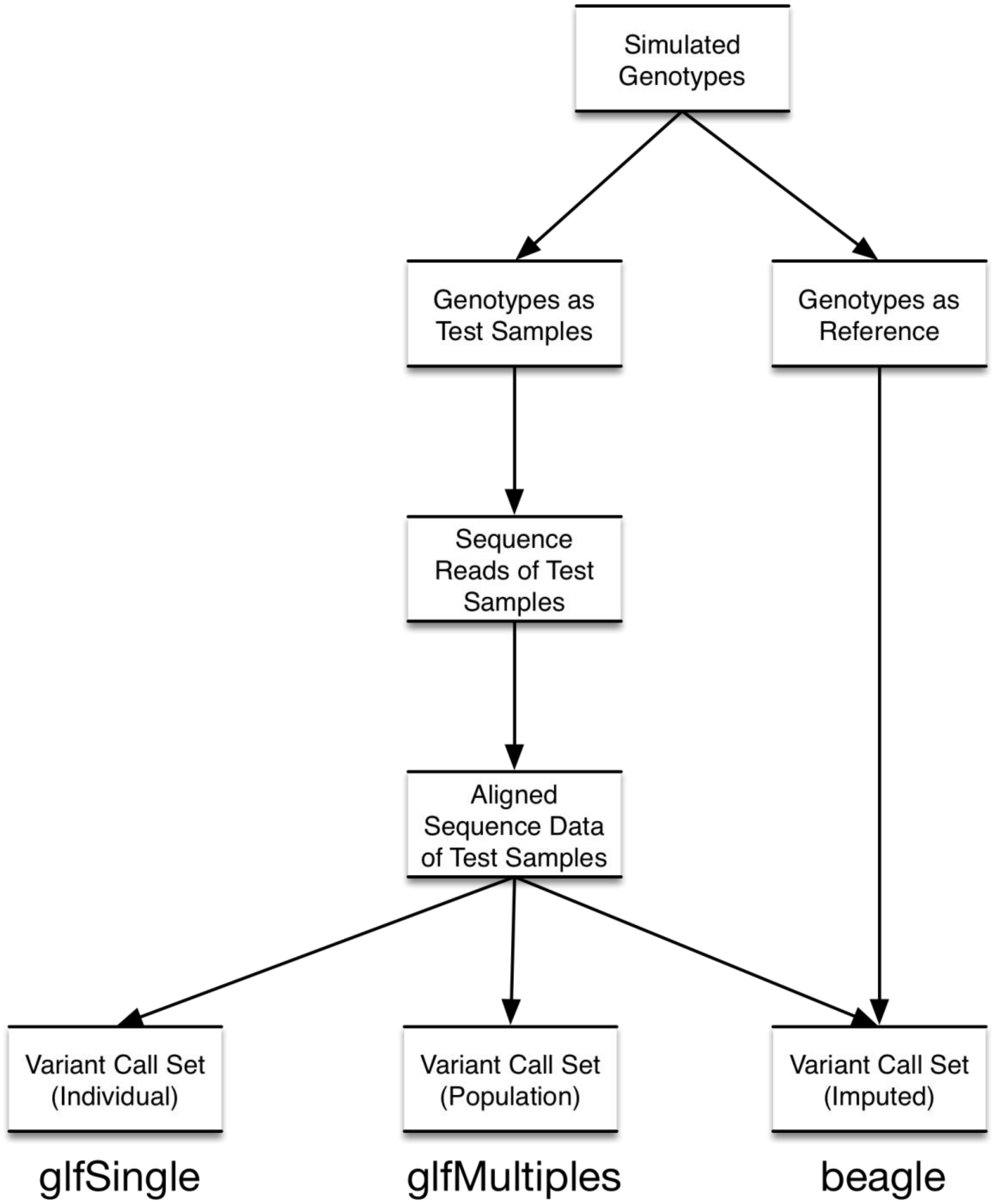
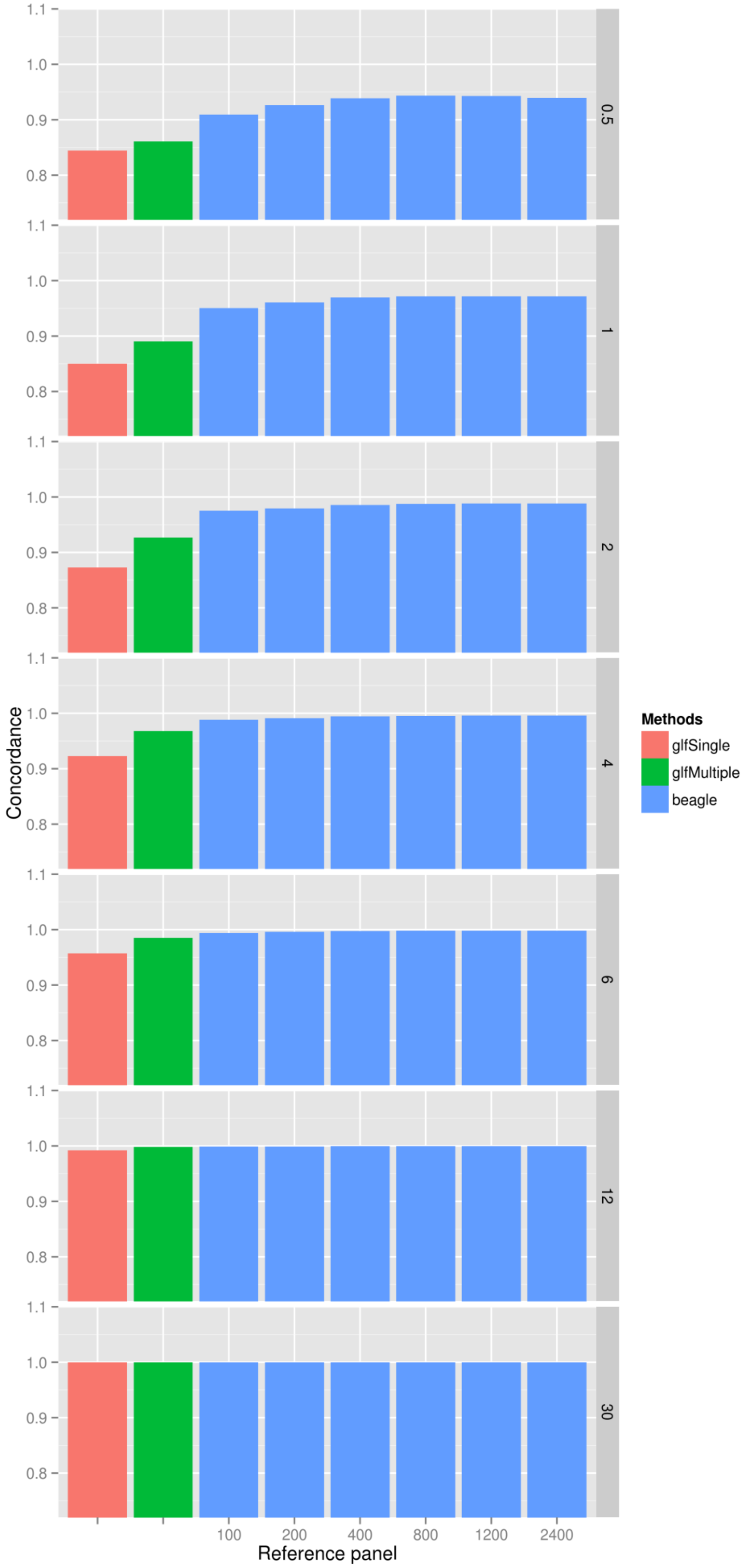

Simulated Genotypes
Genotypes as Test Samples
Genotypes as Reference
Sequence Reads of Test Samples
Aligned Sequence Data of Test Samples
Variant Call Set (Individual)
Variant Call Set (Population)
Variant Call Set (Imputed)
glfSingle
glfMultiples
beagle

Figure 4-5 Marginal genotype calling accuracy.

Genotype accuracy is evaluated by concordances and discovery rates. At various sequence depths (0.5X, 1X, 2X, 4X, 6X, 12X and 30X), we evaluate three approaches: glfSingle, glfMultiples and beagle. In evaluating beagle, we choose different reference panel sizes (100, 200, 400, 800, 1200 and 2400).


Concordance
1.1
1.0
0.9
0.8
0.5
1
2
4
6
12
30
100
200
400
800
1200
2400
Reference panel
Methods
glfSingle
glfMultiple
beagle

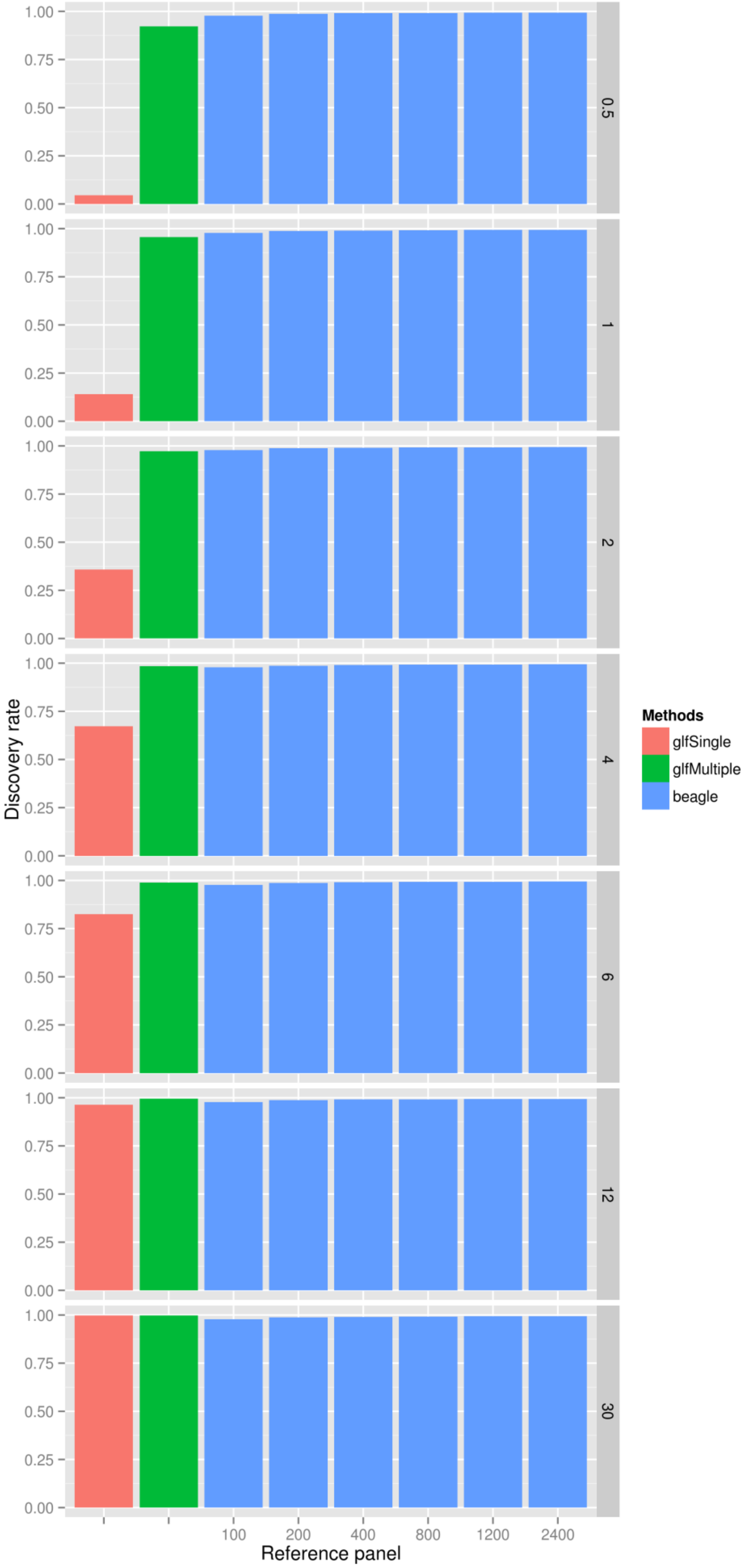
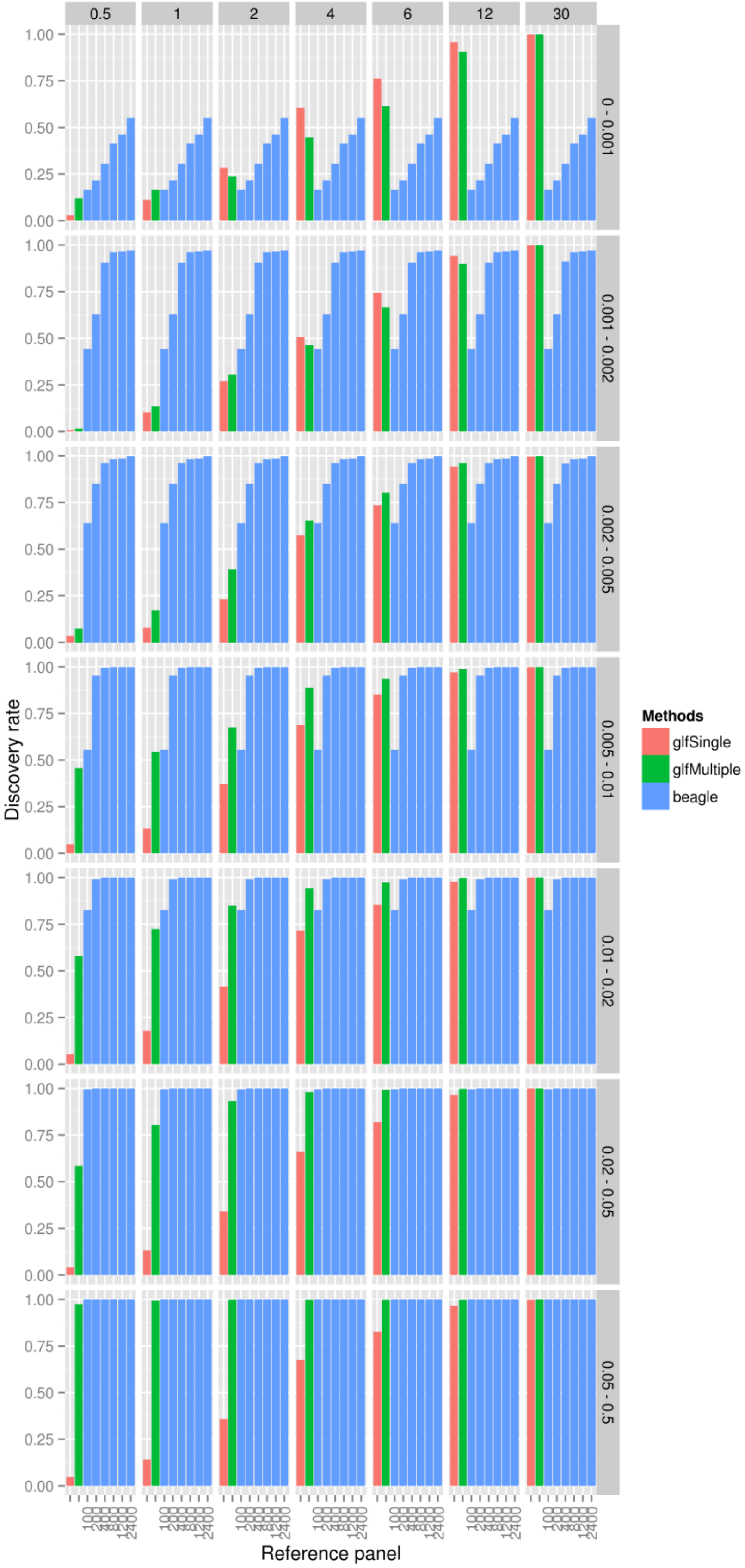

Discovery rate
1.00
0.75
0.50
0.25
0.00
0.5
1
2
4
6
12
30
100
200
400
800
1200
2400
Reference panel
Methods
glfSingle
glfMultiple
beagle

Figure 4-6 Genotype calling quality stratified by MAF at various depths.

We evaluate genotype concordance and discovery rate in various simulation settings. The evaluations are stratified by allele frequency bins: 0-0.1%, 0.1-0.2%, 0.2-0.5%, 0.5-1%, 1-2%, 2-5%, 5-50%. For concordance, we chose 0.5X, 1X, 2X and 4X, as higher depths have similar concordance as in the 4X settings. For discovery rates, we choose sequence depths at 0.5X, 1X, 2X, 4X, 6X, 12X and 30X. We evaluate three approaches: glfSingle, glfMultiples and beagle. In evaluating beagle, we choose different reference panel sizes (100, 200, 400, 800, 1200 and 2400).


0.5
1
2
4
6
12
30
0 - 0.001
0.001 - 0.002
0.002 - 0.005
0.005 - 0.01
0.01 - 0.02
0.02 - 0.05
0.05 - 0.5
Discovery rate
0.00
0.25
0.50
0.75
1.00
Reference panel
100
200
400
800
1200
2400
Methods
glfSingle
glfMultiple
beagle

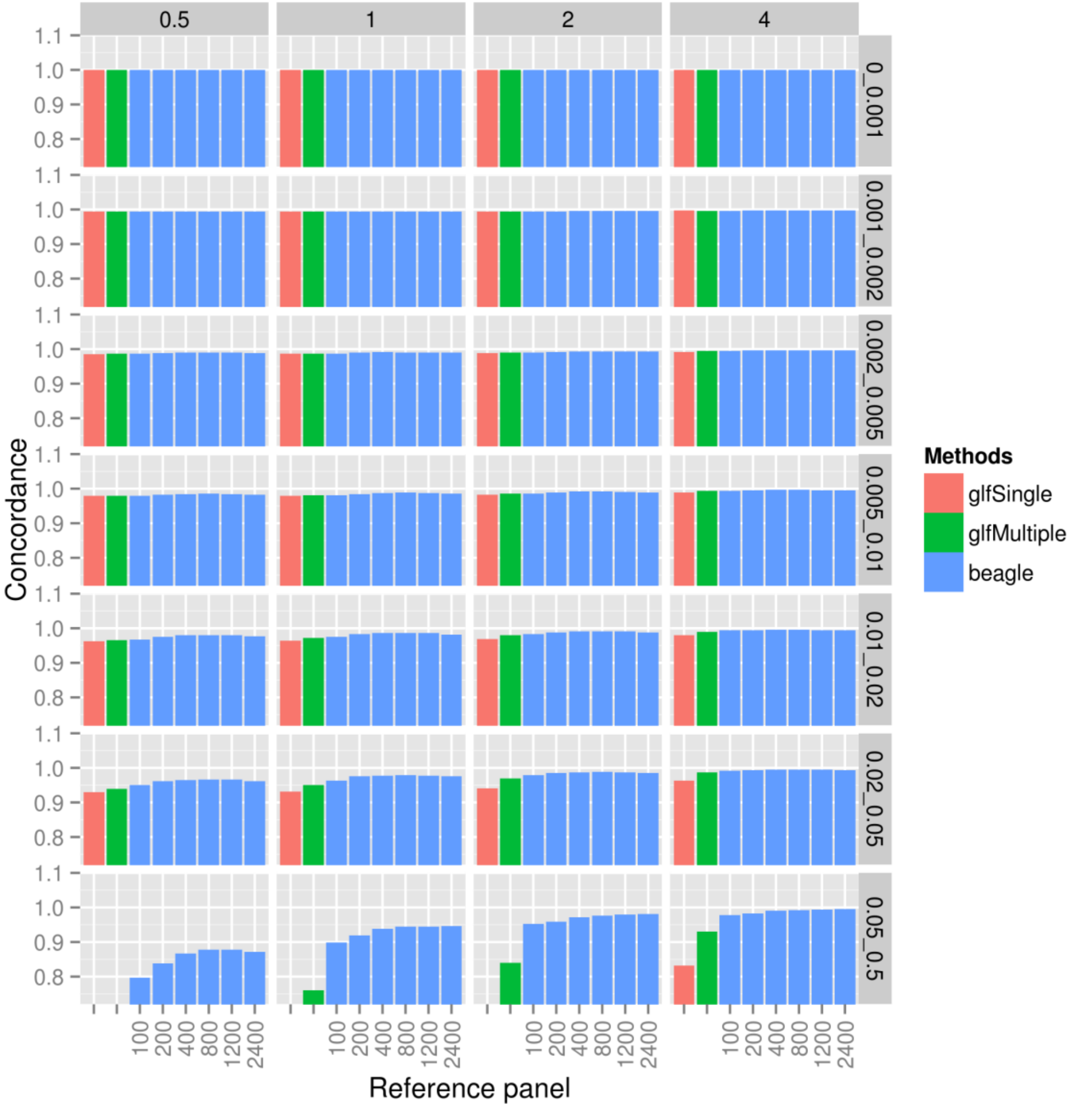

0.5
1
2
4
0_0.001
0.001_0.002
0.002_0.005
0.005_0.01
0.01_0.02
0.02_0.05
0.05_0.5
1.1
1.0
0.9
0.8
Concordance
100
200
400
800
1200
2400
Reference panel
Methods
glfSingle
glfMultiple
beagle

Table 4-1 Comparison of genotype calling quality using different reference panel

We compare two settings of reference panels: (Up) the reference panel consists of one population; (Down) the reference panel consists of three populations (counts of population size is 2:1:1). We evaluate the concordances and discovery rates at various sequence depths (0.5X, 1X, 2X, 4X, 6X, 12X and 30X). The imputation-based approach, beagle, is used in this simulation (50 iterations) and we choose different reference panel sizes (100, 200, 400, 800).

Same Population in Reference Panel

		Reference Panel Size					
		100	200	400	800	1200	2400
Sequence Depth	0.5	0.9094	0.9266	0.9385	0.9433	0.9428	0.9395
	1	0.9507	0.9607	0.9694	0.9719	0.9716	0.9716
	2	0.9751	0.9791	0.9852	0.9873	0.9880	0.9885
	4	0.9886	0.9912	0.9944	0.9955	0.9960	0.9961
	6	0.9941	0.9958	0.9972	0.9979	0.9981	0.9982
	12	0.9990	0.9993	0.9994	0.9995	0.9996	0.9996
	30	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
Discovery Rate		0.9780	0.9875	0.9905	0.9921	0.9928	0.9941

Different Populations in Reference Panel

		Reference Panel Size					
		100	200	400	800	1200	2400
Sequence Depth	0.5	0.9073	0.9177	0.9299	0.9367	0.9371	0.9340
	1	0.9469	0.9581	0.9648	0.9686	0.9702	0.9683
	2	0.9736	0.9791	0.9838	0.9868	0.9880	0.9867
	4	0.9885	0.9919	0.9940	0.9957	0.9960	0.9954
	6	0.9938	0.9957	0.9971	0.9979	0.9981	0.9978
	12	0.9984	0.9993	0.9994	0.9996	0.9996	0.9995
	30	0.9992	0.9998	0.9998	0.9998	0.9998	0.9998
Discover Rate		0.9799	0.9835	0.9890	0.9911	0.9890	0.9911

Table 4-2 Evaluation of genotyping calling using the 1000 Genome Project data

We evaluate the genotype callers (glfMultiples and beagle) using the 1000 Genome Project data. We show the concordance and discover rate stratified by different minor allele frequency bins (MAF are obtained from the 1000 Genome Project Phase 1 release (2011/05/21)). When evaluation uses beagle, we let beagle iterate 50 times.

		glfMultiples	beagle	
			50	100
MAF	0.002 - 0.01	0.8889	0.7143	0.8125
	0.01 - 0.02	0.9801	0.9472	0.9332
	0.02 - 0.05	0.9474	0.9054	0.9029
	0.05 - 0.5	0.8269	0.9287	0.9308
	(all)	0.8281	0.9206	0.9227
Discovery Rate		0.9272	0.9622	0.9674

Chapter 5 Summary and Discussion

Next-generation sequencing technology has improved fast over the past several years. It enables researchers to study a broad range of genetic questions: from characterization of DNA and RNA variations, to metagenomics and transcriptomics in a more systematic manner. However, this new technology brings up new challenges. For example, due to the long processing time and the high error rate of NGS technology, it is challenging to use traditional methods to analyze the mass amount of imperfect sequences. To help overcome these challenges in the sequence era, my thesis includes three chapters. Each chapter focuses on one useful statistical method that can efficiently handle large amount of sequence data, extend related statistical genetics questions, and produce accurate and meaningful results.

Chapter 2 focuses on ancestral inference of sequence samples in an age-relate macular degeneration association study. While the ancestral inference question using genotyping array has been widely studied (Price et al. 2006), the commonly used principal components analysis (PCA) can infer global ancestries or fine-scale European ancestries. However, three factors perplex this specific question: (1) coverage of sequence reads is not uniformly distributed, even in the targeted sequence region; (2) per-base error rate in the sequence reads is higher than in genotype data; (3) the majority of sequence data is close to known disease susceptible loci. These three problems are barriers to applying the PCA method.

Thus, to address these limitations, we have developed the LASER (Locating Ancestries using Sequence Reads) method, which does not require genome-wide sequence coverage. It robustly incorporates error in the estimation procedure, and utilizes both on-target and off-target sequence reads to avoid the unclear influence of the disease susceptible loci. The LASER has enabled us to identify ancestries of sequenced samples in the Age-related Macular Degeneration (AMD) project as well as in the NHLBI exome sequencing project (Fu et al. 2013). We matched sets of 2,268 cases and 2,268 controls and focus on coding variants with deep (minimum 10X) coverage. Subsequent association analysis identified two strongly associated variants, one in the *CFH* gene (control frequency = 0.02%, exact P-value = 2.91x10-6, OR = 23.11) and another in the *C3* gene (control frequency = 0.40%, exact P-value = 2.73x10-4, OR = 2.68). This discovery will help us understand the disease etiology and my help treat this disease more effectively in the future. Although the method still required much computation time, it represents a significant improvement over the PCA method in terms of ancestral inference from the targeted sequencing data.

As an extension to the work in Chapter 2, we investigated a likelihood model in Chapter 3 that can efficiently infer ancestral origins using both genotype data and sequence data. This work utilizes the genetic marker gradients, which describe how allele frequency changes geographically. We extensively evaluated this model using simulated genotypes and sequence data and obtained accurate ancestral locations at modest computation cost. We further evaluated the model using target-sequence data in an AMD study. The method yielded comparable results to the PCA method, but uses a fraction of the computation time. The computational advantages suggest our method will be applicable to a wide range of genetic studies such as “the big data” problem which can include a large number of sequenced individuals. This method could even potentially be commercially available to the public interested in understanding their ancestries.

Besides the ancestral inference questions, in Chapter 4, we advocate a genotyping calling strategy using the ever-growing reference panel. Currently, sequence technology is becoming increasingly adapted so that more genetic studies will benefit from it. Therefore, the size of publicly available reference panels can be expected to grow. Our strategy showed that these reference panels can be incorporated in the sequence genotyping procedure using a Hidden Markov Model (HMM). This finding can be helpful for researchers conducting sequence experiments and seeking to improve existing genotyping results.

Building on this dissertation, there are several aspects to explore in the future. Our likelihood-based ancestral inference method has been evaluated on world-wide populations using the Human Genome Diversity Project (HGDP) data set. To study a finer-scale population structure, it will be interesting to apply this method on European populations. We are in the process of applying the Population Reference Sample (POPRES) data set which provides genotype data of European populations. When this data set becomes available, we expect further evaluations and refinements of our method. Our improved imputation strategy can produce more accurate genotypes. We can then perform association tests. We have explored a Genetic Random Field (GRF) based method based on a novel alternative hypothesis, which models genotypes as random fields (He et al. 2014). Evaluation of the Dallas Heart Study (Victor et al. 2004) has shown potentially interesting outcomes, and we can combine these results with the imputation procedure to closely study disease etiology.

In summary, this dissertation includes ancestral inference algorithms and imputation-based genotyping strategies. These are well suited for next-generation sequence studies as they are accurate and efficient compared to a wide range of existing work. Specifically, the ancestral inference algorithm is useful for genetic association studies to alleviate the so-called population stratification problem. Furthermore, the imputation strategy is widely applicable to researchers hoping to improve genotyping from sequence data. We envision more sequencing experiment will be conducted in the medical genetics fields. We expect that our methods play an essential role in various downstream analyses, and we hope to extend these methods to a broader range of applicable fields. In all, it can be expected that our statistical methods applied in next generation sequence data will greatly reduce costs and lead to more genetic findings.

Bibliography

Adams AM, Hudson RR. 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* **168**(3): 1699-1712.

Age-Related Eye Disease Study Research Group. 2000. Risk factors associated with age-related macular degeneration. A case-control study in the age-related eye disease study: Age-Related Eye Disease Study Report Number 3. *Ophthalmology* **107**(12): 2224-2232.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**(9): 1655-1664.

Arakawa S, Takahashi A, Ashikawa K, Hosono N, Aoi T, Yasuda M, Oshima Y, Yoshida S, Enaida H, Tsuchihashi T et al. 2011. Genome-wide association study identifies two susceptibility loci for exudative age-related macular degeneration in the Japanese population. *Nat Genet* **43**(10): 1001-1004.

Becerra JE, Khoury MJ, Cordero JF, Erickson JD. 1990. Diabetes mellitus during pregnancy and the risks for specific birth defects: a population-based case-control study. *Pediatrics* **85**(1): 1-9.

Bentley DR Balasubramanian S Swerdlow HP Smith GP Milton J Brown CG Hall KP Evers DJ Barnes CL Bignell HR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218): 53-59.

Berthoumieux S, de Jong H, Baptist G, Pinel C, Ranquet C, Ropers D, Geiselmann J. 2013. Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Molecular systems biology* **9**: 634.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Browning BL, Yu Z. 2009. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *The American Journal of Human Genetics* **85**(6): 847-861.

Browning SR. 2006. Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet* **78**(6): 903-913.

Broyden CG. 1970. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics* **6**(1): 76-90.

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**(7145): 661-678.

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN. 2005. Demonstrating stratification in a European American population. *Nature genetics* **37**(8): 868-872.

Cantor RM, Lange K, Sinsheimer JS. 2010. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics* **86**(1): 6-22.

Cardon LR, Palmer LJ. 2003. Population stratification and spurious allelic association. *The Lancet* **361**(9357): 598-604.

Chen W-M, Abecasis GR. 2007. Family-based association tests for genomewide association scans. *The American Journal of Human Genetics* **81**(5): 913-926.

Chen W, Stambolian D, Edwards AO, Branham KE, Othman M, Jakobsdottir J, Tosakulwong N, Pericak-Vance MA, Campochiaro PA, Klein ML et al. 2010. Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. *Proceedings of the National Academy of Sciences of the United States of America* **107**(16): 7401-7406.

Chen Y-H, Chatterjee N, Carroll RJ. 2013. Using shared genetic controls in studies of gene-environment interactions. *Biometrika* **100**(2): 319-338.

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11**(6): 415-425.

Cox DR, Shell EJ. 1970. Analysis of Binary Data.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**(15): 2156-2158.

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

Edwards AO, Ritter R, 3rd, Abel KJ, Manning A, Panhuysen C, Farrer LA. 2005. Complement factor H polymorphism and age-related macular degeneration. *Science* **308**(5720): 421-424.

Fagerness JA, Maller JB, Neale BM, Reynolds RC, Daly MJ, Seddon JM. 2009. Variation near complement factor I is associated with risk of advanced AMD. *Eur J Hum Genet* **17**(1): 100-104.

Ferreira VP, Herbert AP, Cortes C, McKee KA, Blaum BS, Esswein ST, Uhrin D, Barlow PN, Pangburn MK, Kavanagh D. 2009. The binding of factor H to a complex of physiological polyanions and C3b on cells is impaired in atypical hemolytic uremic syndrome. *Journal of immunology* **182**(11): 7009-7018.

Fletcher R. 1970. A new approach to variable metric algorithms. *The Computer Journal* **13**(3): 317-322.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN. 2004. Assessing the impact of population stratification on genetic association studies. *Nature genetics* **36**(4): 388-393.

Friedman DS, O'Colmain BJ, Munoz B, Tomany SC, McCarty C, de Jong PT, Nemesure B, Mitchell P, Kempen J. 2004. Prevalence of age-related macular degeneration in the United States. *Archives of Ophthalmology* **122**(4): 564-572.

Fritsche LG Chen W Schu M Yaspan BL Yu Y Thorleifsson G Zack DJ Arakawa S Cipriani V Ripke S et al. 2013. Seven New Loci Associated with Age-Related Macular Degeneration. *Nature Genetics* **(in press)**.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**(7431): 216-220.

Garrison E. Hapviz: https://github.com/ekg/hapviz.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing.

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang L-Y, Huang W, Liu B, Shen Y. 2003. The international HapMap project. *Nature* **426**(6968): 789-796.

Gold B, Merriam JE, Zernant J, Hancox LS, Taiber AJ, Gehrs K, Cramer K, Neel J, Bergeron J, Barile GR et al. 2006. Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nat Genet* **38**(4): 458-462.

Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR et al. 2005. Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**(5720): 419-421.

He Z, Zhang M, Zhan X, Lu Q. 2014. Modeling and testing for joint association using a genetic random field model. *Biometrics*.

Helgason H, Sulem P, Thorleifsson G, Stefansson H, Masson G, Gudbjartsson D, Walters GB, Magnusson O, Kong A, Rafnar T et al. 2013. A rare nonsynonymous sequence variant in C3 confers high risk of age-related macular degeneration. *Nat Genet*.

Heurich M, Martinez-Barricarte R, Francis NJ, Roberts DL, Rodriguez de Cordoba S, Morgan BP, Harris CL. 2011. Common polymorphisms in C3, factor B, and factor H collaborate to determine systemic complement activity and disease risk. *Proceedings of the National Academy of Sciences of the United States of America* **108**(21): 8761-8766.

Hirji KF, Mehta CR, Patel NR. 1987. Computing Distributions for Exact Logistic-Regression. *J Am Stat Assoc* **82**(400): 1110-1117.

Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, Forbes S, Gilbert JG, Halls K, Harrow JL et al. 2008. Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* **60**(1): 1-18.

Hu Y, Willer C, Zhan X, Kang HM, Abecasis GR. 2013. Accurate local-ancestry inference in exome-sequenced admixed individuals via off-target sequence reads. *Am J Hum Genet* **93**(5): 891-899.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**(2): 337-338.

Jakobsdottir J, Conley YP, Weeks DE, Mah TS, Ferrell RE, Gorin MB. 2005. Susceptibility genes for age-related maculopathy on chromosome 10q26. *Am J Hum Genet* **77**(3): 389-407.

Jozsi M, Heinen S, Hartmann A, Ostrowicz CW, Halbich S, Richter H, Kunert A, Licht C, Saunders RE, Perkins SJ et al. 2006. Factor H and atypical hemolytic uremic syndrome: mutations in the C-terminus cause structural changes and defective recognition functions. *Journal of the American Society of Nephrology : JASN* **17**(1): 170-177.

Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. 2012. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *Am J Hum Genet* **91**(5): 839-848.

Kang HM, Sul JH, Zaitlen NA, Kong S-y, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**(4): 348-354.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. *Genetics* **178**(3): 1709-1723.

Kelley CT. 1999. *Iterative methods for optimization*. Siam.

Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ et al. 2009. A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**(7258): 1011-1015.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**(5720): 385-389.

Le SQ, Durbin R. 2011. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome research* **21**(6): 952-960.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G et al. 2007. The diploid genome sequence of an individual human. *PLoS biology* **5**(10): e254.

Li B, Leal SM. 2009. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS genetics* **5**(5): e1000481.

Li B, Zhan X, Wing M-K, Anderson P, Kang HM, Abecasis GR. 2013. QPLOT: A Quality Assessment Tool for Next Generation Sequencing Data. *BioMed research international* **2013**.

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

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**(11): 1851-1858.

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

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. 2009a. SNP detection for massively parallel whole-genome resequencing. *Genome research* **19**(6): 1124-1132.

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. 2011. Low-coverage sequencing: implications for design of complex trait association studies. *Genome research* **21**(6): 940-951.

Li Y, Willer C, Sanna S, Abecasis G. 2009b. Genotype imputation. *Annual review of genomics and human genetics* **10**: 387.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology* **34**(8): 816-834.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**(10): 833-835.

Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA et al. 2010. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* **362**(13): 1181-1191.

Maller JB, Fagerness JA, Reynolds RC, Neale BM, Daly MJ, Seddon JM. 2007. Variation in complement factor 3 is associated with risk of age-related macular degeneration. *Nat Genet* **39**(10): 1200-1201.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**(7265): 747-753.

Manuelian T, Hellwage J, Meri S, Caprioli J, Noris M, Heinen S, Jozsi M, Neumann HP, Remuzzi G, Zipfel PF. 2003. Mutations in factor H reduce binding affinity to C3b and heparin and surface attachment to endothelial cells in hemolytic uremic syndrome. *The Journal of clinical investigation* **111**(8): 1181-1190.

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**(7): 499-511.

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.

Mardis ER. 2011. A decade/'s perspective on DNA sequencing technology. *Nature* **470**(7333): 198-203.

Mathieson I, McVean G. 2012. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* **44**(3): 243-246.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**(9): 1297-1303.

McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, Olson JJ, Mikkelsen T, Lehman N, Aldape K. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**(7216): 1061-1068.

Menozzi P, Piazza A, Cavalli-Sforza L. 1978. Synthetic maps of human gene frequencies in Europeans. *Science* **201**(4358): 786-792.

Metzker ML. 2009. Sequencing technologies—the next generation. *Nature Reviews Genetics* **11**(1): 31-46.

Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin U-M, Saad M, Simon-Sanchez J, Schulte C, Lesage S, Sveinbjörnsdóttir S. 2011. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* **377**(9766): 641-649.

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**(5925): 387-389.

Nelder JA, Mead R. 1965. A Simplex Method for Function Minimization. *The Computer Journal* **7**(4): 308-313.

Nelis M, Esko T, Magi R, Zimprich F, Zimprich A, Toncheva D, Karachanak S, Piskackova T, Balascak I, Peltonen L et al. 2009. Genetic structure of Europeans: a view from the North-East. *PLoS One* **4**(5): e5472.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature reviews Genetics* **12**(6): 443-451.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR et al. 2008. Genes mirror geography within Europe. *Nature* **456**(7218): 98-101.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics* **40**(5): 646-649.

O'Connell J, Gurdasani D, Delaneau O, Marchini J. 2014. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS genetics* **in press**.

O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K. 2012. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**(6114): 1619-1622.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**(8): 904-909.

Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**(7): 459-463.

Pritchard JK, Rosenberg NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. *The American Journal of Human Genetics* **65**(1): 220-228.

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

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000b. Association mapping in structured populations. *The American Journal of Human Genetics* **67**(1): 170-181.

Priya RR, Chew EY, Swaroop A. 2012. Genetic Studies of Age-related Macular Degeneration: Lessons, Challenges, and Opportunities for Disease Management. *Ophthalmology* **119**(12): 2526-2536.

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. 2010. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**(18): 2336-2337.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**(3): 559-575.

Ratnapriya R, Swaroop A. 2013. Genetic architecture of retinal and macular degenerative diseases: the promise and challenges of next-generation sequencing. *Genome Medicine* **5**(10): 84.

Raychaudhuri S, Iartchouk O, Chin K, Tan PL, Tai AK, Ripke S, Gowrisankar S, Vemuri S, Montgomery K, Yu Y et al. 2011. A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. *Nat Genet* **43**(12): 1232-1236.

Rosenbaum PR, Rubin DB. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70**(1): 41-55.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. *Science* **298**(5602): 2381-2385.

Schönemann PH, Carroll RM. 1970. Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika* **35**(2): 245-255.

Sekhon JS, Mebane WR. 1998. Genetic optimization using derivatives. *Political Analysis* **7**(1): 187-210.

Sforza LLC, Bodmer WF. 1999. *The genetics of human populations*. DoverPublications. com.

Shanno DF, Phua K-H. 1980. Remark on "Algorithm 500: Minimization of unconstrained multivariate functions [e4]". *ACM Transactions on Mathematical Software (TOMS)* **6**(4): 618-622.

Shu XO, Gao YT, Tu JT, Zheng W, Brinton LA, Linet MS, Fraumeni JF. 1988. A population‐based case‐control study of childhood leukemia in shanghai. *Cancer* **62**(3): 635-644.

Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* **38**(11): 1348-1354.

Skoglund P, Malmstrom H, Raghavan M, Stora J, Hall P, Willerslev E, Gilbert MT, Gotherstrom A, Jakobsson M. 2012. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336**(6080): 466-469.

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS. 2012. Rapid variance components-based method for whole-genome association analysis. *Nature genetics*.

Swaroop A, Chew EY, Rickman CB, Abecasis GR. 2009. Unraveling a multifactorial late-onset disease: from genetic susceptibility to disease mechanisms for age-related macular degeneration. *Annual review of genomics and human genetics* **10**: 19-43.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**(6090): 64-69.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319): 1061-1073.

The 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422): 56-65.

The AREDS2 Research Group, Chew EY, Clemons T, Sangiovanni JP, Danis R, Domalpally A, McBee W, Sperduto R, Ferris FL. 2012. The Age-Related Eye Disease Study 2 (AREDS2): study design and baseline characteristics (AREDS2 report number 1). *Ophthalmology* **119**(11): 2282-2289.

Tian C, Kosoy R, Nassir R, Lee A, Villoslada P, Klareskog L, Hammarstrom L, Garchon HJ, Pulver AE, Ransom M et al. 2009. European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. *Molecular medicine* **15**(11-12): 371-383.

Tringe SG, Von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC. 2005. Comparative metagenomics of microbial communities. *Science* **308**(5721): 554-557.

Victor RG, Haley RW, Willett DL, Peshock RM, Vaeth PC, Leonard D, Basit M, Cooper RS, Iannacchione VG, Visscher WA et al. 2004. The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *The American journal of cardiology* **93**(12): 1473-1480.

Wang C, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, Hardy JA, Singleton AB, Rosenberg NA. 2010. Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Statistical applications in genetics and molecular biology* **9**(1): Article 13.

Wang C, Zhan X, Bragg-Gresham J, Stambolian D, Chew EY, Fulton R, Wilson RK, Mardis ER, Swaroop A, Zöllner S et al. 2014. Estimating Individual Ancestry Using Next Generation Sequencing. *Nat Genet*.

Wang C, Zöllner S, Rosenberg NA. 2012. A Quantitative Comparison of the Similarity between Genes and Geography in Worldwide Human Populations. *PLoS genetics* **8**(8): e1002886.

Wang Y, Lu J, Yu J, Gibbs RA, Yu F. 2013. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome research* **23**(5): 833-842.

Weir BS, Cockerham C. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**(6): 1358-1370.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**(7145): 661-678.

Wetterstrand K. 2014. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**(7189): 872-876.

Wolk A, Mantzoros CS, Andersson S-O, Bergström R, Signorello LB, Lagiou P, Adami H-O, Trichopoulos D. 1998. Insulin-like growth factor 1 and prostate cancer risk: a population-based, case-control study. *Journal of the National Cancer Institute* **90**(12): 911-915.

Yang WY, Novembre J, Eskin E, Halperin E. 2012. A model-based approach for analysis of spatial structure in genetic data. *Nat Genet* **44**(6): 725-731.

Yates JR, Sepp T, Matharu BK, Khan JC, Thurlby DA, Shahid H, Clayton DG, Hayward C, Morgan J, Wright AF et al. 2007. Complement C3 variant and the risk of age-related macular degeneration. *N Engl J Med* **357**(6): 553-561.

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics* **38**(2): 203-208.

Zaykin DV, Kozbur DO. 2010. P - value based analysis for shared controls design in genome - wide association studies. *Genetic epidemiology* **34**(7): 725-738.

Zhan X, Larson DE, Wang C, Koboldt DC, Sergeev YV, Fulton RS, Fulton LL, Fronick CC, Branham KE, Bragg-Gresham J et al. 2013. Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nat Genet* **45**(11): 1375-1379.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**(24): 3326-3328.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44**(7): 821-824.