Curating and Combining Big Data from Genetic Studies

by

Sichen Liu

B.Sc., Simon Fraser University, 2018

Project Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in the Department of Statistics and Actuarial Science Faculty of Science

© Sichen Liu 2020

SIMON FRASER UNIVERSITY
Summer 2020

Copyright in this work rests with the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.
Approval

Name: Sichen Liu
Degree: Master of Science (Statistics)
Title: Curating and Combining Big Data from Genetic Studies

Examining Committee: Chair: Joan Hu
Professor

Jinko Graham
Senior Supervisor Professor via videoconferencing

Brad McNeney
Examiner Associate Professor via videoconferencing

Dave Campbell
Supervisor Professor via videoconferencing
School of Mathematics and Statistics
Carleton University

Date Defended: August 19, 2020
Abstract

Big data curation is often underappreciated by users of processed data. With the development of high-throughput genotyping technology, large-scale genome-wide data are available for genetic association analysis with disease. In this project, we describe a data-curation protocol to deal with the genotyping errors and missing values in genetic data. We obtain publicly-available genetic data from three studies in the Alzheimer’s Disease Neuroimaging Initiative (ADNI), and with the aid of the freely-available HapMap3 reference panel, we improve the quality and size of the ADNI genetic data. We use the software PLINK to manage data format, SHAPEIT to check DNA strand alignment and perform phasing of the genetic markers that have been inherited from the same parent, IMPUTE2 to impute missing SNP genotypes, and GTOOL to merge files and convert file formats. After merging the genetic data across these studies, we also use the reference panel to investigate the population structure of the processed data. ADNI’s participants are collected in the U.S, where the majority of the population are descendants of relatively recent immigrants. We use principal component analysis to understand the population structure of the participants, and model-based clustering to investigate the genetic composition of each participant and compare it with self-reported ethnicity information. This project is intended to serve as a guide to future users of the processed data.

Keywords: Big data processing; Alzheimer’s Disease Neuroimaging Initiative (ADNI); Genome-wide association study; Population structure; Principal Component Analysis (PCA); Model-Based Clustering (MBC)
Acknowledgements

My deepest gratitude goes first to my supervisor, Dr. Jinko Graham, for her inspiration and guidance through each stage of the process. Her passion in statistical genetics motivated my interest in learning the fascinating facts about statistical analysis on genetic data. Thanks to Dr. Brad McNeney and Dr. Dave Campbell for appreciating my work and giving me suggestions on my thesis. I would like to thank my parents in their understanding and endless support for all these years. To Payman Nickchi, Charith Karunarathna, Dongmeng Liu, and Pulindu Ratnasekera, thank you for your patience in explaining the genetic terminologies and statistical methodologies to me. I also would like to thank my graduate writing facilitator Eleonora Joensuu, I appreciate her detailed comments and writing suggestions. For all my friends in SFU, thank you for giving me the best memories of my life.
# Table of Contents

Approval ii  
Abstract iii  
Acknowledgements iv  
Table of Contents v  
List of Tables vii  
List of Figures viii  

1 Introduction 1  

2 Data and Software Summary 4  

2.1 Data Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4  

2.2 Software Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6  

3 Data Processing 9  

3.1 Inclusion Criteria . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11  
3.2 Quality Control . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11  
3.3 Pre-Phasing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12  
3.4 Phasing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13  
3.5 Imputation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16  
3.6 Merge and Format . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18  
3.7 Curated Data Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18  

4 Population Structure 20  

4.1 Principal Component Analysis . . . . . . . . . . . . . . . . . . . . . . . . . 20  
4.2 Model-Based Clustering . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24  

5 Conclusion 31  

Bibliography 33
## List of Tables

<table>
<thead>
<tr>
<th>Table 2.1</th>
<th>Data Summary of the ADNI Studies</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.2</td>
<td>Population samples in Hapmap3 <em>(HapMap 3 n.d.)</em></td>
<td>6</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Data Summary for ADNI-1 and ADNI-GO/2</td>
<td>18</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>HapMap3 clustering by population</td>
<td>27</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>ADNI clustering by self-reported race</td>
<td>28</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>ADNI clustering by self-reported ethnicity</td>
<td>29</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Data Pipeline</td>
<td>3</td>
</tr>
<tr>
<td>3.1</td>
<td>Genetic Data Processing Protocol</td>
<td>10</td>
</tr>
<tr>
<td>3.2</td>
<td>Ambiguity in Phasing (Leitwein et al. 2020)</td>
<td>14</td>
</tr>
<tr>
<td>3.3</td>
<td>A simple illustration of the haplotype estimation in SHAPEIT</td>
<td>15</td>
</tr>
<tr>
<td>3.4</td>
<td>Overview of imputation using IMPUTE2 (Howie &amp; Marchini 2014)</td>
<td>16</td>
</tr>
<tr>
<td>3.5</td>
<td>Venn diagram of type 0/1/2 SNPs</td>
<td>17</td>
</tr>
<tr>
<td>3.6</td>
<td>Venn Diagram of ADNI studies and HapMap3 before and after imputation.</td>
<td>19</td>
</tr>
<tr>
<td>4.1</td>
<td>Scree plot of Eigenvalues</td>
<td>22</td>
</tr>
<tr>
<td>4.2</td>
<td>PC plot of HapMap3 Subjects</td>
<td>23</td>
</tr>
<tr>
<td>4.3</td>
<td>PC plot of ADNI Subjects</td>
<td>24</td>
</tr>
<tr>
<td>4.4</td>
<td>Genetic composition of ADNI participants in selected races</td>
<td>28</td>
</tr>
<tr>
<td>4.5</td>
<td>Genetic composition of ADNI participants self-reporting as white, by their</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>self-reported ethnicity</td>
<td></td>
</tr>
<tr>
<td>A.1</td>
<td>Genetic compositions of HapMap3 samples in selected populations: red ASW,</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>gold YRI, green CEU/TSI, blue-green CHB/CHD/JPT, blue GIH/MEX, purple LWK,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pink M KK</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

With recent advanced genotyping technologies, scientists are able to obtain large-scale, genome-wide association data and investigate genetic association of human disease. Generally, the raw genetic data come directly from the genotyping machine, and so genotyping errors and missing genotypes cannot be neglected. These errors and missing genotypes occur due to several causes. First, the genotyping machine can fail to detect the mutation at a specific position on the genome; second, the sample quality can be compromised due to low quantity and quality of DNA extract or contamination of DNA extract; third, low-quality reagents or poor machine accuracy can lead to biochemical artefacts; and fourth, human error, which is the most common error, can occur when handlers mislabel or switch DNA samples, or contaminate samples with foreign DNA (Pompanon et al. 2005). Since the validity of genetic association testing depends on the accuracy of the genotype data, innovative methods are needed to deal with errors in genotyping and missing genotypes for the Single Nucleotide Polymorphisms (SNPs) used as genetic variables in these studies.

In this project, we describe a workflow as a partial remedy for genotyping errors and missing genotypes. Our workflow ranges from data cleaning to exploratory data analysis and can be viewed as part of the overall data pipeline illustrated in Figure 1.1. The steps of our workflow consist of data cleaning, imputing missing data, formatting data and performing exploratory data analysis. After applying these steps, the power of genome-wide association studies (GWAS) can be increased, and users can better understand the background of the processed data. Analysts often take for granted well-prepared data and are unaware of the procedures for preparing the raw data for analysis. In statistical reports, analysts try to interpret p-values or the results from other techniques for statistical inference, without understanding the limitations of the data. However, statistical inference is just the "tip of the iceberg". Figure 1.1 provides an overview of the entire set of procedures, from designing the study to obtaining summary statistics. Before applying any statistical models, analysts need to understand the design of the study, how the raw data are collected, and how the raw data are tidied and processed. Only then can analysts explore the processed data and select potential statistical models (Leek & Peng 2015).
The aim of our project is to curate and combine the raw genetic data for the three studies in the Alzheimer’s Disease Neuroimaging Initiative (ADNI), so that they can be analysed later for association with the imaging and clinical data. ADNI is a multi-site longitudinal study started in 2004, that was designed to collect clinical, imaging, genetic, and biochemical biomarker information of the participants (Alzheimer’s Disease Neuroimaging Initiative 2017). In achieving the objective to fix genotyping errors and combine the genetic data, we anticipate the following challenges. First, the size of the genetic data in each study is very large. The ADNI-1 data has 620,901 genetic variables and 757 samples, and the combined ADNI-GO and ADNI-2 data has 730,525 genetic variables and 793 samples. Second, in the ADNI-GO/2 studies, the genetic variables are defined differently than in the ADNI-1 study because they were processed on a newer high-throughput genotyping chip. As a result, only 306,243 genetic variables overlap between the studies, and yet we want to merge the ADNI-GO/2 genetic data with ADNI-1. Third, owing to genotyping failures or sample corruption, we need to remove some poor quality genetic variables and samples. Next, we have dependencies among the samples, but we want independently and identically distributed observations. Therefore, we need to remove relatives or sample duplicates from the dataset. Last but not least, we have significant amounts of missing data because the high-throughput genotyping chip failed to detect some genetic variables for some individuals.

Understanding the data is a crucial first step in any evidence-based analysis, including learning the decisions made by others to prepare and clean the raw data. This project considers two important stages of the data-curation process shown in the pipeline below: data cleaning and exploratory data analysis. In this project, we first introduce the data and software involved in the data processing. Next, we discuss the data-curation protocol that processes and combines the three genetic datasets from the ADNI studies. Finally, we perform exploratory data analysis to understand the processed data, and compare the participants’ self-reported ethnicities to their genetic information.
Figure 1.1: Data Pipeline
Chapter 2

Data and Software Summary

This project works with raw genetic data and subject information obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI). Alzheimer’s Disease (AD) is the primary cause of dementia in the elderly; patients lose mental function, and their brain tissue deteriorates as the disease progresses. The effect of this disease is irreversible, yet no cures have been discovered. Researchers around the world are trying to investigate and develop methods to slow or stop AD development. Therefore, a multi-site longitudinal study — the Alzheimer’s Disease Neuroimaging Initiative (ADNI) was founded in 2004 under the leadership of Dr. Michael W. Weiner (Alzheimer’s Disease Neuroimaging Initiative 2017). The initiative was designed to collect clinical, imaging, genetic, and biochemical biomarkers of the participants over time, to detect and track the development of AD. In this chapter, we introduce the data and software involved in our project.

2.1 Data Summary

The data used in this project involve three ADNI studies. The initial five-year study ADNI-1 collected data on 200 elderly, cognitively-normal (CN) subjects, 400 mildly cognitively-impaired (MCI) subjects, and 200 AD subjects. The study was then extended as ADNI-GO by another two years; besides the updates of the existing subjects from ADNI-1, another 200 early MCI subjects were examined. In 2011, another five-year study ADNI-2 was funded. Besides the original subjects in ADNI-1/GO, ADNI-2 recruited 150 CN elderly subjects, 100 subjects with early MCI, 150 subjects with late MCI, and 150 AD subjects. The latest ADNI-3 study started in 2016 and is a five-year on-going study (Alzheimer’s Disease Neuroimaging Initiative 2017).

For neuroimaging data, the participants had MRI and PET scans at either 6- or 12-month intervals and were followed up for either 2 or 3 years. In past genome-wide association studies (GWAS) for ADNI, researchers have used neuroimaging data to estimate the rates of change in cortical thickness and regional volumes over the progression of AD. Therefore, the imaging data is used as an inclusion criterion in our project (Szefer et al. 2017). Currently,
ADNI-1 and ADNI-GO/2 have genetic data available, so our project will only focus on these studies. As summarized in table 2.1, 757 ADNI-1 participants were genotyped on an Illumina Human610-Quad BeadChip with 620,901 SNPs and 793 ADNI-GO/2 participants were genotyped on an Illumina HumanOmniExpress BeadChip with 730,525 SNPs. Researchers can combine these genetic data with imaging and clinical data to investigate the association of imaging or clinical phenotypes (or traits) with specific genes.

**Table 2.1: Data Summary of the ADNI Studies**

<table>
<thead>
<tr>
<th>Study Characteristics</th>
<th>ADNI-1</th>
<th>ADNI-GO/2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cohort</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200 elderly CN;</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>400 MCI;</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>200 AD</td>
<td>200</td>
<td>400</td>
</tr>
<tr>
<td><strong>Total Subjects</strong></td>
<td>800</td>
<td>+ 600</td>
</tr>
<tr>
<td><strong>Genotyping Platform</strong></td>
<td>Illumina Human610-Quad BeadChip</td>
<td>Illumina HumanOmniExpress BeadChip</td>
</tr>
<tr>
<td><strong>Number of Subjects Genotyped</strong></td>
<td>757</td>
<td>793</td>
</tr>
<tr>
<td><strong>Number of SNPs</strong></td>
<td>620,901</td>
<td>730,525</td>
</tr>
</tbody>
</table>

ADNI’s genetic data are available in a set of binary files specific to the freely-available software PLINK that is used for processing it. Participants’ genotypes are stored in a .bed file as "00", "01", "02", "03". SNPs have only two alleles or variant forms and we inherit an allele from each parent. Without loss of generality, label the first allele of a SNP as "A" and the second as "B". Then, for each participant at a specific SNP along the genome, "00" represents missing genotype, "01" represents homozygous for the first allele "A/A", "02" represents heterozygous "A/B", and "03" represents homozygous for the second allele "B/B". The accompanying file .bim contains information on the chromosome code, SNP ID, the genetic distance of the SNP from the end of the chromosome, the base-pair position of the SNP, the minor allele, and the major allele for each SNP. The accompanying .fam file records the participants’ information such as family ID, individual ID, whether the father or mother was sampled in the data, sex code, and phenotype value.

In addition, an R package ADNIMERGE (https://adni.bitbucket.io) was used for data processing and inference of population structure. The ADNIMERGE package contains demographic, clinical, and imaging data for the participants across the ADNI-1, ADNI-GO, ADNI-2 and ADNI-3 studies. The combined dataset has 14,461 rows for each visit/participant combination, and 113 variables in columns. In our project, the main variables we used were "PTID", "ORIGPROT", "DX.bl", "VISCODE", "PTETHCAT", "PTR-ACCAT", which represent patient ID, protocol from which subject originated (ADNI-1/GO/2/3), disease baseline, visit code, self-reported ethnicity, and self-reported race, respectively. Disease baseline records the patient’s initial diagnosis, and visit code records the baseline visit and consecutive visits in number of months from the first visit. At the stage of inclusion criteria, we also updated the patients’ disease status in ADNI-1 based on the CONVERSION file on
the ADNI website, and found that one person had converted from CN (Cognitively normal) to LMCI (Low mild cognitively impairment) during the study period.

To impute the missing genotypes in our genetic data, we used the publicly-available HapMap3 reference panel based on the NCBI build 36/hg18 of the human genome. This reference panel can be downloaded from https://mathgen.stats.ox.ac.uk/impute/data_download_hapmap3_r2.html. HapMap3 contains 1,301 DNA samples with 1,440,616 SNPs from a variety of human populations. The samples were collected from populations with ancestry from parts of Africa, Asia, America, and Europe (HapMap 3 n.d.). Table 2.2 shows the number of samples collected from different populations. HapMap3 is preferred over other publicly-available reference panels because more SNPs overlap between ADNI-1 and HapMap3. In this project, the HapMap3 haplotypes (sequences of SNPs inherited together from the same parent) were used to fill in the missing genotypes in the ADNI studies during imputation. The HapMap3 genotype data were also used for inferring population structure in Chapter 4. The HapMap3 haplotype data were stored in a compressed .haps file in which the rows correspond to SNPs and the columns correspond to attributes of the SNPs. Specifically, the first five columns of the .haps file are the chromosome number, SNP ID, base-pair position, minor and major allele. The following columns come in pairs, with each pair representing the two haplotypes of a given study participant. The corresponding .legend file also has SNPs as rows and has four columns for the SNP ID, base-pair position, minor and major alleles. The genotype information and sample information for HapMap3 is in the form of PLINK binary files.

Table 2.2: Population samples in Hapmap3 (HapMap 3 n.d.)

<table>
<thead>
<tr>
<th>Population</th>
<th>Representation</th>
<th>Number of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASW</td>
<td>African ancestry in Southwest USA</td>
<td>48</td>
</tr>
<tr>
<td>CEU</td>
<td>Utah residents with Northern and Western European ancestry from the CEPH collection</td>
<td>111</td>
</tr>
<tr>
<td>CHB</td>
<td>Han Chinese in Beijing, China</td>
<td>84</td>
</tr>
<tr>
<td>CHD</td>
<td>Chinese in Metropolitan Denver, Colorado</td>
<td>84</td>
</tr>
<tr>
<td>GIH</td>
<td>Gujarati Indians in Houston, Texas</td>
<td>85</td>
</tr>
<tr>
<td>JPT</td>
<td>Japanese in Tokyo, Japan</td>
<td>86</td>
</tr>
<tr>
<td>LWK</td>
<td>Luhya in Webuye, Kenya</td>
<td>84</td>
</tr>
<tr>
<td>MEX</td>
<td>Mexican ancestry in Los Angeles, California</td>
<td>47</td>
</tr>
<tr>
<td>MKK</td>
<td>Maasai in Kinyawa, Kenya</td>
<td>135</td>
</tr>
<tr>
<td>TSI</td>
<td>Toscani in Italia</td>
<td>88</td>
</tr>
<tr>
<td>YRI</td>
<td>Yoruba in Ibadan, Nigeria</td>
<td>112</td>
</tr>
</tbody>
</table>

2.2 Software Summary

As genetic datasets are very large, handling them requires specialized software to manage and merge data, estimate haplotypes from genotypes, impute missing genotypes inferred from haplotypes, and convert file formats. The following software tools are crucial to our data-curation process, and we give a brief review here.
PLINK

PLINK is an open-source toolset that can handle large-scale genetic data analyses efficiently. We used PLINK to manage and summarize our genetic data. PLINK plays an essential role during quality control of genome-wide data, as it has tools to calculate the rate of missing genotype data for individuals and SNPs, SNP allele frequencies, and test statistics to detect non-random mating, inbreeding, hidden relatedness, and errors in SNP genotyping or reported sex. Throughout this project, we used PLINK to read the genetic data in PLINK file format, then recoded, reordered, split, extracted, and merged data. For genomic quality control, we used PLINK to calculate missing genotype rates, minor allele frequencies, summaries of identity-by-descent and test statistics for Hardy–Weinberg equilibrium (non-random mating) and sex checks based on X-chromosome SNPs.

SHAPEIT

SHAPEIT is free software for academic use and only available on the Linux platform. For imputation purposes, we need to estimate the haplotypes from the genotype data. SHAPEIT is the abbreviation of “Segmented Haplotype Estimation and Imputation Tool”. This software estimates haplotypes, also referred to as phasing, in a precise and computationally efficient manner. In this project, we used SHAPEIT to check strand alignment between our genetic data and HapMap3 reference panel. The software also estimated haplotypes from genotypes in our data and with the help of the reference panel. For large-scale, genome-wide datasets, the phasing methods implemented in SHAPEIT have a linear complexity in the number of SNPs.

IMPUTE2

IMPUTE2 is known for its flexible and accurate genotype imputation in genome-wide association studies (GWAS). This software is available for free download over different platforms such as Linux, Mac, Windows, and Solaris. We input the haplotypes of our study from phasing and the haplotypes from the reference panel to IMPUTE2. The software can find shared segments of recent common ancestry between the study panel and reference panel, then use these segments to impute the missing alleles and output the resulting genotypes.

GTOOL

GTOOL works closely with the imputed genotype data from IMPUTE2. We need GTOOL to convert the file format of the imputed and non-imputed data, merge the genotype datasets, and then output them as the PLINK formats. GTOOL is open-source and available for download on the Linux and Mac platforms.
These software packages are essential to data processing. As the size of GWAS data is very large, software like SHAPEIT and IMPUTE2 are much more computationally efficient in handling phasing and imputation. PLINK and GTOOL can effectively manage the format of data, and the conversions between file types. We will discuss the details of how the software aids our data curation protocol in the next chapter.
Chapter 3

Data Processing

As discussed in the introduction, our challenge is to build a protocol to curate and combine the genetic data in the ADNI studies (ADNI-1 and ADNI-GO/2), so that these data can be used in later genetic analysis of imaging data. The protocol for genetic-data processing consists of six steps as illustrated in Figure 3.1. In this chapter, we will discuss the details of the data-processing protocol section-by-section. First, the inclusion criteria reduce the size of the data; second, genomic quality control removes poor samples and SNPs; third, pre-phasing prepares the data for haplotype estimation or phasing; fourth, phasing estimates the haplotypes of each individual with the aid of the HapMap3 reference panel; fifth, genotype imputation fills in the missing genotypes by inferring them based on the haplotypes in both the study panel and the reference panel; finally, merging the imputed and non-imputed data, and combining the genetic data between the ADNI studies increases the power of subsequent analyses.
Figure 3.1: Genetic Data Processing Protocol

**Inclusion Criteria**
- At least a baseline MRI scan, and one follow-up scan

**Genomic Quality Control**
1. Remove individuals with genotyping rate < 90% and SNPs with call rate < 95%
2. Remove SNPs with MAF < 0.01 and HWE-test p-value < $10^{-6}$
3. Remove individuals who failed sex check
4. Remove first-degree relatives or duplicate samples
5. Filter out sex chromosome SNPs

**Pre-Phasing**
1. Split data by chromosome
2. Remove duplicate SNPs
3. Check strand alignment
4. Flip SNPs

**Phasing**
- Estimate the haplotypes of the study panel based on the reference panel

**Imputation**
- Impute the missing genotypes of the study panel based on the haplotypes of the study panel and reference panel

**Merge and Format**
1. Merge imputed data with the study-panel-only SNPs in Pre-Phasing, change the data format back to PLINK files.
2. Take the intersection of SNPs between the three studies and the reference panel, then merge the SNPs across these studies together.
3.1 Inclusion Criteria

As discussed in Chapter 3 (Data and Software Summary), 757 subjects from ADNI-1 and 793 subjects from ADNI-GO/2 have genetic data available. In this section, only the subjects with imaging data available are selected for further curation. In ADNI-1, 696 out of 757 subjects met our inclusion criteria of having a baseline MRI scan and at least one follow-up scan. In ADNI-GO/2, 538 out of 793 subjects met the inclusion criteria.

3.2 Quality Control

The fundamental goal of a genetic-association study is to find SNPs associated with disease susceptibility. The common steps for quality control of genomic data were applied to both the ADNI-1 and ADNI-GO/2 datasets:

1. We removed the samples and SNPs with poor quality. As mentioned in the introduction, genotyping and sample-processing errors can be misleading in genetic-association studies. The individuals with genotyping rate less than 90%, and SNPs with call rate less than 95% are considered as poor quality (Anderson et al. 2010). We removed 42,800 SNPs from ADNI-1 and 17,311 SNPs from ADNI-GO/2. No samples were removed since all the individuals had genotyping rate over 90%.

2. We excluded SNPs with sample Minor Allele Frequency (MAF) less than 0.01, because SNPs are chosen to have population MAF > 0.01 and a low sample MAF is consistent with genotyping errors (Anderson et al. 2010). We removed 21,538 SNPs from ADNI-1, and 63,319 SNPs from ADNI-GO/2. We also excluded SNPs whose sample genotype proportions were inconsistent with random mating. To assess the assumption of random mating or Hardy-Weinberg equilibrium (HWE), we used a Fisher’s exact test of goodness of fit at level 10^-6 (Purcell et al. 2007). SNPs with sample genotype proportions that depart from the expected genotype proportions under HWE are interpreted to be of poor quality and removed from further analysis (Anderson et al. 2010). We removed 209 SNPs from ADNI-1 and 1,151 SNPs from ADNI-GO/2.

3. We checked whether the self-reported gender is consistent with SNPs on the X-chromosome and removed any inconsistent individuals. Males have an X- and a Y-chromosome, whereas females have two X-chromosomes. Since only the X-chromosome carries genetic information, males are referred to as hemizygous. In males, automated genotype-calling algorithms tend to interpret genotypes for SNPs on the X-chromosome as homozygous. For SNPs on the X-chromosome, high homozygosity rates are consistent with a subject being male, whereas low homozygosity rates are consistent with a subject being female. The homozygosity rate of an individual is their proportion of homozygous SNPs (Anderson et al. 2010). In PLINK, homozygosity rates > 0.8 and <
are interpreted to be consistent with the male and female sex, respectively (Purcell et al. 2007). In ADNI-1, two subjects had ambiguous homozygosity rates between 0.2 and 0.8 and so they were removed.

4. To ensure that study samples are independent, we removed suspected first-degree relatives and duplicate samples in the data. Pairs of the first-degree relatives such as parents, children or siblings are identified through their identity-by-descent (IBD) values, defined as the number of identical copies inherited from a common ancestor. Of specific interest is the coefficient of relatedness of a pair of individuals, which is defined to be \( P(\text{IBD} = 2) + 0.5 \times P(\text{IBD} = 1) \). In pairs of first-degree relatives, this coefficient is 0.5 (Anderson et al. 2010). We declare a pair of individuals to be first-degree relatives or sample duplicates if their estimated coefficient of relatedness is \( \geq 0.5 \) (Purcell et al. 2007). Each study has 3 pairs of declared first-degree relatives or sample duplicates, and so we randomly removed one individual from each pair.

5. Finally, we filtered out the sex chromosome SNPs because their genetic transmission differs from the transmission of SNPs on non-sex chromosomes (collectively known as the autosomes, or chromosomes 1-22).

After the above steps, we have 521,059 SNPs left in ADNI-1 and 648,744 SNPs left in ADNI-GO/2. In ADNI-1, two subjects of ambiguous sex were removed and 3 random individuals from 3 pairs of declared first-degree relatives or sample duplicates were removed, leaving 691 subjects. In ADNI-GO/2, 3 random individuals from 3 pairs of declared first-degree relatives or sample duplicates were removed, leaving 535 subjects.

The genomic positions and names of the SNPs in the ADNI-1 study and the HapMap3 reference panel are with respect to the same (older) build of the human genome, hg18. By contrast, the names and positions of the ADNI-GO/2 SNPs are from a newer build of the human genome, hg19. We performed a "LiftOver" step for the ADNI-GO/2 genotypes, to match the SNP name (variant ID) and genomic position to the names and positions in the older genome build. LiftOver is a software tool that converts the SNP names and genome coordinates between genotyping platforms that use different builds of the human genome (Zhan 2015). New SNPs in ADNI-GO/2 that were not in the older genome build used for the ADNI-1 or HapMap3 genotyping platforms are lost. After LiftOver, we have 632,073 SNPs left in ADNI-GO/2.

3.3 Pre-Phasing

Due to limitations in genotyping technologies, missing genotype values in genetic data are unavoidable. There are two kinds of missing genotypes, sporadically missing and completely missing. For sporadically missing genotypes, the SNPs for some study individuals fail to be genotyped; for completely missing genotypes, the chip does not assay a given SNP and so
no subjects have genotypes for it. In order to avoid missing data problems and increase the power of analysis, we will be using a reference panel for imputation of the missing SNPs. Before the actual imputation, the genetic data for both studies needs to be pre-phased and then phased. The following steps prepare our data for phasing:

1. For both ADNI-1 and ADNI-GO/2, we split the data by each chromosome, so we can perform analysis on chromosomes 1 to 22 separately. The purpose of splitting the data into chromosomes is to work with smaller chunks of data for computational efficiency.

2. Check for duplicate SNPs that share the same base-pair position in each chromosome, and then randomly remove one of the duplicates to ensure we do not have repeated variables.

3. Check the strand alignment of SNPs in our data with the HapMap3 reference panel. DNA is double-stranded and the top and bottom strands are oriented in opposite directions. Sometimes we do not know which direction the chip is reading the double-stranded DNA when it detects a SNP. The direction that the study chip reads the SNP may differ from the direction of the reference-panel chip, leading to different variants recorded for that SNP in the study versus the reference panel. We need to flip the SNPs with such strand issues in our data to match their strand alignments in the reference panel.

4. The SHAPEIT software cannot phase SNPs that are not in the reference panel and returns an error. Therefore, we remove the study-panel-only SNPs. Note that we do not discard these study-panel-only SNPs; they are retained and inserted in the data after the imputation. We will discuss this point in section 3.6, Merge and Format.

After pre-phasing, we have 493,743 SNPs in ADNI-1 and 489,239 SNPs in ADNI-GO/2 which are also in the reference panel. To further increase the size of data, we also retain the study-panel-only SNPs to merge them with the imputed SNPs later on.

### 3.4 Phasing

Phasing is also referred to as haplotype estimation because an individual’s haplotypes, or strings of SNPs inherited from the same parent, are inferred based on their genotypes. We inherit two haplotypes or sequences of alleles together from our parents, and the combination of these alleles forms genotypes. The point of phasing is to convert our study data from genotypes to haplotypes, so that we can improve imputation efficiency in the next step. However, phasing can be ambiguous. If we have $N$ heterozygous SNP genotypes, then there are $2^N$ possible haplotypes and $2^{N-1}$ possible pairs of haplotypes that are consistent with the $N$ heterozygous SNP genotypes. In Figure 3.2, there are variations in the unphased sequence
at SNP-1: C/G, SNP-2: T/C, and SNP-3: A/C. Thus, the haplotype estimation based on these three SNPs is ambiguous. There are four possible pairs of unlabeled haplotypes that are consistent with the genotypes. However, if we have genetic information about the individual’s relatives, then we would be able to estimate the individual’s haplotypes more accurately. In the diagram below, the second unlabeled pair of haplotypes (from the middle row labeled as "possible haplotypes") is shown as the paternal and maternal haplotypes at the bottom.

SHAPEIT uses a novel approach based on a graphical model of the space of all possible haplotypes that are consistent with an individual’s genotypes. The software infers the pair of haplotypes that is most consistent with the genotypes by matching the current guesses of haplotypes with all known haplotypes in the model, and then filling in the gaps in-between the study-panel SNPs with the SNPs found on the fully-sequenced reference haplotypes. For each target genotype, the two haplotypes that are the most consistent with the target genotype are selected using hidden Markov models. SHAPEIT’s algorithm starts by partitioning a set of known haplotypes in the sample, represented by $H$, into non-overlapping consecutive segments. A graph $H_g$ represents the distinct haplotypes in $H$ as segments. In Figure 3.3, $H$ contains $K = 8$ haplotypes in the rows of panel a and the haplotypes are split into two segments. The first segment contains SNPS 1 to 4 and the second segment contains SNPS 5 to 8. The graph $H_g$ represents the $J = 3$ distinct haplotypes in these two segments. Each path through $H_g$ represents a distinct haplotype in $H$. For example, in panel a of the figure, the highlighted haplotype in $H$ has a corresponding path through $H_g$ coloured in pink. When we sample a template haplotype from $H$, it is the same as sampling a path through $H_g$. In $H_g$, the alleles 0 or 1 are represented as nodes, and the edges connecting these nodes are weighted by the frequency of the haplotype in $H$ whose paths traverse them.
In panel a, $H$ represents all known haplotypes in the sample, and $H_g$ is a graph that represents all the distinct haplotypes from $H$ in segments. In panel b, $G$ represents an individual’s genotypes, and $S_g$ represents their possible haplotypes, in segments. (Delaneau et al. 2012).

Suppose we have an individual’s genotype $G$ as illustrated in panel b, which contains four heterozygous SNPs defined over $M = 8$ markers. Then the individual-level graph $S_g$ is built of the template haplotypes sampled from $H_g$ in segments that are consistent with $G$. In Figure 3.3, $G$ is split into two segments, each of which contains two heterozygous markers (both with genotypic value $G = 1$, which represents A/B), and has four possible haplotypes that are consistent with $G$. Typically, SHAPEIT splits an individual’s observed genotypes into genomic segments of three heterozygous markers, such that each segment has 8 possible haplotypes. The individual’s haplotypes are estimated from the available haplotypes in the graph $H_g$. In practice, SHAPEIT uses HMM-based methods to sample haplotypes that are consistent with $G$ from $H_g$. In these HMMs, the hidden states are the template haplotypes which are modelled by a Markov chain. The template haplotypes are not directly observed, but rather are a subset of haplotypes that is consistent with the individual’s genotype. When sampling a pair of paths through for the haplotypes of an individual, the transition probabilities determine how the path moves from one chromosomal position to the next, and the emission probabilities link the unobserved states to the target genotype. The observed target genotype is obtained by combining the final pairs of the inferred haplotypes (Li & Stephens 2003). In Figure 3.3, the final pair of paths in $S_g$ that is compatible with genotype $G$ is coloured in blue and green. With the aid of the distinct populations in the HapMap3 reference panel, the accuracy of phasing and imputation has increased substantially (Delaneau et al. 2012).
3.5 Imputation

Imputation methods predict the unobserved genotypes in a study sample. These missing genotypes are predicted with the help of the observed genotypes of individuals in a reference panel at a dense set of SNPs. The purpose of this project is to fill in as many missing genotype values as possible in both the ADNI-1 and ADNIGO/2 study data, so that the power of later analyses of the combined data is increased. In this project, we used the HapMap3 reference panel to impute sporadically missing genotypes in the study sample and also completely missing genotypes for SNPs that aren’t on the study’s genotyping chip. Figure 3.4 provides an overview of the SNPs to be imputed. In the figure, sporadically missing genotypes are for SNPs that have been genotyped in the study panel but have missing values for some subjects, whereas completely missing genotypes are for SNPs that are not genotyped in the study panel at all. The software IMPUTE2 imputes missing genotypes by estimating their marginal posterior probabilities given the observed genotype data in the study and in the reference panel. The IMPUTE2 documentation refers to SNPs that are genotyped and not genotyped in the study panel as type 2 and type 0 SNPs, respectively (Howie & Marchini 2014). In figure 3.5, type 2 SNPs are in both the study panel and the reference panel, type 0 SNPs are in the reference panel only, and type 1 SNPs are in the study panel only. Note that type 1 SNPs are removed in the pre-phasing step, and merged later on with the imputed SNPs. For the ADNI1 study, there are 870,345 type 0 SNPs in the HapMap3 reference panel.
only; 48,824 type 1 SNPs in the ADNI1 study panel only; and 516,988 type 2 SNPs in both the ADNI1 study panel and the HapMap3 reference panel. For the ADNI-GO/2 study, there are 848,508 type 0 SNPs; 142,784 type 1 SNPs and 538,730 type 2 SNPs.

Figure 3.5: Venn diagram of type 0/1/2 SNPs

If samples have not been pre-phased, the IMPUTE2 algorithm has two steps, which are iterated in a Markov chain Monte Carlo (MCMC) framework. If the study data has been pre-phased, the algorithm skips step 1. In step 1, IMPUTE2 samples a new haplotype pair for an individual based on the known haplotypes in both the study panel and the reference panel that are the most consistent with the individual’s observed genotype. For each individual, the haplotype pair is sampled conditional on the individual’s observed genotype, the current guesses of haplotypes, and the recombination rate between SNPs. This sampling technique is the same as the one used in SHAPEIT by sampling a pair of paths through the hidden states (template haplotypes) of the HMM model. In step 2, IMPUTE2 estimates the marginal posterior probabilities of the missing alleles by matching the incomplete haplotype pairs from step 1 with the fully-sequenced haplotypes in the reference panel. In this second step, the state space of the HMM includes only the reference-panel haplotypes, and the gaps between the SNPs in the study panel are filled in with the matching alleles on the reference-panel haplotypes. Next, the software converts the estimated allelic probabilities to genotypic probabilities for each individual (Howie et al. 2009). Steps 1 and 2 are iterated in the MCMC algorithm. In this project, we set a threshold of 0.9 for the MCMC estimates of the posterior genotypic probabilities. We retain the inferred genotypes with marginal posterior probabilities $\geq 0.9$, and consider the genotypes that did not pass the threshold as missing. After imputation, we have 1,387,333 SNPs in ADNI-1, 1,387,238 SNPs in ADNI-GO/2.
3.6 Merge and Format

In this section, we discuss the use of the software GTOOL. We used GTOOL to convert between file formats and to merge the imputed data with the non-imputed data that contains the study-panel-only SNPs. Recall that in the pre-phasing step we remove the study-panel-only SNPs because SHAPEIT cannot phase SNPs unless they are in the reference panel. Referring to Figure 3.6, we have only 516,988 SNPs in ADNI-1 shared with HapMap3, and 538,730 SNPs in ADNI-GO/2 shared with HapMap3. We imputed sporadically missing genotypes for these SNPs shared by the study panel and the reference panel. However, to retain the data on the study-panel-only SNPs that cannot be imputed, we merge them with the imputed data using GTOOL. Specifically, we use GTOOL to convert the PLINK files that contain the study-panel-only SNPs into the GEN and SAMPLE format files used by the IMPUTE2 software. Then we merge the study-panel-only SNPs in GEN and SAMPLE format with the imputed SNPs in GEN and SAMPLE format. After merging, we have 1,436,157 SNPs in ADNI-1, and 1,530,022 SNPs in ADNI-GO/2. Next, we use GTOOL to convert the GEN and SAMPLE files back to the PLINK format. Finally, we use PLINK to merge the 22 chromosomes for each study into a single dataset to conduct further analysis.

3.7 Curated Data Summary

In the original studies, the ADNI-1 data has 757 individuals and 620,901 SNPs, and the combined ADNI-GO and ADNI-2 data has 793 individuals and 730,525 SNPs. After applying the inclusion criteria, the quality control steps, the imputation and the merging with the study-panel-only SNPs, we have 691 individuals and 1,436,157 SNPs in ADNI-1, and 535 individuals and 1,530,022 SNPs in ADNI-GO/2. A total of 66 individuals in ADNI-1 and 258 individuals in ADNI-GO/2 are lost because their samples did not pass the inclusion criteria and quality control steps. Table 3.1 shows the detailed changes in number of individuals and SNPs in each step.

<table>
<thead>
<tr>
<th></th>
<th>ADNI-1</th>
<th>ADNI-1</th>
<th>ADNI-GO/2</th>
<th>ADNI-GO/2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of Subjects</td>
<td>Number of SNPs</td>
<td>Number of Subjects</td>
<td>Number of SNPs</td>
</tr>
<tr>
<td>Original</td>
<td>757</td>
<td>620,901</td>
<td>793</td>
<td>730,525</td>
</tr>
<tr>
<td>Inclusion Criteria</td>
<td>691</td>
<td>(-42,800)</td>
<td>691</td>
<td>(-700)</td>
</tr>
<tr>
<td>Samples with call rate &lt; 90%</td>
<td>(0)</td>
<td>535</td>
<td>(-1,151)</td>
<td></td>
</tr>
<tr>
<td>SNPs with call rate &lt; 95%</td>
<td>(-21,538)</td>
<td>(-16,671)</td>
<td>(-43,319)</td>
<td></td>
</tr>
<tr>
<td>HWE $p &lt; 10^{-6}$</td>
<td>(-1,209)</td>
<td>(-0)</td>
<td>(-1,151)</td>
<td></td>
</tr>
<tr>
<td>MAF &lt; 0.01</td>
<td>(-0.01)</td>
<td>(-0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remove subjects with unidentifiable sex</td>
<td>(-2)</td>
<td>(-0)</td>
<td>(-16,671)</td>
<td></td>
</tr>
<tr>
<td>Remove relatives and sample duplicates</td>
<td>(-3)</td>
<td>(-3)</td>
<td>(-16,671)</td>
<td></td>
</tr>
<tr>
<td>LiftOver</td>
<td>-</td>
<td>(-0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>After QC</td>
<td>691</td>
<td>521,059</td>
<td>535</td>
<td>632,073</td>
</tr>
<tr>
<td>After imputation</td>
<td>691</td>
<td>1,387,333</td>
<td>535</td>
<td>1,387,258</td>
</tr>
<tr>
<td>After merging with study-panel-only SNPs</td>
<td>691</td>
<td>1,436,157</td>
<td>535</td>
<td>1,530,022</td>
</tr>
</tbody>
</table>
The purpose of this project is to combine the genetic data of ADNI-1 and ADNI-GO/2, for improved power of future genetic analyses of the imaging data. After the quality-control steps but before imputation, we have only 342,603 SNPs shared by the ADNI studies. After imputation, the number of shared SNPs between the ADNI studies increases markedly, to 1,409,835, as shown in Figure 3.6.

**Figure 3.6:** Venn Diagram of ADNI studies and HapMap3 before and after imputation.

iADNI-1 and iADNI-GO/2 denote the merged and imputed ADNI-1 data respectively.

One crucial step that we have omitted is the quality control of SNPs that have been imputed into the study panel. These reference-panel-only SNPs (or type 0 SNPs in IMPUTE2) need to be checked for their reliability in the study panel. For each SNP in the imputed study panel, IMPUTE2 returns an "info metric" measuring the reliability of its imputed genotypes. This SNP-specific metric takes a value of 1 when there is no genotype uncertainty in the study sample. When the metric is less than 0.5, the sample mean of the posterior variance of imputed genotypes is at least half the variance that would be expected if alleles were sampled at random from the reference population (Howie & Marchini 2014). Using the 0.5 threshold retains 96.92 percent of the 893,641 completely missing SNPs in ADNI-1 and 96.04 percent of the 898,145 completely missing SNPs in ADNI-GO/2. For now, we omit this final quality-control step but note that future work should eliminate unreliably imputed SNPs with info metrics < 0.5. We return to this point in the Conclusions.
Chapter 4

Population Structure

In this chapter, we explore the imputed data described previously. Our goal is to investigate the population structure and of the ADNI studies. Recall that the HapMap3 reference panel has samples coming from different continents. By contrast, the participants in the ADNI studies are collected in the U.S., and the major composition of this population is immigrants or their descendants. ADNI participants self-reported on a variable called race and a variable called ethnicity. The variable race has seven categories: "White", "Black", "Asian", "Hawaiian/Other PI", "American-Indian/Alaskan", "More than one", "Unknown". The variable ethnicity has three categories: "Not Hispanic/Latino", "Hispanic/Latino", and "Unknown". We use principal-component analysis and model-based clustering to investigate the population structure of the participants in the ADNI studies, and in particular the meaning of self-reporting as ethnicity 'Hispanic' in these studies.

4.1 Principal Component Analysis

Principal component analysis (PCA) is a dimension-reduction method that projects the high-dimensional data onto a lower dimensional subspace explaining the most variance. PCA constructs uncorrelated linear combinations of the variables called principal components (PCs), which explain the most variability in the original data. The first PC has the largest variance and, normally, the first few PCs that explain the most variance (80-90%) can be used for further analysis instead of all the PCs (Johnson et al. 2002). In genetics, PCA is widely used to study the population structure of genetic data. The continuous axes of the first and second PCs best reflect the genotypic variability between sampled individuals.

HapMap3 collected over 1,301 DNA samples from a variety of global populations. We have used the diverse HapMap3 reference panel to impute the missing SNPs in the ADNI individuals. We can also use HapMap3 as an anchor to gain insight into the genetic ancestry of the ADNI subjects. Before PCA, we first need to make sure the individuals in HapMap3 are independent (i.e. not related). After using PLINK to remove known relatives in the HapMap3 reference panel, we have 988 out of 1184 individuals left to work with. Second,
we take the intersection of SNPs shared by the imputed ADNI data sets and the HapMap3 reference data set and the union of subjects across the three data sets. This leads to 1,387,166 SNPs and 2,214 subjects in total. Third, we use PLINK to infer pairs of first-degree relatives or duplicate samples on the basis of the genetic data and remove a random member in each of the 28 pairs found across the three data sets. Next, we exclude the SNPs that are correlated or in linkage disequilibrium (LD). Specifically, we use PLINK to prune SNPs in a moving windows of size 50 SNPs (Purcell et al. 2007). For each window of 50 SNPs, we regress each SNP against the remaining SNPs simultaneously, to obtain the $R^2$ value that measures the proportion of variance of that SNP explained by the other SNPs in the window. The variance inflation factor (VIF) measures how well the other SNPs in the window predict the target SNP and is given by $\frac{1}{1-R^2}$. If the VIF of a target SNP exceeds 1.5, the SNP is called a prune-out SNP, and removed. After checking each SNP in the current window, we shift our window by 50 SNPs so that the new window has no overlaps with the current window. We continue this process until all the SNPs have been checked. After removing the pruned-out SNPs, we have 2,186 individuals and 481,166 out of 1,387,166 SNPs left to be passed on to PCA.

Methodology

PLINK offers PCA based on the variance-standardized relationship matrix (Purcell et al. 2007). Suppose our genetic data consist of $N$ individuals and $S$ SNPs. For SNP $s$, we define $G_s = (G^s_1, ..., G^s_n)^T$ as an $n \times 1$ vector, where $G^s_i = 0, 0.5, 1$ is the genotype for the $i^{th}$ individual with 0, 1, or 2 copies of the reference allele at SNP $s$, respectively. The expected value of $G^s_i$ is the allele frequency estimate $\hat{p}_s$ for SNP $s$; the standard deviation of $G^s_i$ is $\sqrt{\hat{p}_s(1-\hat{p}_s)}$. We can define $Z$ to be an $N \times S$ standardized genotype matrix with entry $i, s$ that takes the form (Thornton & Wu 2015)

$$Z_{is} = \frac{G^s_i - \hat{p}_s}{\sqrt{\hat{p}_s(1-\hat{p}_s)}}$$

(4.1)

The genetic relationship matrix (GRM) $\hat{\psi} = \frac{1}{S}ZZ^T$ measures the average genetic relationships between individuals. For example, the average genetic difference between individual $i$ and $j$ takes the form (Yang et al. 2011)

$$\hat{\psi}_{ij} = \frac{1}{S} \sum_{s=1}^{S} \frac{(G^s_i - \hat{p}_s)(G^s_j - \hat{p}_s)}{\sqrt{\hat{p}_s(1-\hat{p}_s)}}$$

(4.2)

PLINK applies PCA to the GRM to calculate the eigenvectors and eigenvalues. The spectral decomposition gives us $\hat{\psi} = VDV^T$, where $V$ is an $N \times N$ matrix with $i^{th}$ orthogonal column vector being the $i^{th}$ PC and $D$ is a diagonal matrix with $(i, i)$ entry being the eigenvalue $\lambda_i \geq 0$ for $i = 1, 2, ..., N$. The $i^{th}$ largest eigenvalue represents the variance explained by the
$i^{th}$ ordered PC. PCs are also referred to as continuous axes of variation that capture the genetic variation between sampled subjects. Frequently, the PCs reflect geographic origins (Price et al. 2006).

Results

We used PLINK to calculate the first 20 PCs for the individuals in the three data sets. As PCA reflects the genetic similarities between individuals, individuals from the same population tend to form a cluster in the subspace. The first three PCs explained 80.68% of variance, as shown in the scree plot Figure 4.1. The clear elbow at the third PC suggests that it is safe to use the first three PCs to summarise the data.

![Scree plot of Eigenvalues](image)

**Figure 4.1:** Scree plot of Eigenvalues

We use plots of the first two PCs because they reveal more distinct population structure of the reference panel and ADNI studies compared to the other PCs. As mentioned above, we will be using HapMap3 as an anchor. Thus, in our first PC plot, we project the HapMap3 individuals only onto the first two PCs. Recall the population samples labels in Table 2.2. We plot the first two PCs coloured by the samples’ population. The resulting clustering patterns in Figure 4.2 are very distinct. Han Chinese in Beijing, Chinese in Denver, and Japanese in Tokyo are clustered closely together at the top-left corner, as they are clearly East Asians. Toscani individuals from Italy, and Utah residents with Northern and Western European ancestry are clustered together at the bottom-left corner, as they are Europeans. Individuals of African American ancestry in the U.S. are clustered closely with different African populations at the bottom-right of the plot. Interestingly, Gujarati Indians in
Houston are clustered with individuals of Mexican ancestry in Los Angeles, between the East Asians and Europeans on the plot, which suggests genetic similarities between Indians and Mexicans relative to the other populations considered.

![HapMap3 Population Structure](image)

**Figure 4.2:** PC plot of HapMap3 Subjects

The clustering in the PC plot of the HapMap3 participants can assist in understanding the genetic structure of the participants from the ADNI studies. Figure 4.3 shows the PC plots for the ADNI participants coloured by their self-reported race and ethnicity. Most ADNI participants who self-reported as White cluster in the bottom-left corner, suggesting they are of European ancestry. Most ADNI participants who self-reported as Asian cluster in the top-left of the figure, suggesting they are of East-Asian ancestry. Most ADNI participants who self-reported as Black tend to cluster with the African-American HapMap3 samples. However, some ADNI participants who self-reported as Asian or White cluster with the HapMap3 samples of Mexican or Indian ancestry, and one ADNI participant who self-reported as American-Indian/Alaskan clusters with the HapMap3 European samples. In the PC plots coloured by self-reported ethnicity, most ADNI participants who self-reported as Hispanic/Latino cluster with the European and Mexicans/Indian populations in HapMap3. We next look into the genetic composition of individual ADNI participants through model-based clustering (MBC).
4.2 Model-Based Clustering

As discussed above, the PC plots of ADNI participants and HapMap3 samples give an idea of the population structure in the ADNI studies. Since the U.S. is comprised of many recent immigrants and their descendants, we are also interested in understanding the genetic composition of individual participants. ADNI participants have self-reported race, but people of complex ethnic ancestry may not be aware of it or at least may not be able to adequately describe it in terms of the limited choices offered by study questionnaires. Therefore, we want to investigate how self-reported race and ethnicity correspond to the genetic data of ADNI participants. We use model-based cluster analysis with the R package mclust and the first three genetic PCs to assign the ADNI participants to the HapMap3-defined ethnic clusters. Our clustering approach is "quick-and-dirty" and meant as a first step towards data exploration.
Methodology

Model-based clustering consists of three stages. First, hierarchical clustering is used to initialize clusters. Second, the parameters of a multivariate mixture model are estimated by maximum likelihood. Third, the clustering model and the number of clusters are selected using the Bayesian information criterion (BIC), an approximation to the Bayes factor (Fraley & Raftery 2002, 2007). In model-based clustering, we assume that the population for our data consists of a mixture of subpopulations, and that each subpopulation represents a cluster (Banfield & Raftery 1993). Given an \( N \)-dimensional dataset with independently and identically distributed observations \( x = (x_1, ..., x_n) \), let \( f_k(x_i; \theta_k) \) be the probability density function for observation \( x_i \) from the \( k^{th} \) cluster, where \( \theta_k \) are the corresponding parameters (Fraley & Raftery 1998). The finite mixture model takes the form (Scrucca et al. 2016)

\[
f(x; \psi) = \sum_{k=1}^{G} \pi_k f_k(x_i; \theta_k),
\]

where \( G \) is the number of components or clusters, \( \Psi = \{\pi_1, ..., \pi_{G-1}, \theta_1, ..., \theta_G\} \), and \( \pi_k \) is the probability that an observation belongs to the \( k^{th} \) cluster (i.e., \( \pi_k > 0 \), and \( \sum_{k=1}^{G} \pi_k = 1 \)). The most popular model for \( f_k \) is the multivariate normal density, where \( \mu_k \) is the mean vector, \( \Sigma_k \) is the variance-covariance matrix, and \( \theta_k = (\mu_k, \Sigma_k) \). Based on the geometric properties of the multivariate-normal (MVN) density, the probability density of each cluster is ellipsoidal, with center at \( \mu_k \). In contrast, the cluster-specific parameter \( \Sigma_k \) determines the geometric features of orientation, volume, and shape (Scrucca et al. 2016). These geometric features are controlled by parametrizing the covariance matrix in terms of the eigenvalue decomposition

\[
\Sigma_k = \lambda_k A_k D_k D_k^T,
\]

where \( \lambda_k A_k \) is a diagonal matrix of the eigenvalues of \( \Sigma_k \), \( A_k \) is a diagonal matrix with elements that are proportional to the eigenvalues and which specifies the shape of the density contours, \( \lambda_k \) is a constant of proportionality controlling the volume of the ellipsoid, and \( D_k \) is an orthogonal matrix of eigenvectors which determines the orientation of the \( k^{th} \) cluster’s ellipsoid (Banfield & Raftery 1993, Scrucca et al. 2016). For instance, if \( \Sigma_k = \lambda I \), where \( I \) is the identity matrix, the clusters are spherical and have the same volume. If \( \Sigma_k = \lambda_k I \), the clusters are spherical but of different volumes. If \( \Sigma_k = \Sigma \), the clusters have equal variance, which means they have the same geometric features of orientation, volume, and shape. If the clusters have unconstrained variance \( \Sigma_k \), then the geometric features vary between clusters (Fraley & Raftery 2002).

The multivariate mixture distributions have multiple modes corresponding to different cluster memberships. Cluster membership is a latent variable and so we find the maximum-likelihood estimators using the Expectation-Maximization (EM) algorithm. An hierarchical agglomerative clustering approach is taken in which the two clusters that provide the
smallest decrease in classification likelihood are merged recursively, and then the classification likelihood is estimated by the EM algorithm (Scrucca et al. 2016). The EM algorithm consists of two steps: an 'E step' that creates a conditional expectation function of the log-likelihood with the current parameter estimates; then an 'M step' that estimates the optimal values for the parameters by maximizing the function in the 'E step'.

We use Bayesian model selection to choose the parameterization of the model and the number of clusters (Fraley & Raftery 2002). The Bayes factor for model comparison is the posterior odds of one model against another regardless of their priors (Kass & Raftery 1995). For example, the Bayes factor between models $M_1$ and $M_2$ is $B_{12} = \frac{p(D|M_1)}{p(D|M_2)}$, where $B_{12} > 1$ suggests substantial evidence favoring $M_1$, and $B_{12} > 100$ suggests strong evidence for $M_1$ (Kass & Raftery 1995). Suppose we have data $D$, and models $M_1,...,M_k$, with prior probabilities $p(M_k)$ and posterior probabilities $p(M_k|D)$, where $k = 1,...,G$ and $G$ is the number of clusters. Then, by Bayes’s theorem, the posterior probability of model $M_k$ is proportional to the probability density of data $D$ given model $M_k$ times the prior probability of that model, which takes the form

$$p(M_k|D) \propto p(D|M_k)p(M_k)$$

(4.5)

As the parameters of the model are unknown, the density $p(D|M_k)$ is obtained by integrating over the parameter vector $\theta_k$. The resulting integrated likelihood takes the form

$$p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)d\theta_k,$$

(4.6)

where $p(\theta_k|M_k)$ is the prior distribution of the parameter vector for model $M_k$. Evaluating the integral that defines the integrated likelihood is difficult in practice, and also the prior distribution of the parameter vector $p(\theta_k|M_k)$ might be hard to set accurately. We can, however, avoid the integration by using BIC to approximate the integrated likelihood

$$2 \log p(D|M_k) \approx 2 \log p(D|\hat{\theta}_k, M_k) - \nu_k \log(n) = BIC_k,$$

(4.7)

where $\hat{\theta}_k$ is the maximum likelihood estimate of the parameter vector, and $\nu_k$ is the number of independent parameters to be estimated in model $M_k$. The model and number of clusters are selected based on the maximum BIC value (Fraley & Raftery 2002).

Results

In the R package mclust, 14 model-combinations are considered that vary the volume, shape and orientation of the clusters in the data. The volume controls whether each cluster has the same number of observations, shape controls whether each cluster has the same variance, and orientation controls whether each cluster should be forced to be axis-aligned (Greenwell 2020). The optimal model and the number of clusters that maximize the BIC are
selected. In PCA, we have removed the first-degree relatives among the HapMap3 and ADNI individuals, so they are independent. We fit the first three PCs from HapMap3 in \texttt{Mclust()} to obtain the initial clusters. The model that maximizes the BIC has 9 clusters, and is ellipsoidal with unequal variance and differing numbers of individuals across the clusters. In terms of the HapMap3 samples, Cluster 1 corresponds to ASW, cluster 2 to YRI, cluster 3 to European samples (CEU and TSI), cluster 4 to East-Asian samples (CHB, CHD, and JPT), cluster 5 to GIH, cluster 6 to LWK, cluster 7 to MEX, cluster 8 and cluster 9 to MKK. Table 4.1 summarizes the cluster memberships for the HapMap3 samples after combining the two MKK clusters and the GIH and MEX clusters. Figure A.1 illustrates their posterior probabilities of cluster membership. One ASW sample clusters with both the ASW and YRI samples. Similarly, one YRI sample clusters with both ASW and YRI. These two samples appear to be mixtures between ASW and YRI.

Table 4.1: HapMap3 clustering by population

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASW</td>
<td>47.586</td>
<td>0.414</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CEU</td>
<td>0</td>
<td>0</td>
<td>111</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CHB</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>84</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CHD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>84</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GIH</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>84</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>JPT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>86</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LWK</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>84</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MEX</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>47</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MKK</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>135</td>
<td>0</td>
</tr>
<tr>
<td>TSI</td>
<td>0</td>
<td>0</td>
<td>88</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YRI</td>
<td>0.121</td>
<td>111.879</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In further analyses, we combine the two MKK clusters and also the GIH and MEX clusters to get the seven clusters shown in the columns of Table 4.1: ASW, YRI, CEU/TSI (European), CHB/CHD/JPT (East Asian), GIH/MEX, LWK and MKK. These seven simplified clusters are used to estimate the posterior probabilities of cluster membership. These posterior probabilities sum to 1 for each individual. Instead of assigning each individual to the cluster that has the highest membership probability, we allow admixture, so that each individual can have more than one cluster membership (Lee et al. 2009). Table 4.2 shows the joint posterior (frequency) distribution of the self-reported race of ADNI participants and their genetic clustering, allowing for individuals to cluster with more than one genetically-defined group.
Table 4.2: ADNI clustering by self-reported race

<table>
<thead>
<tr>
<th>Race</th>
<th>ASW</th>
<th>EURO</th>
<th>EA</th>
<th>GIH/MEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Am Indian/Alaskan</td>
<td>0</td>
<td>3.000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Asian</td>
<td>0</td>
<td>1.000</td>
<td>12.000</td>
<td>3.000</td>
</tr>
<tr>
<td>Hawaiian/Other PI</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.000</td>
</tr>
<tr>
<td>Black</td>
<td>48.000</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>White</td>
<td>3.000</td>
<td>1100.770</td>
<td>0</td>
<td>40.230</td>
</tr>
<tr>
<td>More than one</td>
<td>1.997</td>
<td>6.000</td>
<td>0</td>
<td>0.003</td>
</tr>
<tr>
<td>Unknown</td>
<td>2.074</td>
<td>0</td>
<td>0</td>
<td>0.926</td>
</tr>
</tbody>
</table>

Figure 4.4: Genetic composition of ADNI participants in selected races

Table 4.2 suggests that the three ADNI participants who self-reported as 'American-Indian/Alaskan' cluster genetically with the European HapMap3 samples. Among ADNI participants self-reporting as 'Asian', one clusters genetically with the European samples, and three with the GIH/MEX samples (may be South Asian); the others cluster with the East Asian samples. The ADNI participant who self-reported as 'Hawaiian/Other PI' clusters with the GIH/MEX samples. The 48 ADNI participants who self-reported as Black cluster genetically with the ASW samples. Of the 1144 ADNI participants who self-reported as White, three cluster with the ASW samples and 24 with the GIH/MEX samples.
samples; some also cluster with both the European samples and the GIH/MEX samples, suggesting mixed ancestry. Of the 8 ADNI participants who self-reported as having "More than one" origin, 6 cluster entirely with the European samples, one with the ASW samples, and one with the ASW and GIH/MEX samples. Of the 3 ADNI participants who self-reported unknown origin, two cluster with the ASW samples and one clusters with both the ASW and GIH/MEX samples. Figure 4.4 shows the individual genetic compositions for selected ADNI participants based on their posterior probabilities of cluster membership.

Table 4.3: ADNI clustering by self-reported ethnicity

<table>
<thead>
<tr>
<th></th>
<th>ASW</th>
<th>EURO</th>
<th>GIH/MEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>White and Not Hisp/Latino</td>
<td>0</td>
<td>1089.937</td>
<td>20.063</td>
</tr>
<tr>
<td>White and Hisp/Latino</td>
<td>3.000</td>
<td>5.867</td>
<td>20.133</td>
</tr>
</tbody>
</table>

Figure 4.5: Genetic composition of ADNI participants self-reporting as white, by their self-reported ethnicity

In the ADNI studies, participants also reported their ethnicity as either "Hispanic/Latino" or "Not Hispanic/Latino". We want to explore the participants who self-reported as "White" and see if the genetic data has anything to say about what makes them Hispanic/Latino or not. Table 4.3 and Figure 4.5 suggests that the ADNI participants who self-reported as White and Non-Hispanic are mainly of European descent. Only a small portion of these participants cluster with the GIH/MEX HapMap3 samples or with a combination of the GIH/MEX and European HapMap3 samples. Of the ADNI participants who self-reported as White and Hispanic, three cluster with the ASW samples, some cluster with the GIH/MEX samples, some cluster with the European samples and some with various combinations of these. Past genetic studies have considered people who self-identify as White and Non-
Hispanic to be a homogeneous group of so-called white Caucasians. However, the genetic compositions of this self-reported group in the ADNI studies suggest that people may not fully understand their genetic ancestry. This limitation of self-reported race and ethnicity information is well known and shows why these data explorations are useful. Statisticians who prepare the data should investigate population structure and notify potential analysts and users of the data about any discrepancies between genetic ancestry and self-reported race and/or ethnicity.
Chapter 5

Conclusion

This project describes a protocol for the curation of large-scale genetic data. The protocol is illustrated in Figure 3.1, which outlines steps to prepare and combine genetic data for further analysis. To improve the quality and quantity of genetic data, the protocol involves applying inclusion criteria and genomic quality control to filter out SNPs and samples of poor quality, pre-phasing of genotype data in the study for computational efficiency and accuracy of haplotype estimation, estimation of haplotypes for imputation, missing-value imputation to increase the size of data, and finally data merging and formatting to prepare data for further analysis. The curation process is nontrivial and minor mistakes can corrupt the data and lead to unexpected results in further analysis. The proposed protocol successfully increased the size of the data and therefore the power of further analysis. At the beginning, we had 620,901 SNPs from ADNI-1 and 730,525 SNPs from ADNI-GO/2. After the curation process, we have 1,436,157 SNPs from ANDI-1 and 1,530,022 SNPs from ADNI-GO/2. Thus, we have almost doubled the number of SNPs in both the studies. The new SNPs that have been imputed into the study data allow us to combine the study subjects over a common and larger set of SNPs than previously defined for either of the studies. This affords more flexibility for testing and training different statistical models with these datasets.

After processing the genetic data, we explored it to investigate the population structure within the ADNI studies. PCA can uncover the population structure, and MBC reveals the genetic composition of each individual. Our exploratory data analysis lead to insights into questions of self-reported race and ethnicity of the ADNI participants. Some participants failed to identify themselves properly due to the limited choices on the study questionnaire or incomplete knowledge about their origins. We found that our initial understanding that participants self-reporting as white and non-Hispanic are of European ancestry (from, e.g., Spain) is oversimplified. Our results suggest that some of these individuals cluster genetically with European as well as Mexican samples in the HapMap3 reference populations.

The work of data processing should be appreciated by any user of the curated data. The study design and the data collection process need to be taken into consideration before

31
choosing an analysis question and the appropriate statistical method. Exploratory data analysis can also give insights into the data and its limitations before choosing a statistical analysis. This project gives an introduction to the study design, informative descriptions of the genetic data and the software involved in managing the data, and a detailed description of the steps in the curation protocol as well as the methods behind each of these steps. We hope that users of the processed data will appreciate the work involved in data processing. This project can be considered a manual to inform users about how the genetic data are collected and formatted, and how the data-curation process works with the essential software.

The data-curation protocol described in this project only applies to the specific data format for ADNI studies and the HapMap3 reference panel used to impute the missing genotypes. One direction for future research is to redo the exploratory data analysis after removing unreliable SNPs that were imputed into the ADNI studies. In the future, we would like to extend this data-curation protocol to other genetic data sets and updated reference panels. We would also like to use the processed data for further analysis with clinical and imaging data, to understand genetic associations with disease.
Bibliography

URL: http://adni.loni.usc.edu/


URL: https://bradleyboehmke.github.io/HOML/model-clustering.html

HapMap 3 (n.d.).
URL: https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html

URL: https://mathgen.stats.ox.ac.uk/impute/impute_v2.html


Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M. & Sham, P. (2007), ‘PLINK: a toolset for whole-genome association and population-based linkage analysis.’. URL: http://zzz.bwh.harvard.edu/plink/


34
Appendix A

Figures

Figure A.1: Genetic compositions of HapMap3 samples in selected populations: red ASW, gold YRI, green CEU/TSI, blue-green CHB/CHD/JPT, blue GIH/MEX, purple LWK, pink MKK