2019

Discovery and characterization of genetic variants associated with extreme longevity

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES AND COLLEGE

OF ENGINEERING

Dissertation

DISCOVERY AND CHARACTERIZATION OF GENETIC VARIANTS ASSOCIATED WITH EXTREME LONGEVITY

by

ANASTASIA GURINOVICH

B.S., Dubna International University, 2008
M.S., Dubna International University, 2010

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2019

Approved by

First Reader

Paola Sebastiani, PhD
Professor, Biostatistics

Second Reader

Thomas T. Perls, MD MPH
Professor, Medicine

Third Reader

Stefano Monti, PhD
Associate Professor, Medicine

DEDICATION

I would like to dedicate this dissertation to my parents, Alexander and Irina Gurinovich.

ACKNOWLEDGMENTS

I would like to thank my advisor, Paola Sebastiani. This dissertation would not have been possible without her. She is the best mentor I have ever had. I have learned so much from her, not just how to be a better scientist, but also how to be a better leader and a person. I would also like to thank my co-advisor, Thomas Perls for his constant support and kindness, and the opportunity to work in the centenarian study group. I would like to thank the rest of my committee members, John Farrell, Stefano Monti and Gary Benson, for their support, encouragement and useful feedback throughout my dissertation work.

I would also like to thank the Bioinformatics Program, administrative staff including David, Caroline, Johanna, Mary Ellen, and all the students and professors I have interacted with throughout my PhD.

I would like to thank my friends, who are close and afar. Knowing I have them and that I can rely on them anytime have given me the strength to finish my PhD and to keep going during the most difficult and stressful times. Specifically, I would like to thank Polina Molotkova, Georgia Skardasi, Bahar Pourazar, Sevtap Duman, Tanya Kariagannis, Michelle Dow, Marzie Rasekh, Leila Prijahi, Demarcus Briers, David Liu, Oleksander Semenov and many many others.

I would also like to thank my newly acquired family, my in-laws, Vasilika, Nike, Eleni, Ngjello, and Pandeli. You made my last few years in Boston indescribably better, and have truly become part of my family.

I would like to thank my brother-in law Matt Kuntz, who in the last seven years since I've known him has truly become like a big brother to me. I know he always has my back, and I feel comfortable to contact him for anything at any time. Also thank you to Matt for creating and raising the best nieces and nephew I could ever

ask for, Fiona, Rowan and Bodie, I love you guys so much, and I wish I could see you more often.

I would like to thank my husband Alban Kristo Cobi who I met during my second year of the PhD; but it feels that I have known him forever. He made my life complete. His constant support, care, encouragement and love give me strength every day.

Last, but not least, I would like to thank the people who have known me since day one, and who have been with me through bad and good. Because of their unconditional love, sacrifices and encouragement, I am the person who I am today. Thank you to my dad Alexander Gurinovich, a person with the biggest heart and the biggest curiosity for everything, my mom Irina Gurinovich, for her patience and love and support (even when she disagrees with my choices), my sister Marina Kuntz, who no matter the distance and differences (or similarities) in personalities, I love and miss every day; my babushka Nina Pernovskaya, who is the most hardworking person I have ever known with the biggest selfless heart who sacrificed her life for others; my dedushka Evgeniy Pernovskiy, who I miss very much and I hope he is at peace wherever his soul is now.

DISCOVERY AND CHARACTERIZATION OF GENETIC VARIANTS ASSOCIATED WITH EXTREME LONGEVITY

ANASTASIA GURINOVICH

Boston University, Graduate School of Arts and Sciences and College of Engineering, 2019

Major Professor: Paola Sebastiani, Professor of Biostatistics

ABSTRACT

Over the last decade, there have been multiple genome-wide association studies (GWASs) of human extreme longevity (EL). However, only a limited number of genetic variants have been identified as significant, and only few of these variants have been replicated in independent studies. There are two possible reasons for this limitation. First, genetic variants might have a varying effect on EL in different populations, and GWAS applied to a dataset as a whole may not pinpoint such differences. Second, EL is a very rare trait in a population, and rare and uncommon variants might be important factors in explaining its heritability but GWASs have focused on the analyses of variants that are relatively common in the population. In this dissertation, I present three projects that address these issues. First, I propose PopCluster: an algorithm that automatically discovers subsets of individuals in which the genetic effects of a variant are statistically different. PopCluster provides a simple framework to directly analyze genotype data without prior knowledge of subjects ethnicities. Second, I investigate ethnic-specific effects of *APOE* alleles on EL in Europeans. *APOE* is a well-studied gene with multiple effects on aging and longevity. The gene has 3 alleles: e2, e3 and e4, whose fre-

quencies vary by ethnicity. I identify several ethnically different clusters in which the effect of the e2 and e4 alleles on EL changes substantially. Furthermore, I investigate the interaction of *APOE* alleles with the country of residence. Results of this analysis suggest possible interaction of this gene with dietary habits or other environmental factors. For the third project, I perform a GWAS of rare variants and EL in a case-control dataset with median age of cases 104 years old. I analyze 4.5 million high-imputation quality rare SNPs imputed with HRC panel with minor allele frequency < 0.05. The analysis replicates all previous genome-wide level significant SNPs and identifies a few more potential targets. Additionally, I use serum protein data available for a subset of subjects and find significant pQTLs which have potential functional role. Based on these analyses, both genetic and environmental factors appear to be important factors for EL.

CONTENTS

Dedication iv

Acknowledgements v

Abstract vii

List of Tables xii

List of Figures xv

List of Symbols and Abbreviations xviii

1 Introduction 1

1.1 Human extreme longevity 2

1.2 Summary 4

2 PopCluster: an algorithm to identify genetic variants with ethnicity-dependent effects 7

2.1 Introduction 7

2.2 Materials and methods 9

2.2.1 Methodology 9

2.2.2 Genotype and phenotype data 15

2.2.3 Evaluation 18

2.3 Results 23

2.3.1 Evaluation results 23

2.3.2 Application to real data 29

2.4 Discussion 47

3 Varying effect of *APOE* alleles on extreme longevity in European ethnicities 51

3.1 Introduction 51

3.2 Materials and methods 52

3.2.1 Study Populations 52

3.2.2 Genotype Data 53

3.2.3 Statistical Analysis 53

3.3 Results 54

3.3.1 Effect of *APOE* in Italians 61

3.3.2 Effect of *APOE* in Danish 63

3.4 Discussion 64

4 GWAS of rare variants and extreme longevity 67

4.1 Introduction 67

4.2 Materials and methods 68

4.2.1 Study populations 68

4.2.2 GWAS dataset 69

4.2.3 Protein data 71

4.2.4 Statistical analysis 72

4.3 Results 75

4.3.1 GENESIS 75

4.3.2 Bayesian logistic regression 75

4.3.3 Replication 79

4.3.4 pQTL analysis 80

4.4 Discussion 92

5 Conclusions **96**

Bibliography **99**

Curriculum Vitae **110**

LIST OF TABLES

2.1 Summary of studies of extreme longevity included in the analysis. . 17

2.2 Subset of SNPs associated with EL. 17

2.3 Percentage of simulation runs (FPR) that returned n number of significant associations (Clusters) ($n = \{0, 1, 2, 3\}$). 25

2.4 Percentage of simulation runs (TPR) that returned at least one cluster with more than 80% subjects in the region simulated to have an association between allele A and the phenotype (scenario 1). 27

2.5 Average number of returned clusters with more than 80% subjects in the region simulated to have an association between allele A and the phenotype (scenario 1). 27

2.6 Percentage of simulation runs (TPR) of PopCluster that returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset (scenario 2). 29

2.7 Percentage of simulation runs (TPR) of PopCluster that returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset with re-shuffled case/control labels (scenario 2). 29

2.8 Percentage of simulation runs (TPR) of PopCluster that returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset with balanced number of re-shuffled case/control labels (scenario 2). 30

2.9 Percentage of simulation runs (TPR) of PopCluster that returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset with balanced number of re-shuffled case/control labels in each cluster (scenario 2). 30
2.10 Complete list of clusters detected by PopCluster for 371 SNPs and EL. 31
2.11 Complete list of clusters for rs3764814 and EL. 42
2.12 Complete list of clusters for rs72834698 returned as an output from PopCluster run on HRS dataset with phenotype of surviving past age 90. 45
2.13 Complete list of clusters detected by PopCluster for 11 SNPs and HRS. 45

3.1 *APOE* genotype distribution in the studies of extreme longevity. . . . 54
3.2 Associations between *APOE* e2 and EL in ethnic-specific clusters. . . 56
3.3 Associations between *APOE* e4 and EL in ethnic-specific clusters. . . 57
3.4 Gene-environment model parameters (logistic regression) testing association between *APOE* e4 and EL in cluster 1309 enriched of subjects of South Italian descent. 62
3.5 Distribution of countries where subjects live for clusters enriched of Danish ethnicity. 64

4.1 Subjects with SOMAscan protein data. 72
4.2 Genome-wide significant loci as returned by GENESIS. 75
4.3 SNPs for which GWAS results generated by GENESIS are not reliable. 78
4.4 Genome-wide significant loci as identified by Bayesian analysis. . . . 78
4.5 Signature of 16 biomarkers associated with *APOE* genotypes. 86

4.6 Description of SNPs not on chromosome 19 which were identified as significant pQTLs. 93
4.7 Protein-SNP pQTLs for SNPs not on chromosome 19. 94

LIST OF FIGURES

2.1 Generation of clusters using genome-wide principal components . . 11

2.2 Test of the associations between the phenotype and SNPs in each cluster 12

2.3 A schematic of the recursive pruning of redundant clusters 14

2.4 Scatter plot of the first two principal components with the selected region for the TPR evaluation 21

2.5 Boxplots of the FPR in six different simulations 24

2.6 Boxplots of the TPR for various combinations of probabilities of allele A in cases and controls 26

2.7 Boxplots of the differences between true effect β and estimated effect $\hat{\beta}$ in the clusters where allele A was simulated to be associated with EL (scenario 1) 28

2.8 Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 1 31

2.9 Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 2 32

2.10 Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 3 33

2.11 Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 4 34

2.12 Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 5 35

2.13 Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 6 36
2.14 Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 7 37
2.15 Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 8 38
2.16 Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 9 39
2.17 Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 10 40
2.18 Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of subjects in the NECS 41
2.19 Full hierarchical tree structure returned for the analysis of EL dataset before pruning of redundant clusters step 43
2.20 Ethnic groups in which the effect of SNP rs3764814 on EL did not reach statistical significance 44
2.21 Full hierarchical tree structure returned for the analysis of survival past age 90 HRS dataset before pruning of redundant clusters step . 46
2.22 Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the HRS 47

3.1 Scatter plots of principal components PC1-PC2 and PC3-PC4 from genome-wide genotype data of all subjects in the study of EL with labels - set 1 58

3.2 Scatter plots of principal components PC1-PC2 and PC3-PC4 from genome-wide genotype data of all subjects in the study of EL with labels - set 2 59
3.3 Scatter plots of principal components PC1-PC2 and PC3-PC4 from genome-wide genotype data of all subjects in the study of EL with labels - set 3 60
3.4 Scatter plots of principal components PC1-PC2 and PC3-PC4 calculated from genome-wide genotype data of all subjects in the EL study for the cluster with 1309 subjects 63

4.1 Scatter plots of principal components PC1-PC2 and PC3-PC4 from genome-wide genotype data of 6,088 genotyped and imputed subjects from the NECS and Illumina control repository 70
4.2 Scatter plots of re-calculated principal components PC1-PC2 and PC3-PC4 from genome-wide genotype data of European subjects in NECS and Illumina control repository 71
4.3 GENESIS GWAS flow chart 73
4.4 QQ plot of the GWAS of EL as implemented by GENESIS 76
4.5 Manhattan plot of the GWAS of EL as implemented by GENESIS . . 77
4.6 Distribution of protein intensity in log scale by *APOE* genotypes . . . 87
4.7 Boxplots of distribution of protein intensity in log scale by significant pQTLs that are not on chromosome 19 93

LIST OF SYMBOLS AND ABBREVIATIONS

APOE .	Apolipoprotein E
bps ..	Base pairs
CA ..	Coded Allele
CAF ..	Coded Allele Frequency
CI ...	Confidence Interval
dbGaP .	Database of Genotypes and Phenotypes
EL ...	Extreme Longevity
FDR ..	False Discovery Rate
FN ..	False Negative
FP ...	False Positive
FPR ..	False Positive Rate
GEE ..	Generalized Estimating Equation
GMMAT	Generalized Linear Mixed Model Association Test
GWAS .	Genome-wide Association Study
HRC ..	Haplotype Reference Consortium
HRS ..	Health and Retirement Study
IBD ..	Identity-by-descent
IRB ..	Institutional Review Board
LD ..	Linkage Disequilibrium
LGP ..	Longevity Gene Project
LLFS .	Long Life Family Study
MAF .	Minor Allele Frequency

NECS .	New England Centenarian Study
OR ..	Odds Ratio
p ...	P-value
PC ...	Principal Component
PCA ..	Principal Component Analysis
pQTL .	Protein Quantitative Trait Loci
QC ..	Quality Control
SE ...	Standard Error
SICS ..	Southern Italian Centenarian Study
SNP ..	Single Nucleotide Polymorphism
TPR ..	True Positive Rate

CHAPTER 1

Introduction

In a genome-wide association study (GWAS) of human extreme longevity (EL) with a case-control study design, cases are defined as subjects who have attained a pre-defined age cutoff and controls are often general population controls. Using this study design, the genetic effect is the odds ratio (OR) for extreme longevity estimated typically via a logistic regression model, which compares the carriers and non-carriers of a specific allele of a SNP. One of the main challenges in the genetic studies of EL is lack of replication of many of the findings. There are two possible explanations for this issue. First, genetic variants associated with EL might be population-specific. Spurious association between phenotype and a SNP could stem from ethnic-specific differences in allele frequency (Solovieff et al., 2010). There are several strategies to control for population structure (Price et al., 2006; Epstein et al., 2007; Kimmel et al., 2007; Wang, 2009). The most commonly used approach is to adjust regression model by principal components (PCs) calculated from the genome-wide genotype data (Price et al., 2006). However, simply adjusting by population structure does not identify important variants specific to each group defined by different genetic compositions, which are shaped through complex environmental influences and population dynamics (Giuliani et al., 2018). A second reason for the lack of replication in genetic studies of EL could be that most of the focus in the field has been on identifying common variants. EL is a rare trait in the population; thus, less common variants might be important targets that would explain the heritability and replicate better.

In this chapter I will first review what extreme longevity is and why studying it is important. Next, I will outline the goal and projects of this dissertation.

1.1 HUMAN EXTREME LONGEVITY

Aging and longevity have been popular topics of study because almost all of us want to know how to live longer, and most importantly - healthier. The research focus of our group is extreme longevity in humans. In this subsection I will review what extreme longevity is and why it is important to study it.

Centenarians are known to delay or sometimes completely avoid some age-related diseases [1] (Andersen et al., 2012). This phenomenon is known as compression of morbidity and was first proposed by James Fries in 1980 (Fries, 1980). In short, the theory states that as the limit of life span is approached the proportion of time we experience morbidity and disability gets smaller.

This hypothesis was subsequently investigated with from centenarians and supercentenarians in the New England Centenarian Study (NECS) (Evert et al., 2003). An initial analysis showed that there are three profiles that centenarians fall into with respect to the compression of morbidity hypothesis, specifically: survivors, delayers, and escapers. In addition, the investigators demonstrated that survival and delay of or escape from morbidity can be gender, environment and health-choices dependent. *Survivors* were defined as individuals who were diagnosed with at least one age-related disease before the age of 80. *Delayers* were defined as individuals who were diagnosed with at least one age-related disease between ages of 80 and 90. *Escapers* were defined as the "lucky ones" who escaped (were not diagnosed with) any age-related diseases before 100 years old.

After the NECS collected the largest sample to-date of semi-supercentenarians (age at death between 105 and 109 years) and supercentenarians (age at death 110

[1] Some of the age-related diseases: heart disease, diabetes, cancer, skin cancer, osteoporosis, thyroid condition, hypertension, stroke, dementia, chronic obstructive pulmonary disease, Alzheimer's, Parkinson's, etc.

and older) in the world, Dr. Perls and co-authors observed compression of morbidity among semi-supercentenarians and especially supercentenarians (Andersen et al., 2012). This result subsequently has been replicated in the Long Life Family Study (LLFS) (Sebastiani et al., 2013) and the Longevity Genes Project (LGP) (Ismail et al., 2016).

Another important aspect of extreme longevity is whether it is a hereditary or an environmental trait. Investigators from the NECS have put together a large collection of multigeneration pedigrees of centenarians (Sebastiani et al., 2016a) and were able to demonstrate that exceptional longevity strongly clusters in families. In addition, centenarians are more genetically homogenous compared to random population controls (Sebastiani et al., 2016b).

Even though in the last few decades the proportion of centenarians has been growing, the frequency of supercentenarians has remained the same. This may be an indication that there are some rare genetic variations that contribute to extreme longevity. GWASs of EL have shown that individual single nucleotide polymorphisms (SNPs) have relatively small effects on exceptional longevity (Sebastiani et al., 2012). However, when several SNPs are combined into a genetic signature, a stronger effect is achieved.

Centenarians and supercentenarians are a great, but complex, model for healthy aging. There is still a lot to learn, and one must take genetics, environment and behavioral choices into account. Some interesting observations by Dr. Thomas Perls in his book "Living to 100: Lessons in Living to your Maximum Potential at any Age" is that healthy centenarians and supercentenarians manage stress very well and are usually non-smokers (Thomas T. Perls, 1999).

1.2 SUMMARY

The overall goal of my dissertation is to discover and analyze ethnic-specific and rare genetic variants associated with EL. To achieve this goal, I first develop the PopCluster algorithm that facilitates the discovery of genetic variants with ethnicity-dependent effects on a phenotype. Second, I investigate varying effects of *APOE* alleles on extreme longevity in various ethnicities. Third, I perform a GWAS of rare variants of extreme longevity.

In the first part of my dissertation, I propose PopCluster: an algorithm to automatically discover subsets of individuals in which the genetic effects of a variant are statistically different (Gurinovich et al., 2019). Over the last decade, more diverse populations have been included in GWAS. If a genetic variant has a varying effect on a phenotype in different populations, GWAS applied to a dataset as a whole may not pinpoint such differences. It is especially important to be able to identify population-specific effects of genetic variants in studies that would eventually lead to development of diagnostic tests or drug discovery. PopCluster provides a simple framework to directly analyze genotype data without prior knowledge of subjects' ethnicities. PopCluster combines logistic regression modeling, principal component analysis, hierarchical clustering, and a recursive bottom-up tree parsing procedure. The evaluation of PopCluster suggests that the algorithm has a stable low false positive rate ($\sim$4%) and high true positive rate ($>$80%) in simulations with large differences in allele frequencies between cases and controls. Application of PopCluster to data from genetic studies of longevity discovers ethnicity-dependent heterogeneity in the association of rs3764814 (*USP42*) with the phenotype. PopCluster was implemented using the R programming language (R Core Team, 2018), PLINK (Chang et al., 2015; Purcell et al., 2007) and EIGENSOFT

software (Price et al., 2006), and can be found at the following GitHub repository: https://github.com/gurinovich/PopCluster with instructions on its installation and usage.

In the second part of my dissertation, I investigate ethnic-specific effects of *APOE* alleles on extreme longevity in Europeans. *APOE* is a well-studied gene with multiple effects on aging and longevity. The gene has 3 alleles: e2, e3 and e4, whose frequencies vary by ethnicity. While the e2 is associated with healthy aging, the e4 allele has a deleterious effect and its prevalence among people with EL is low. Using the PopCluster algorithm, I identified several ethnically different clusters in which the effect of the e2 and e4 alleles on EL changed substantially. For example, PopCluster discovered a large group of 1309 subjects enriched of Southern Italian genetic ancestry with weaker protective effect of e2 and weaker damaging effect of e4 on EL compared to other European ethnicities. Further analysis of this cluster suggests that the odds for EL in carriers of the e4 allele with Southern Italian genetic ancestry differ depending on whether they live in the U.S.A. or Italy. PopCluster also found clusters enriched of subjects with Danish ancestry with varying effect of e2 on EL. The country of residence (Denmark or U.S.A.) appears to change the odds for EL in the e2 carriers. These results suggest possible interaction of this gene with dietary habits or other environmental factors.

In the third part of my dissertation, I conduct a genome-wide association study of 4216 individuals including 1317 centenarians from the NECS (median age = 104 years) using >9M genetic variants imputed to the HRC panel of 65,000 haplotypes. The strong heritability of extreme human longevity supports the hypothesis that this is a genetically-regulated trait. However, association studies focused on common genetic variants have discovered a limited number of longevity-associated

genes. The set for the analysis includes approximately 5M uncommon variants. The associations are tested using a mixed effect logistic regression model with genotype-based kinship covariance of the random effects to adjust for cryptic relations using the package GENESIS. The analysis discovers 61 genome-wide significant SNPs ($p < 5E-08$) including fifteen new loci in chromosomes 4, 6, 7, 8, 9, 10, 14 and 15 in addition to the APOE locus. The list includes new protein quantitative trait loci (pQTLs) in serum that suggest new biological mechanism involved in extreme human longevity.

CHAPTER 2

PopCluster: an algorithm to identify genetic variants with ethnicity-dependent effects

2.1 INTRODUCTION

In many genetic association studies, the phenotype is a binary variable indicating the presence or absence of a trait, and logistic regression is a popular model used to test the associations between SNPs and the phenotype. The model can be used to adjust the association between each SNP and the phenotype by various covariates, including genome-wide principal components that describe the genetic architecture of different ethnic groups (Solovieff et al., 2010).

While non-European ethnicities have been under-represented in GWASes, the number of diverse ethnicities is increasing (Popejoy & Fullerton, 2016; Petrovski & Goldstein, 2016). Comparison of the ancestry distribution of the GWAS catalog from 2009 to 2016 shows, for example, that the percentage of subjects of European and Jewish ancestry has decreased from 96% to 81%, and the number of subjects of Asian descent has increased from 3% to 14% (Need & Goldstein, 2009; Popejoy & Fullerton, 2016). Although some other ethnic groups are still highly underrepresented, their inclusion continues to increase (Mathew et al., 2017).

Population stratification can challenge genetic association studies when the magnitude and/or direction of the effects of the allele as well as the allele frequency vary according to ethnicity (Popejoy & Fullerton, 2016; The PLOS Medicine Editors et al., 2016; Torkamani et al., 2012). For example, the *APOE* e4 allele, which is a known risk factor of Alzheimer's disease, has different allele frequencies and effects in Europeans, Africans and Hispanics (Corbo & Scacchi, 1999; Liu et al.,

2013; Hendrie et al., 2014; Campos et al., 2013). Similarly, it has been shown that for 25% of the SNPs associated with body mass index, type 2 diabetes and lipid levels in Europeans, the strength of association varies substantially in at least one non-European population (Carlson et al., 2013). Even though a large number of these SNPs may be in linkage disequilibrium (LD) with causal SNPs, it is important to investigate whether any of the associations are due to true population differences rather than differences in LD between populations.

If the association between a SNP and a trait is tested in a group of subjects in which the genetic effect of the SNP varies with ethnicity, ignoring the interaction between the genetic effect and the ethnicity may produce either a false positive (FP) or a false negative (FN) result. For example, if the effects of the SNP are in opposite directions in some ethnic groups, ignoring these antagonistic effects may result in a FN result. An alternative and common situation is when the genetic effect is significant only in a particular genetic background that is over-represented in the analysis. Ignoring the ethnicity effect may produce a FP association in ethnicities in which there is no association between the SNP and phenotype.

In this chapter, I introduce PopCluster - an algorithm that finds subpopulations of study subjects in which the genetic effects of a SNP are different. I thoroughly evaluate the false and true positive rates of PopCluster using real and simulated genetic data. I also apply the algorithm to real data from four studies of extreme longevity and the Health and Retirement Study (HRS) (Sonnega et al., 2014). I conclude by reviewing usefulness and limitations of PopCluster, and suggest potential applications.

2.2 MATERIALS AND METHODS

2.2.1 Methodology

The algorithm takes the following variables as its input: genome-wide genotype data for each subject, a list of SNPs of interest to test, phenotype information for each subject, and a list of covariates to be included in the model, for example sex and age. PopCluster takes this information to discover ethnic specific effects of the list of SNPs of interest by performing the following analyses, which are described in detail in the next sub-sections. First, PopCluster performs PCA of the genome-wide genotype data and hierarchical clustering of the most informative principal components to discover a set of nested clusters of genetic ethnicity. Next, genome-wide principal components are recalculated in each cluster of subjects, followed by test of the associations between the phenotype and SNPs in each cluster. The final step of PopCluster is pruning of redundant clusters to generate the final list of SNPs and clusters with varying genetic effects on the phenotype.

2.2.1.1 Cluster generation

The cluster generation step is depicted in Figure 2.1. First, PopCluster computes genome-wide principal components using the EIGENSOFT package on the genome-wide genotype data (Price et al., 2006). Next, hierarchical clustering is performed on subjects using the most informative number of principal components. Scree plot is a good way to decide on how many principal components to use (Solovieff et al., 2010). Typically, 6 principal components are sufficient to characterize the major European ethnic groups, while up to 20 principal components may be needed to characterize more heterogeneous ethnic groups. Since the dendrogram

associated with hierarchical clustering is a binary tree, each node (cluster) has at most two children nodes, one parent node, and one sibling node, while the ancestors of a node are the parent node and the recursive set of parent nodes. Therefore, a set of nested clusters is generated by sequentially cutting all edges of the dendrogram that describe the agglomerative clustering procedure. Only the clusters with more than 100 subjects, and with a sibling node cluster with more than 100 subjects are included in the subsequent analyses. Figure 2.1 contains an example of a dendrogram showing hierarchical clusters of 500 subjects. Each node in the dendrogram represents a cluster and the number at each node is the size of the cluster. Clusters 110, 240, 350, 150, 500 above the red, dashed line have over 100 subjects and have a sibling node with over 100 subjects and are used in the next step of the algorithm. We chose 100 as the default minimum number of subjects in a cluster to be taken in the next step of the analysis in order to have an average of 25 observations in a 2×2 table for allelic association. This threshold can be easily set to a different value if needed in the input argument list to PopCluster.

In each selected cluster, PopCluster recalculates new principal components using the EIGENSOFT package (Price et al., 2006) in order to more specifically describe the genetic structure of the individuals in every new sub-cluster. In our example in Figure 2.1, PopCluster recalculates the new principal components for clusters: 500, 150, 350, 110, and 240.

2.2.1.2 Test of the associations between the phenotype and SNPs

Next, PopCluster fits logistic regression models to test the associations between the phenotype and each SNP in every cluster:

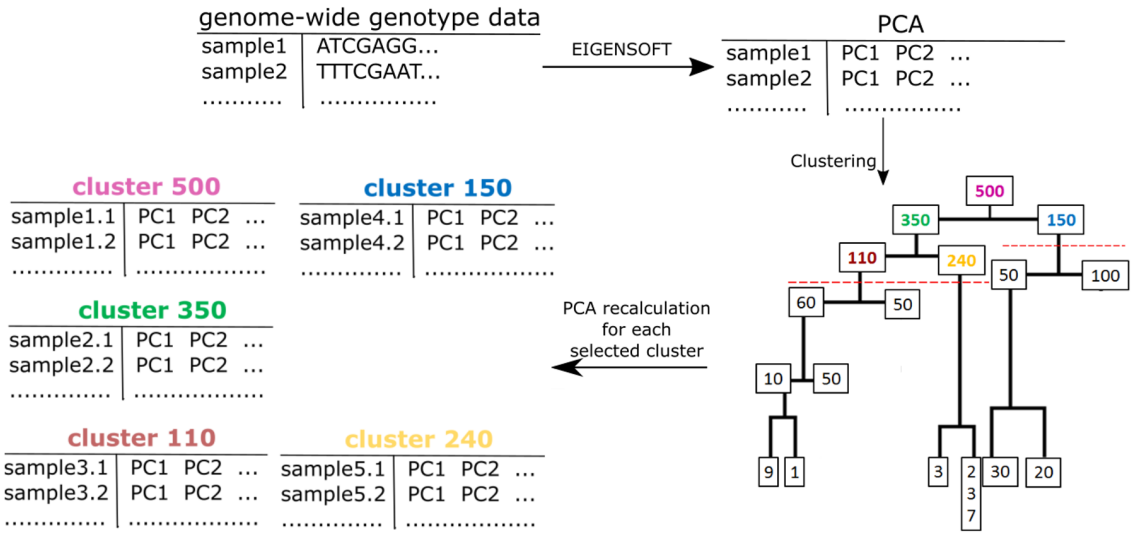

Figure 2.1: Generation of clusters using genome-wide principal components. (top left-to-right arrow): PopCluster calculates principal components from the genome-wide genotype data using the EIGENSOFT software. (middle top-to-down arrow): Subjects are clustered based on a set of principal components using the hierarchical clustering. (bottom left-to-right arrow): PopCluster recalculates principal components for each selected cluster.

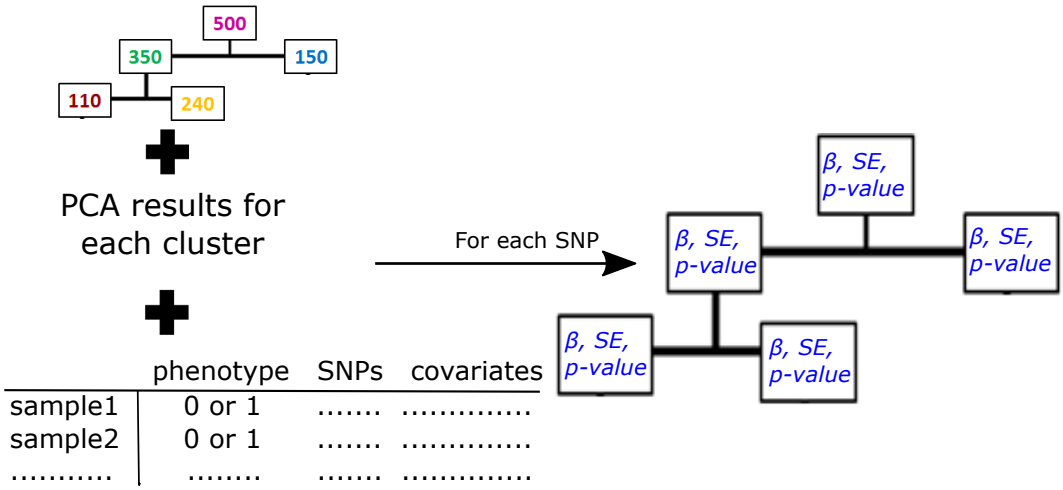

Figure 2.2: Test of the associations between the phenotype and SNPs in each cluster. Logistic regression models are fit for each SNP-cluster combinations, and the respective statistics from the models are saved for the next step of PopCluster.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 SNP + \beta_2 PC_1 + ... + \beta_{n+1} PC_n + \beta_{n+2} x_1 + ... + \beta_{n+m+1} x_m, \tag{2.1}$$

where p is the probability of a subject having the phenotype usually expressed as 0 for its absence and 1 for presence; $\beta_0, \beta_1, ..., \beta_{n+m+1}$ are model parameters; and the variable SNP is typically coded by the number of coded alleles in the genotypes, i.e. additive genetic model. The model is adjusted by PC_1, ..., PC_n and additional covariates x_1, ..., x_m. The statistics from the logistic regression models, such as parameter estimates, standard errors and p-values, are saved by PopCluster for further analysis. The summary of this step is shown in Figure 2.2.

2.2.1.3 *Pruning of redundant clusters*

PopCluster was developed to identify SNPs that have varying effects in different ethnic groups, or sub-populations. Therefore, the core of PopCluster is a recursive algorithm to discover such clusters by comparing the genetic effect of each SNP in the sub-populations represented by two sibling clusters. PopCluster recursively parses the dendrogram bottom-up for every SNP under investigation by comparing the genetic effects of each pair of sibling clusters that have no children (Figure 2.3).

The algorithm first checks the following conditions for each pair of sibling clusters that have no children: (1) each cluster has at least 5 cases and 5 controls; (2) the minor allele frequency (MAF) of a SNP in each cluster is greater than 0.05; (3) one or both of the phenotype-SNP associations are statistically significant (p-value < 0.05). All of these conditions are user defined input parameters. If at least one of these conditions does not hold, PopCluster removes these sibling nodes from the list of clusters. Otherwise, PopCluster compares the SNPs' effects in the two sub-populations by calculating the statistic:

$$z = \frac{\hat{\beta_{1.1}} - \hat{\beta_{1.2}}}{\sqrt{\delta_{1.1}^2 + \delta_{1.2}^2}}, \tag{2.2}$$

where $\hat{\beta_{1.1}}$ and $\hat{\beta_{1.2}}$ are SNP effect estimates for two sibling clusters using the logistic regression model in Equation 2.1, and $\delta_{1.1}$ and $\delta_{1.2}$ are their standard errors. Under the assumption of at least 100 observations per cluster, the estimates $\hat{\beta_{1.1}}$ and $\hat{\beta_{1.2}}$ are approximately normally distributed and independent and therefore $z \sim N(0, 1)$ under the null hypothesis of no difference of the genetic effects.

Therefore, if $|z| < z_{\alpha/2}$, where $z_{\alpha/2}$ is the $(1 - \alpha/2)$ percentile of the standard

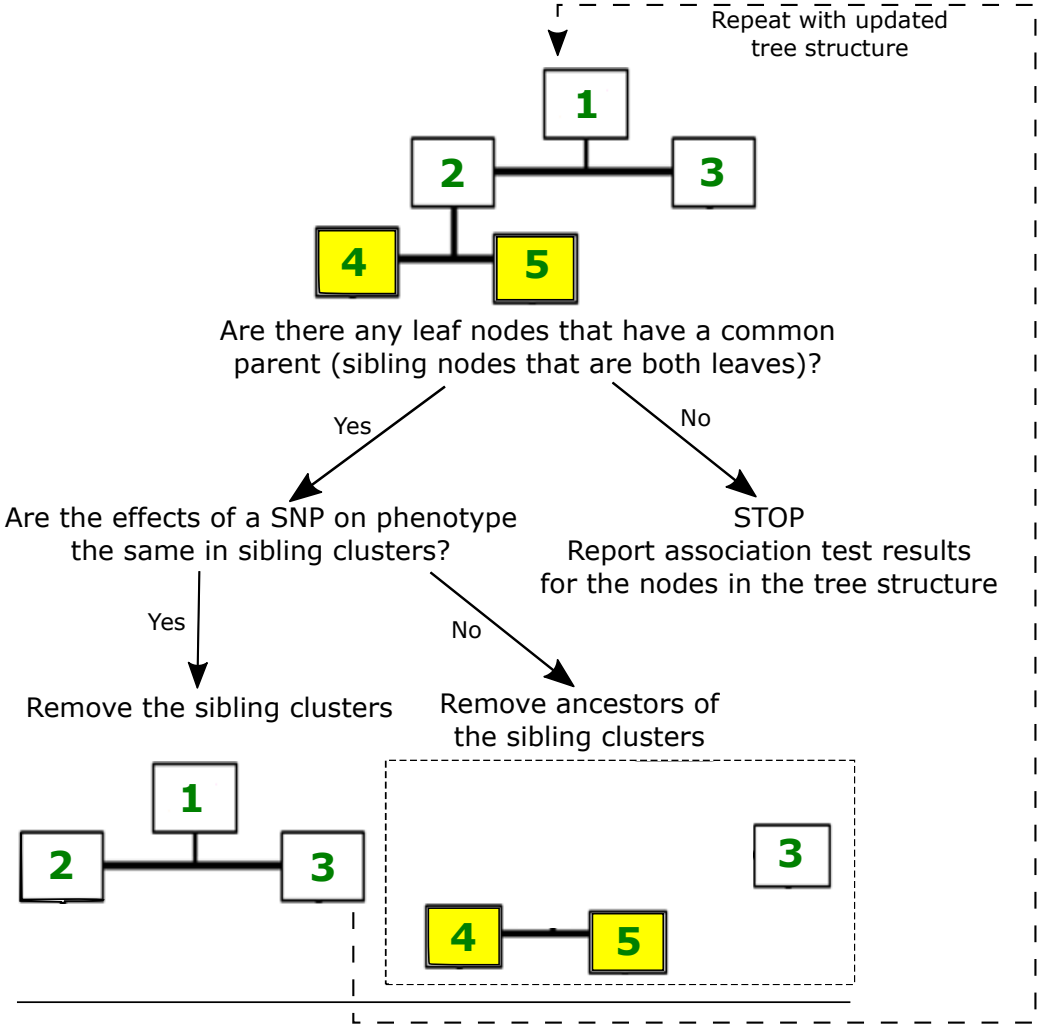

Figure 2.3: A schematic of the recursive pruning of redundant clusters. The dendrogram describing the final cluster in Figure 2.2 is recursively parsed bottom-up to identify clusters in which the genetic effects are not statistically different. Here, numbers in the dendrogram (1, 2, 3, 4, 5) are simply labels to distinguish between different clusters.

normal distribution, then we fail to reject the null hypothesis. In this case, $\hat{\beta_{1.1}}$ is statistically equivalent to $\hat{\beta_{1.2}}$, thus implying that the effects of the tested SNP in the two sibling clusters are equivalent, and PopCluster merges these nodes into their parent cluster, and removes them from the dendrogram. If $|z| \geq z_{\alpha/2}$ then $\hat{\beta_{1.1}} \neq \hat{\beta_{1.1}}$, and the results from the sibling nodes are included in the list of final results, and all the ancestors of these nodes are removed from the dendrogram.

The procedure parses the dendrogram until there are no sibling nodes that are both leaf nodes. The procedure is repeated separately for each SNP, and the output of PopCluster is a list of clusters for each SNP with all the relevant statistics. These clusters are non-overlapping, meaning no cluster has subjects that are in another cluster and each of the subject of the initial dataset is included in one of the clusters. If no population-specific effects are identified, the algorithm returns the original top cluster.

Reported SNP-phenotype associations are considered significant if the association between a SNP and the phenotype (β_1 in the Equation 2.1) in a cluster has a p-value less than a threshold α:

$$\alpha = \frac{0.05}{M \times N}, \tag{2.3}$$

where M is the total number of clusters that was reported by PopCluster for the SNP and N is the number of SNPs tested. By dividing 0.05 by M and N, we adjust the result for multiple comparisons.

2.2.2 Genotype and phenotype data

We used two different phenotypes and two distinct genome-wide genotype datasets to evaluate our algorithm. The first dataset is compiled from four case-

control studies of EL: the New England Centenarian Study (NECS) (Sebastiani & Perls, 2012), the Southern Italian Centenarian Study (SICS) (Malovini et al., 2011), the Longevity Gene Project (LGP) (Atzmon et al., 2004), and the Long Life Family Study (LLFS) (Newman et al., 2011) (Table 2.1). LLFS data are available via the database of Genotypes and Phenotypes (dbGaP) (dbGaP Study Accession: phs000397.v1.p1). The genotype data for all studies were generated using Illumina SNP arrays (Sebastiani et al., 2012) and imputed to the 1000 Genomes haplotypes phase I using IMPUTE2 following the standard protocol and quality control (Howie et al., 2012). All subjects provided informed consent approved by the study institutional review boards. The combined datasets contain several European ethnicities that have been well characterized. See Supplement Figure 1 in (Sebastiani et al., 2017b) for a characterization of European ethnicities in this data set using PCA. Cases are defined as individuals who lived past the 1 percentile survival age from the 1900 birth year cohort based on US Social Security Administration cohort life tables (Bell & Miller, 2005), i.e. age 96 and greater for males, and 100 years and greater for females. The details of the genotype data and the phenotype of EL are presented in (Sebastiani et al., 2017b; Andersen et al., 2012; Sebastiani et al., 2016a,b, 2017a).

In this dataset we used PopCluster to analyze a list of 371 SNPs that were previously found to be associated with EL with p-value < 5E-05 (Sebastiani et al., 2017b). To limit the problem of multiple comparisons, we also used PopCluster to re-analyze the association between the 11 SNPs in Table 2.2 that have been associated with EL with genome-wide significance (p-value < 5E-07) in (Sebastiani et al., 2017b).

In addition, we applied PopCluster to the multi-ethnic HRS (Sonnega et al.,

Table 2.1: Summary of studies of extreme longevity included in the analysis.

Study	Cases (median age, range)	Controls
SICS	174 (100, 96-109)	540
LGP	308 (102, 96-113)	621
LLFS	572 (100, 96-111)	2560
NECS	1084 (103, 96-119)	3102
Total	2138	6823

SICS: South Italian Centenarian Study; LGP: Longevity Genes Project; LLFS: Long Life Family Study; NECS: New England Centenarian Study; Cases (median age, range): number of cases with their median age and the range; Controls: number of controls.

Table 2.2: Subset of SNPs associated with EL.

SNP	Chr	Pos (hg38)	Ref/Alt	Genes
rs2008465	2	10014127	A/G	*GRHL1*, *KLF11*
rs28391193	4	110236842	G/A	*ELOVL6*, *HSBP1P2*
rs72834698	6	26176289	G/A	*HIST1H2BD*, *HIST1H2BE*
rs3764814	7	6150149	T/C	*USP42*
rs7976168	12	83044780	A/G	*TMTC2*
rs7185374	16	48416457	A/C	*SIAH1*
rs5882	16	44888997	A/G	*CETP*
rs6857	19	44888997	C/T	*APOE*
rs59007384	19	44893408	G/T	*TOMM40*
rs405509	19	44905579	T/G	*TOMM40*, *APOE*
rs769449	19	44906745	G/A	*APOE*

Chr: chromosome; Pos (hg38): position of a SNP in the Genome Reference Consortium Human Reference 38; Ref/Alt: reference and alternative alleles; Genes: closest gene/genes (annotation was done using SnpEff (Cingolani et al., 2012)).

2014) on the SNPs from Table 2.2 to search for ethnic-specific genetic effects on surviving past age 90. The HRS includes self-identified “White/Caucasian", “Black or African-American", and a few different groups of “Hispanic" subjects. Controls were subjects with age at last contact < 81. With this definition of cases and controls, the HRS dataset included 866 cases and 8469 controls. The HRS dataset is available through the HRS website (http://hrsonline.isr.umich.edu/) and dbGaP (dbGaP Study Accession: phs000428.v1.p1).

2.2.3 Evaluation

I evaluated PopCluster using a combination of real and simulated datasets. Here I outline the datasets and metrics used for the evaluation.

2.2.3.1 False positive rate

We used genotype data of SNPs in Table 2.2 from EL studies (Table 2.1) as one of the input parameters to PopCluster to evaluate its false positive rate (FPR). This list is a subset of the 371 SNPs described in Section 2.2.2.

In each simulation, we reshuffled the original labels of cases and controls or randomly generated the case/control labels before applying PopCluster. Therefore, by design, all significant associations detected are false positives. Specifically, we used four versions of the original dataset: the original dataset (8961 subjects) with (1) either the same number of cases and controls as in the original data (2138 cases and 6823 controls), and the case/control labels randomly reshuffled in each run, or (2) a randomly assigned even number of case and control labels: 4480 cases and 4481 controls. In addition, from the original dataset of 8961 subjects, related subjects from the same families were removed by selecting only one case and one

control for each family resulting in 7689 subjects. This reduced dataset was used with (3) either the same number of cases and controls as in the original data (1961 cases and 5728 controls), and the case/control labels randomly reshuffled in each run, or (4) a randomly assigned even number of case and control labels: 3844/3845 cases and controls. In addition to permuting case/control status in the overall datasets, we also performed two additional simulations in which the permutation of phenotype labels was done within each cluster in (1) the original dataset (8961 subjects), and (2) the reduced dataset without related individuals (7689 subjects).

We calculated the FPR for a simulation run as

$$FPR = \frac{FP}{FP + TN} = \frac{\sum_{i=1}^{N} \frac{s_i}{k_i}}{N} \tag{2.4}$$

where FP is the number of False Positives, TN is the number of True Negatives, N is the number of SNPs provided to the algorithm (11 in the case of our particular evaluation), s_i is the number of clusters (subpopulations) that were detected by PopCluster to have significant associations between a phenotype and an i-th SNP (FP), k_i is the total number of clusters detected by PopCluster for an i-th SNP ($FP + TN$). Correction for multiple comparisons was incorporated in the FPR evaluation by dividing the nominal significance level α by the total number of clusters detected by PopCluster for each SNP (k_i in Equation 2.4).

2.2.3.2 True positive rate

To estimate the true positive rate (TPR) of PopCluster, we simulated two scenarios with (1) a true association only in a selected subpopulation of subjects, and (2) a true association in the whole dataset. We compared the performances of PopCluster and traditional analysis without clustering in both simulated datasets.

In the first scenario, we simulated an allele A to be associated with EL in the selected group of 1905 subjects with 503 cases and 1402 controls characterized by two first genome-wide principal components calculated with the data of the studies in Table 2.1: $PC_1 \leq -0.005$ and $PC_2 \leq 0$. For the rest of the subjects, the allele was simulated not to be associated with the phenotype (Figure 2.4). Specifically, for the subjects with $PC_1 \leq -0.005$ and $PC_2 \leq 0$ different allele probabilities were assigned to cases and controls as

$$\begin{cases} Pr(A \mid EL) = p_1; \\ Pr(A \mid \overline{EL}) = p_2, \end{cases} \tag{2.5}$$

where p_1 is the probability of allele A in cases; p_2 is the probability of allele A in controls. The probabilities of allele dosages $0, 1, 2$ were generated assuming Hardy-Weinberg equilibrium. Various combinations of probabilities p_1 and p_2 (Equation 2.5) were tested to evaluate sensitivity and specificity of the algorithm to different risk differences. We chose $p_1 = \{0.05, 0.1, 0.25, 0.5\}$ to cover various scenarios with sufficient power with our sample size. For each p_1 value, we set the probability of allele A in controls to be $p_2 = p_1 + g$, where g is the difference in the allele frequency between cases and controls and $g = \{0.05, 0.075, 0.1, 0.125, 0.15\}$. Varying p_1 and g resulted in 20 different combinations of probabilities p_1 and p_2. For the rest of the subjects ($PC_1 > -0.005$ or ($PC_1 < -0.005$ and $PC_2 > 0$)), the allele A was simulated to be associated with PC_1 and PC_2, but not with the phenotype by setting

$$p_3 = Pr(A) = \frac{e^{\beta_0 + \beta_1 * PC_1 + \beta_2 * PC_2}}{1 + e^{\beta_0 + \beta_1 * PC_1 + \beta_2 * PC_2}}, \tag{2.6}$$

with $\beta_0 = -1$, $\beta_1 = -75$, and $\beta_2 = -50$ such that probabilities p_3 are not too

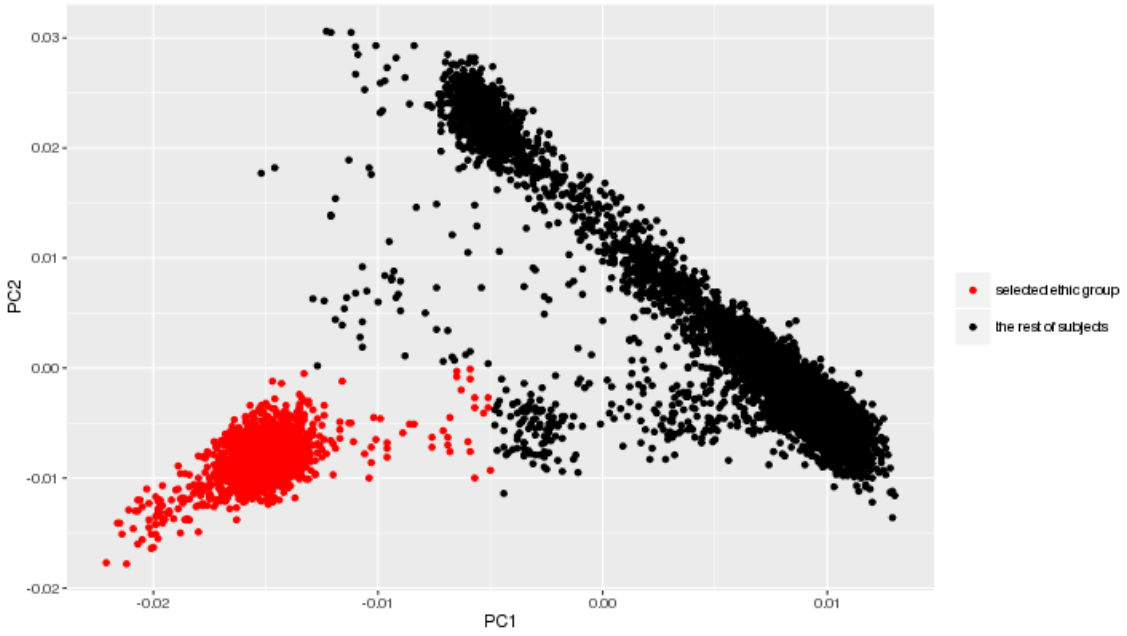

Figure 2.4: Scatter plot of the first two principal components with the selected region for the TPR evaluation. To evaluate the TPR we simulated an allele A to be associated with a phenotype of interest but not associated with principal components in a selected ethnic group: subjects with $PC_1 \leq -0.005$ and $PC_2 \leq 0$ (red dots). For the rest of the subjects (black dots), the allele was simulated to be significantly associated with PC_1 and PC_2.

extreme. The probabilities of allele dosages $0, 1, 2$ were again calculated assuming Hardy-Weinberg equilibrium.

Using the simulated allele data, we estimated the rate of PopCluster to discover the true clusters using the proportion of times the algorithm returned at least one cluster with more than 80% subjects from the region of association. In these cases, we evaluated the TPR of PopCluster for each of the simulation sets as

$$TPR = \frac{TP}{TP + FN}, \tag{2.7}$$

where TP is the number of true associations that PopCluster predicted to be sig-

nificant (positive). FN is the number of true associations that PopCluster found to be insignificant (negative). We define an association in a cluster to be true if more than 80% of subjects in the cluster are from the region of association. An association was significant if the p-value was less than a threshold α (Equation 2.3). We also compared the true effect size β and the estimated parameter value in each simulated data set to evaluate the precision of PopCluster.

To compare the performance of PopCluster with the traditional analysis, we also analyzed each simulated dataset using logistic regression adjusted for sex and the 4 principal components, and we calculated the proportion of associations found significant for each of the parameters combinations. Note that each significant association found with the traditional analysis is a true positive association in the subpopulation in which we simulated a true association, but a false positive association in the remaining subset.

In the second scenario, we simulated an allele A to be associated with EL in the whole dataset using the probabilities p_1 and p_2 (Equation 2.5). We conducted this analysis to compare the TPR of PopCluster and the traditional analysis when there is no heterogeneity in the association between the SNP and the phenotype in different clusters. To evaluate PopCluster's performance, we calculated the proportion of times PopCluster returned exactly one top cluster, and how often this cluster was identified as significant. We calculated the TPR of the traditional analysis as the proportion of significant associations detected in the simulated datasets.

2.3 RESULTS

2.3.1 Evaluation results

2.3.1.1 False positive rate

Figure 2.5 summarizes the results of the FPR evaluation. Each simulation was run 1000 times. On average, the estimated FPR in all six different simulations was $\sim 4\%$. This low FPR shows that the correction for multiple comparisons incorporated in Equation 2.4 is sufficient to bound the family-wise error rate by the level of significance used in the algorithm. Additionally, for each of the first four different simulation set-ups (when the permutation was done on a whole dataset), we report the information on how many significant clusters each run returned (Table 2.3). We define significant clusters here as clusters in which the association between the simulated phenotype and the SNP has a p-value less than 0.05 divided by the total number of clusters returned. By design, any association returned by the algorithm is a false positive, because we reshuffle the labels of cases and controls. As expected, the majority of runs returned no significant associations.

I also evaluated the FPR of PopCluster on a homogeneous subset of our data - LGP (Table 2.1). I did this to verify the FPR when there are no clusters in the study populations. In 100% of simulations, PopCluster returned one cluster - the whole LGP dataset - as a final result, and the FPR in this case is equivalent to the FPR of a traditional analysis that adjust for the population structure. On average, the estimated FPR in this evaluation was $\sim 5\%$.

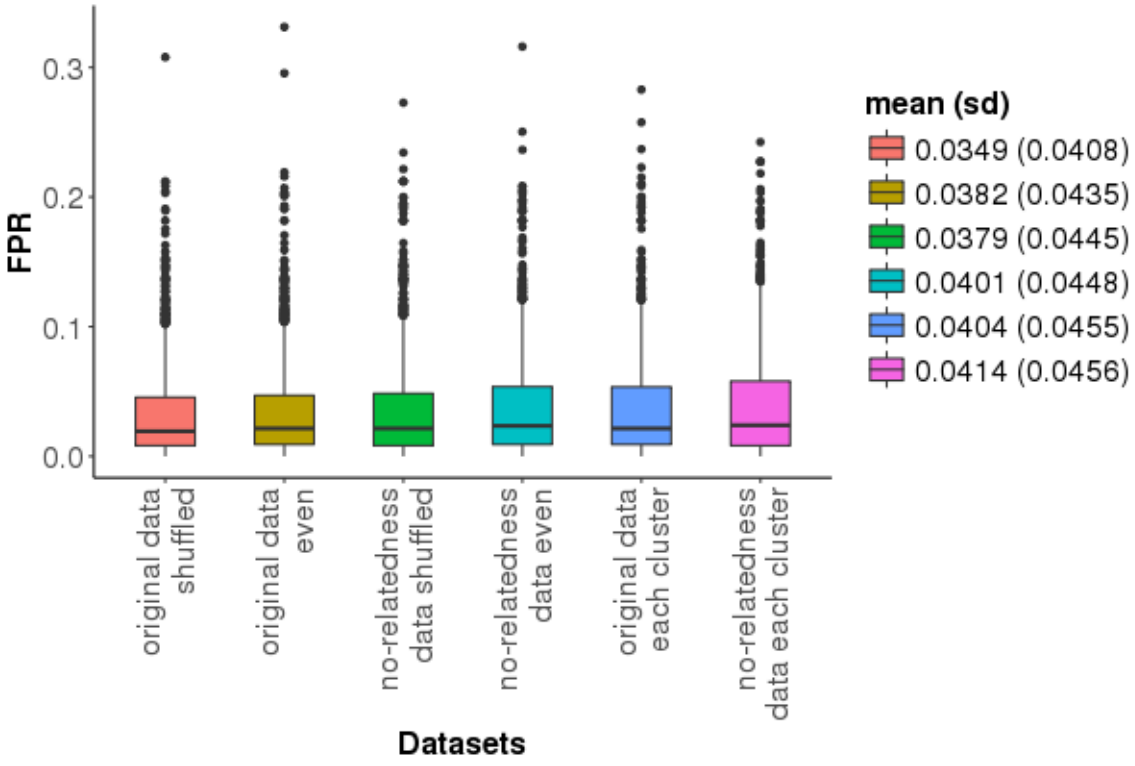

Figure 2.5: Boxplots of the FPR in six different simulations. Mean FPR and standard deviations (in parentheses) for each of the simulations are shown on the right of the Figure. "Original data shuffled": original dataset with random reshuffling of cases and controls in a whole dataset. "Original data even": original dataset with equal number of cases and controls randomly generated. "No-relatedness data shuffled": as "original data shuffled" after we removed related individuals. "No-relatedness data even": as "original data even" after we removed related individuals. "Original data each cluster": original dataset with random reshuffling of cases and controls in each cluster. "No-relatedness data each cluster": as "original data each cluster" after we removed related individuals.

Table 2.3: Percentage of simulation runs (FPR) that returned n number of significant associations (Clusters) ($n = \{0, 1, 2, 3\}$).

Clusters	Simulation 1	Simulation 2	Simulation 3	Simulation 4
0	83.8	83.2	84.1	83.8
1	15.6	15.8	15.1	15.4
2	0.6	0.3	0.8	0.8
3	–	0.02	–	–

Simulation 1: "Original data shuffled" (original dataset with random reshuffling of cases and controls). Simulation 2: "Original data even" (original dataset with equal number of cases and controls randomly generated). Simulation 3: "No-relatedness data shuffled" (as "original data shuffled" after we removed related individuals). Simulation 4: "No-relatedness data even" (as "original data even" after we removed related individuals).

2.3.1.2 True positive rate

The boxplots in Figure 2.6 summarize the results of the evaluation of the PopCluster's TPR for all the combinations of probabilities of allele A in cases and controls when allele A was simulated to be associated with phenotype only in selected region (scenario 1). For each combination of parameters, simulations were run 1000 times. The percent of simulation runs that returned at least one cluster with more than 80% subjects in the region of association was 97.6% (Table 2.4), and the average number of these "true association" clusters was 2.6 (Table 2.5). The TPR of PopCluster increases with the increase in difference in allele frequencies between cases and controls. High TPR values in the simulations with larger differences in allele frequencies suggest that the algorithm can detect clusters of significant association. Low TPR values for smaller differences in allele frequencies indicate that the dataset does not have enough power to detect those fine associations. In addition, we find that the differences between true effect size β and estimated $\hat{\beta}$ are symmetrically distributed around 0 as expected (Figure 2.7). The black circles with

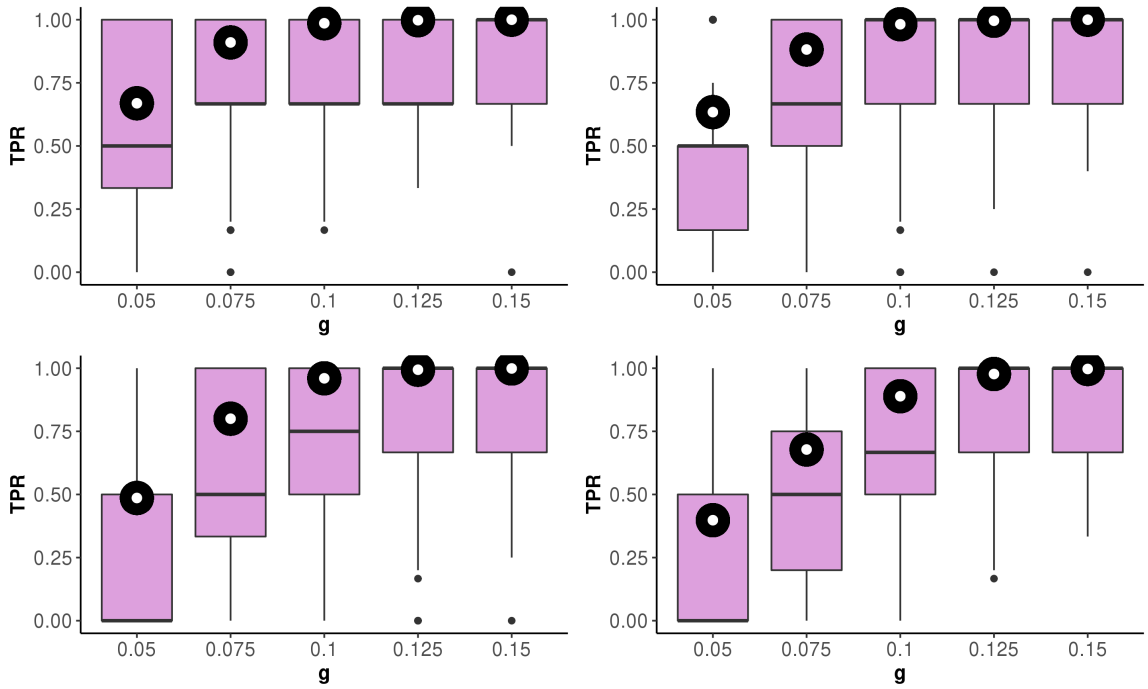

Figure 2.6: Boxplots of the TPR for various combinations of probabilities of allele A in cases and controls. g is the difference in allele probabilities between cases and controls. $P(A)$ denotes the probability of allele A in cases. (A): $P(A) = 0.05$. (B): $P(A) = 0.1$. (C): $P(A) = 0.25$. (D): $P(A) = 0.5$. Black circles with white centers represent how often the traditional analysis finds a general association to be significant.

white centers on boxplots in Figure 2.6 depict the proportion of general associations found by the traditional analysis. These proportions are comparable to the TPR of PopCluster; however they represent only TPR for finding a general association, but every TP in this case is a FP for a group of subjects in which allele A was simulated not to be be associated with a phenotype.

The average TPR of the traditional analysis in datasets simulated to have an association between allele A and phenotype in the whole dataset (scenario 2) was 100%, meaning all of the runs returned an association as significant. The average number of times PopCluster returned only one cluster as a result was 30% (Table

Table 2.4: Percentage of simulation runs (TPR) that returned at least one cluster with more than 80% subjects in the region simulated to have an association between allele A and the phenotype (scenario 1).

p_1/g	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
0.05	99.4	99.9	100	100	99.9
0.1	95.8	100	99.8	99.9	99.9
0.25	86.5	96.4	99.4	99.7	99.9
0.5	82.4	94	99.5	100	100

p_1 is the probability of an allele A in cases in the region where allele A was simulated to be associated with the phenotype. $p_2 = p_1 + g$ is the probability of an allele A in controls in the same region. g is the difference in allele frequencies A between cases and controls in this region.

Table 2.5: Average number of returned clusters with more than 80% subjects in the region simulated to have an association between allele A and the phenotype (scenario 1).

p_1/g	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
0.05	2.4	2.7	2.8	3.0	3.2
0.1	2.3	2.6	2.8	2.9	3.0
0.25	2.0	2.4	2.5	2.7	2.8
0.5	1.9	2.2	2.5	2.6	2.8

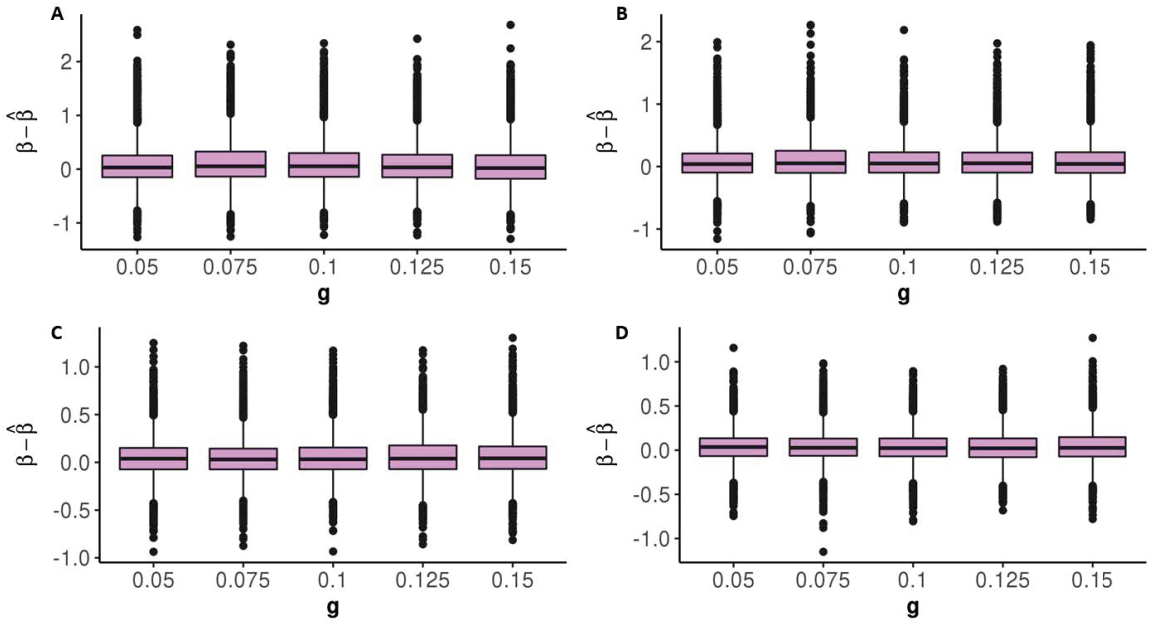

Figure 2.7: Boxplots of the differences between true effect β and estimated effect $\hat{\beta}$ in the clusters where allele A was simulated to be associated with EL (scenario 1). Each set of boxplots corresponds to the boxplots in Figure 2.6.

Table 2.6: Percentage of simulation runs (TPR) of PopCluster that returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset (scenario 2).

p_1/g	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
0.05	39.4	39.5	36.1	39	38
0.1	30.5	31	31.8	29.9	30.3
0.25	26.5	26.5	20.7	25.7	25.7
0.5	27.3	22.6	26.5	22.6	22

Table 2.7: Percentage of simulation runs (TPR) of PopCluster that returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset with re-shuffled case/control labels (scenario 2).

p_1/g	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
0.05	38.9	41.4	41.9	40.3	39.7
0.1	29.1	28.8	27	25.4	26.2
0.25	25.7	21.9	23.9	23.3	21.6
0.5	25.6	24.2	21.9	21.4	20.3

2.6). Among those single clusters, 100% of them were found to have a significant association between the simulated allele and phenotype. We evaluated PopCluster's performance in the case of allele-phenotype association simulated in the whole dataset in three more additional simulation set-ups: (1) re-shuffled case/-control labels (Table 2.7); (2) balanced number of re-shuffled case/control labels (Table 2.8); (3) balanced number of re-shuffled case/control labels in each cluster (Table 2.9).

2.3.2 Application to real data

We used PopCluster to re-analyze the association of the set of 371 SNPs with EL in the data summarized in Table 2.1. We assessed whether the algorithm could detect more significant associations than the analysis that adjusts for population structure, and identify subpopulations in which the associations were not significant.

Table 2.8: Percentage of simulation runs (TPR) of PopCluster that returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset with balanced number of re-shuffled case/control labels (scenario 2).

p_1/g	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
0.05	30.3	24.9	26	27	25.5
0.1	22.8	22.5	23.9	22.2	22.6
0.25	24.5	22.9	25.8	21.1	22.2
0.5	22.9	22.8	21.7	20.8	22.5

Table 2.9: Percentage of simulation runs (TPR) of PopCluster that returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset with balanced number of re-shuffled case/control labels in each cluster (scenario 2).

p_1/g	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
0.05	31.4	25.2	24.5	26	26
0.1	23.9	22.3	23.2	20.4	22.9
0.25	23.9	21.2	21.9	23.2	21.5
0.5	23.7	22.6	23	21.3	20

The analysis identified 14 SNPs in the *APOE* region that reached genome-wide level of significance in at least one cluster and although none of these cluster-specific associations was more significant than the results in the meta-analysis in (Sebastiani et al., 2017b), the analysis suggests that the effect of *APOE* on EL may vary with ethnicity. In addition, PopCluster identified a large cluster of 7401 subjects in which the association between SNP rs2008465 (Table 2.2) and EL was more significant than in the meta-analysis, and smaller clusters comprising mainly North East Europeans in which the association between rs2008465 and EL was not significant. For complete results returned by PopCluster on the analysis of 371 SNPs and EL see Table 2.10 and Figures 2.8-2.17. To interpret ethnic groups from PCA plots, please refer to Figure 2.18.

Below we present an example of SNP, rs3764814, with ethnic-specific effect on

Table 2.10: Complete list of clusters detected by PopCluster for 371 SNPs and EL.

https://open.bu.edu/handle/2144/29809

Following the link you will find *El-results.csv* file with all the results for the analysis of PopCluster on 371 SNPs and EL. Column *Cluster* contains labels for the clusters, all of which are visualized on the PCA plots in Figures 2.8-2.17. Labels reflect cluster sizes, e.g. cluster labeled *118* has 118 subjects, and sorted by their size. Legend for the table in the *EL-results.csv* file: OR: odds ratio for EL in carriers of the allele; P-value: p-value of the association.

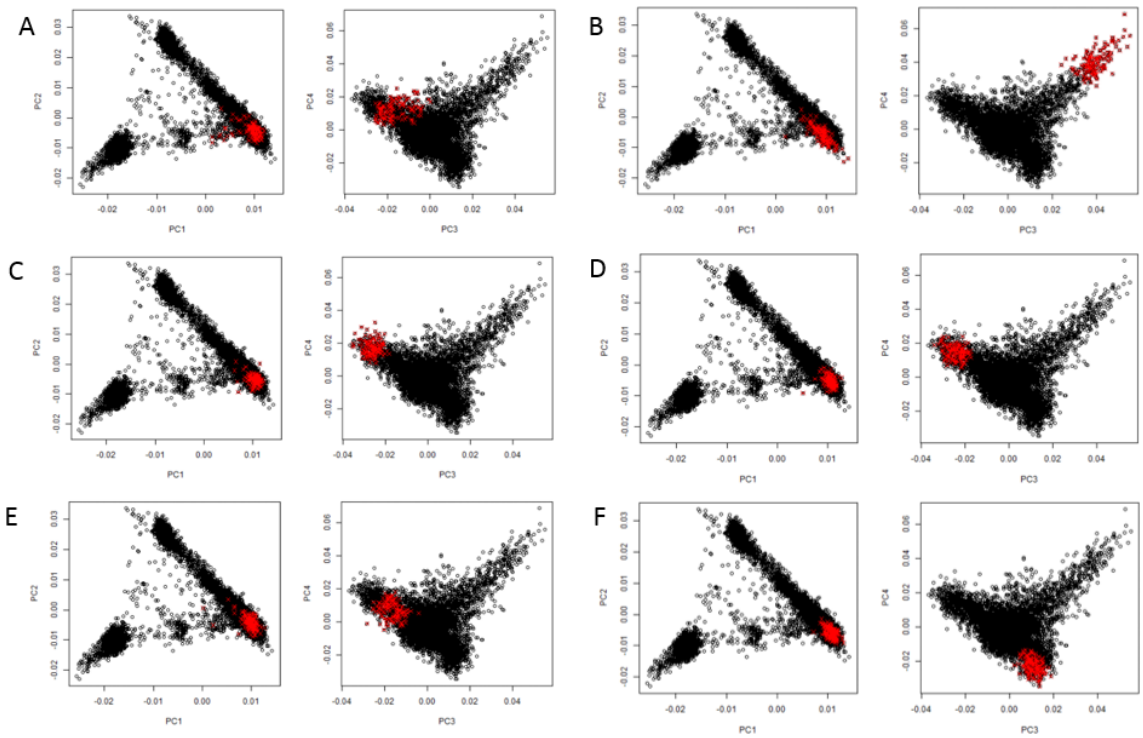

Figure 2.8: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 1. Subjects are colored red if they belong to (A): cluster 118, (B): cluster 126, (C): cluster 128, (D): cluster 129, (E): cluster 133, (F): cluster 134. Subjects not belonging to respective clusters are colored black in every plot.

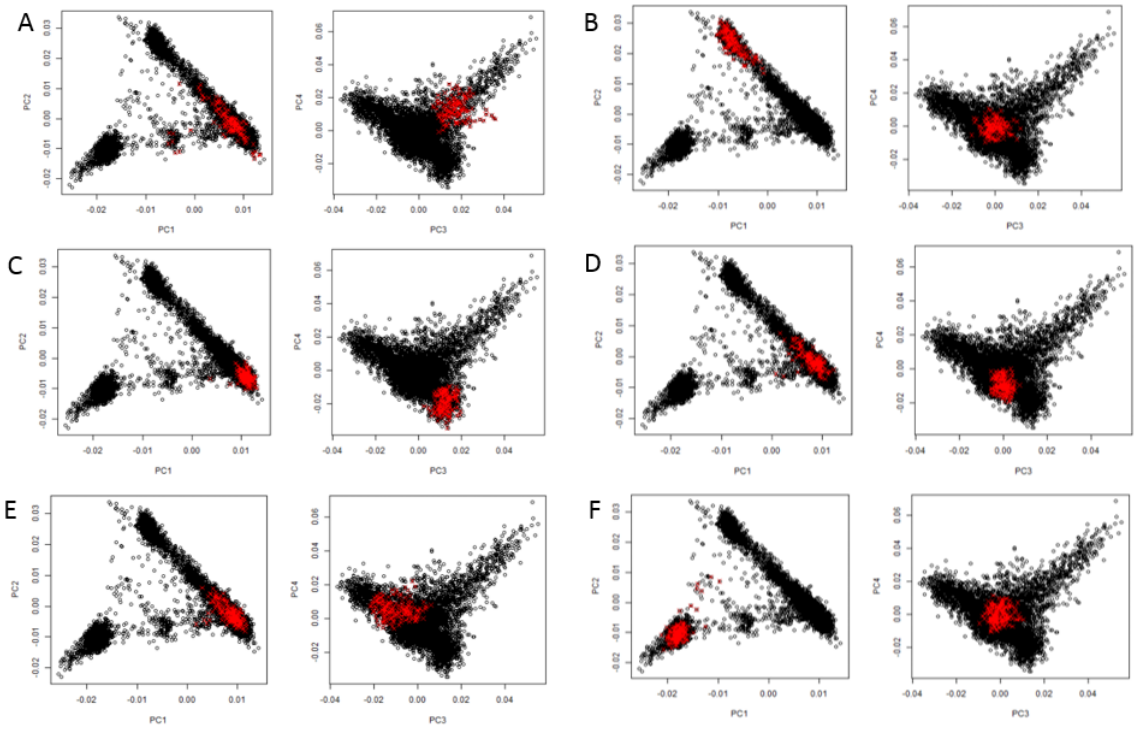

Figure 2.9: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 2. Subjects are colored red if they belong to (A): cluster 141, (B): cluster 148, (C): cluster 153, (D): cluster 160, (E): cluster 170, (F): cluster 176. Subjects not belonging to respective clusters are colored black in every plot.

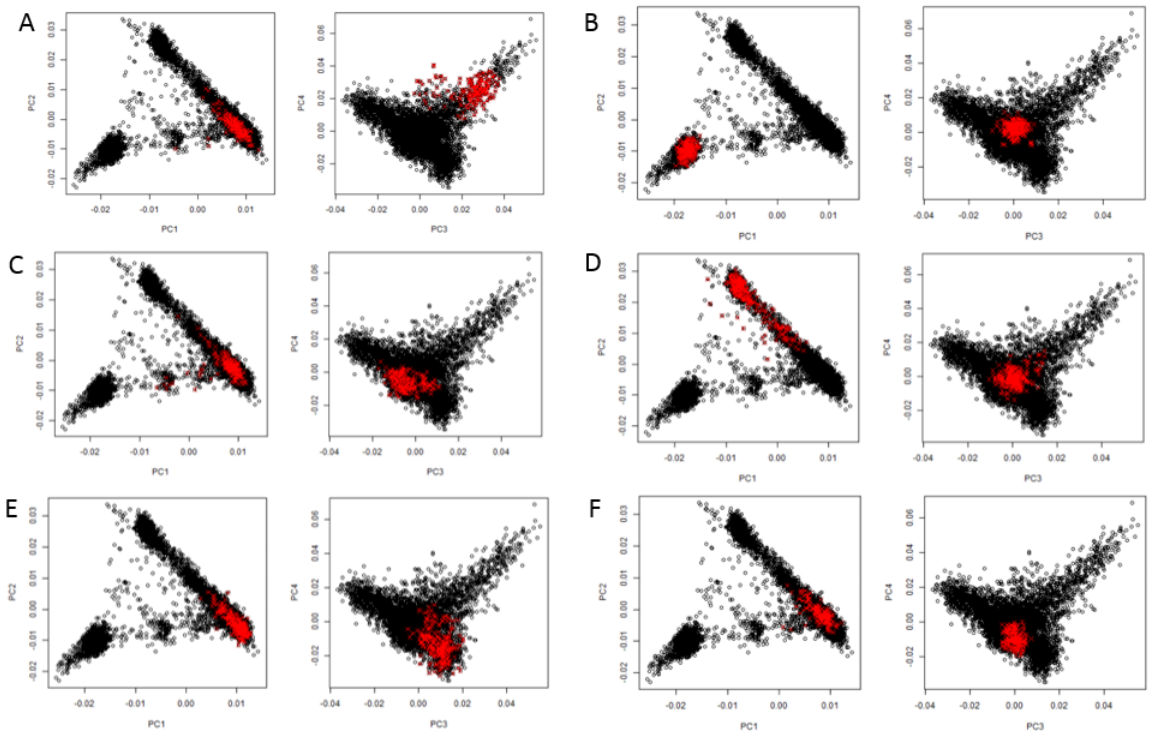

Figure 2.10: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 3. Subjects are colored red if they belong to (A): cluster 180, (B): cluster 193, (C): cluster 194, (D): cluster 240, (E): cluster 249, (F): cluster 253. Subjects not belonging to respective clusters are colored black in every plot.

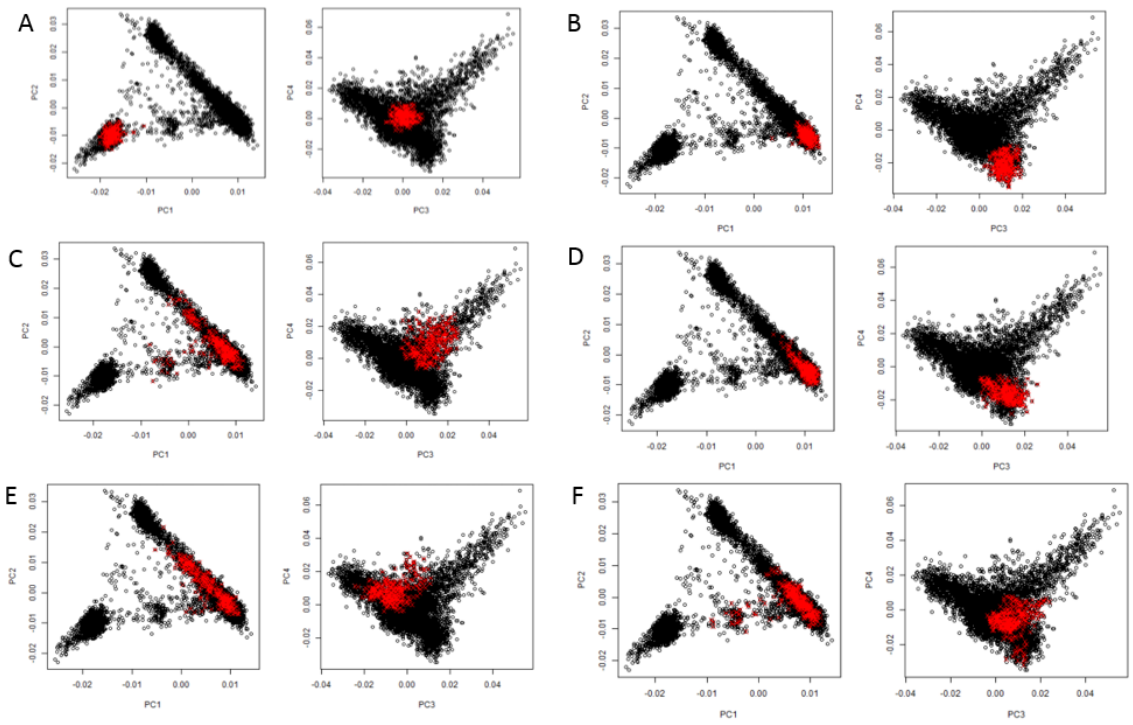

Figure 2.11: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 4. Subjects are colored red if they belong to (A): cluster 274, (B): cluster 287, (C): cluster 290, (D): cluster 296, (E): cluster 297, (F): cluster 316. Subjects not belonging to respective clusters are colored black in every plot.

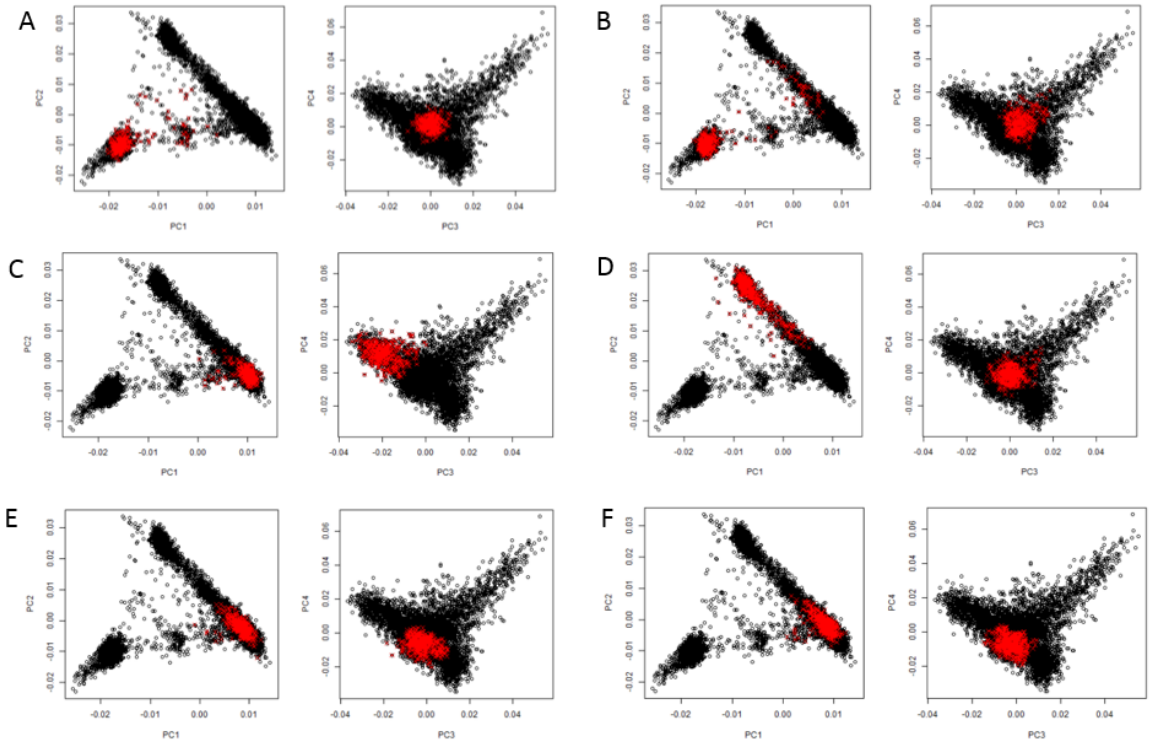

Figure 2.12: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 5. Subjects are colored red if they belong to (A): cluster 330, (B): cluster 348, (C): cluster 380, (D): cluster 388, (E): cluster 396, (F): cluster 413. Subjects not belonging to respective clusters are colored black in every plot.

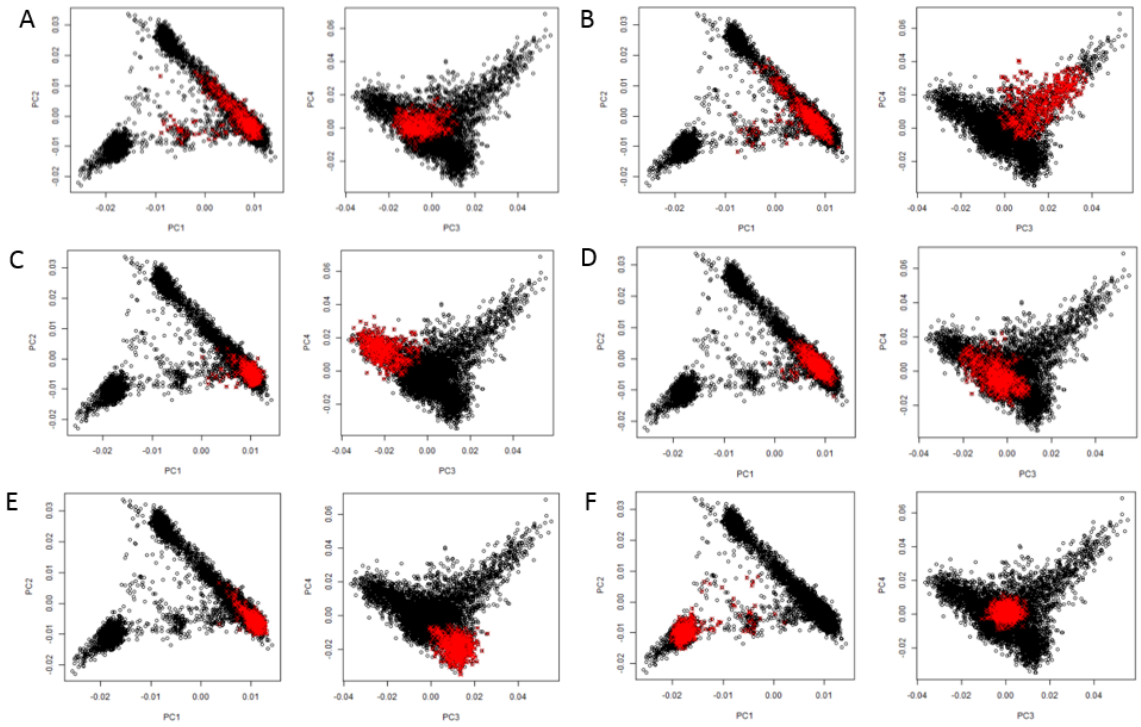

Figure 2.13: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 6. Subjects are colored red if they belong to (A): cluster 416, (B): cluster 470, (C): cluster 508, (D): cluster 566, (E): cluster 583, (F): cluster 604. Subjects not belonging to respective clusters are colored black in every plot.

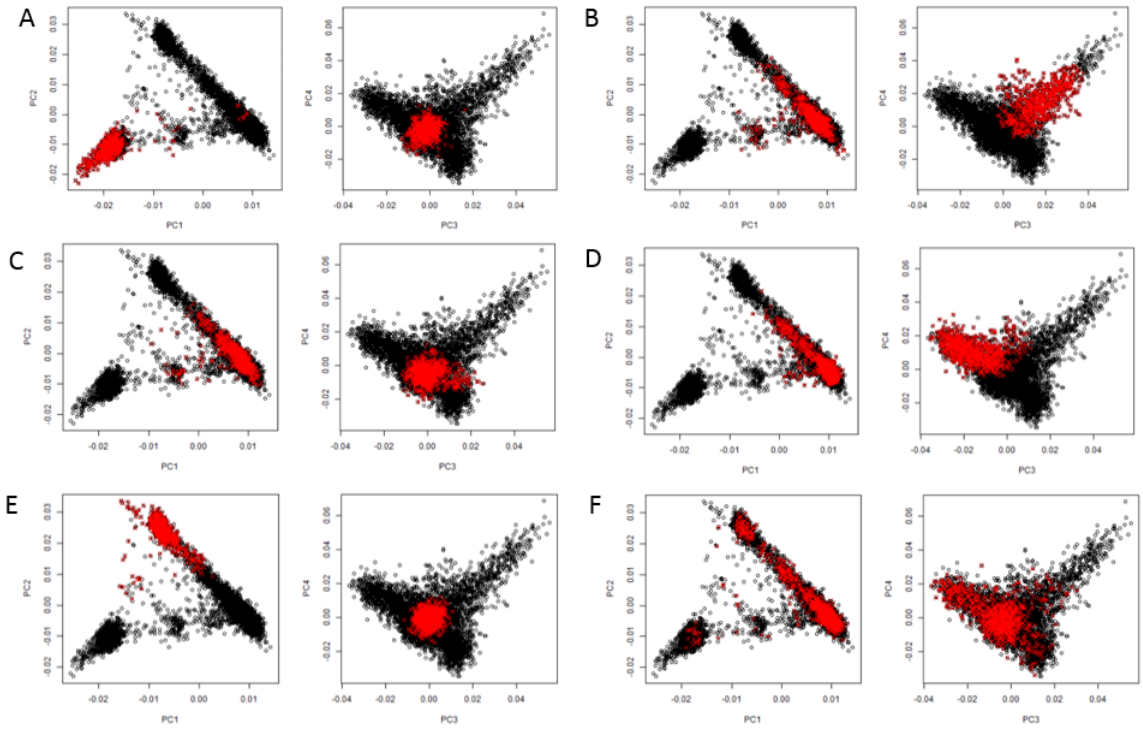

Figure 2.14: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 7. Subjects are colored red if they belong to (A): cluster 606, (B): cluster 611, (C): cluster 721, (D): cluster 805, (E): cluster 818, (F): cluster 828. Subjects not belonging to respective clusters are colored black in every plot.

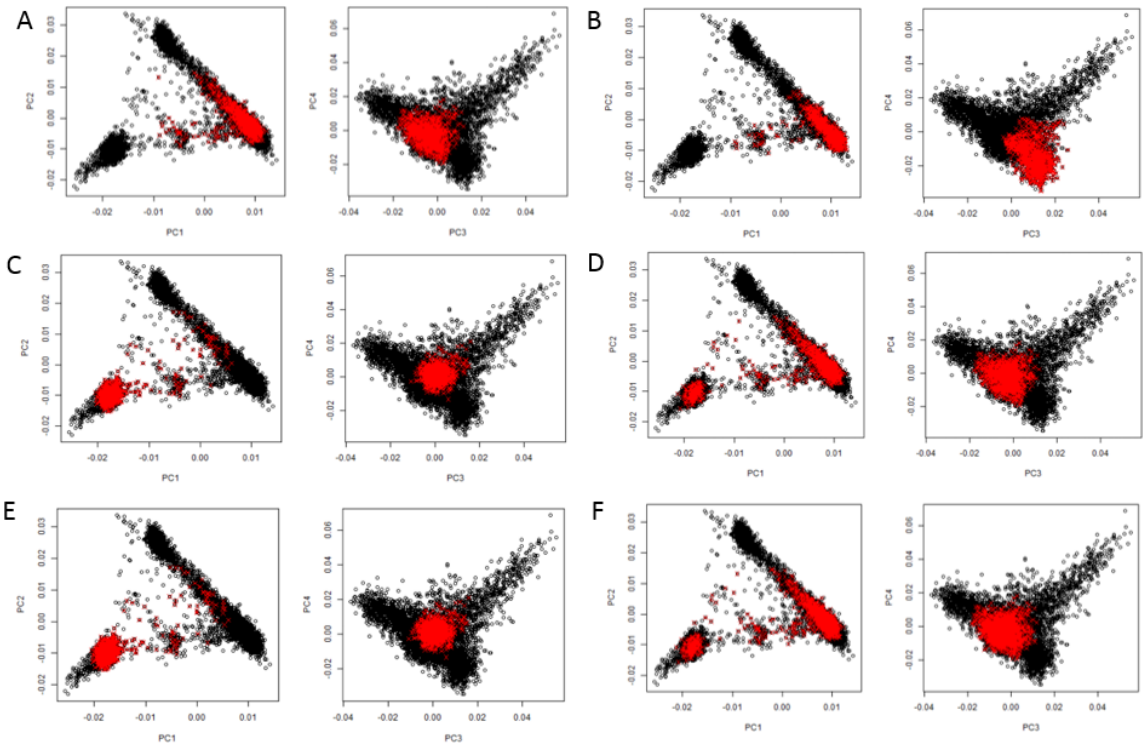

Figure 2.15: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 8. Subjects are colored red if they belong to (A): cluster 829, (B): cluster 899, (C): cluster 952, (D): cluster 1005, (E): cluster 1145, (F): cluster 1199. Subjects not belonging to respective clusters are colored black in every plot.

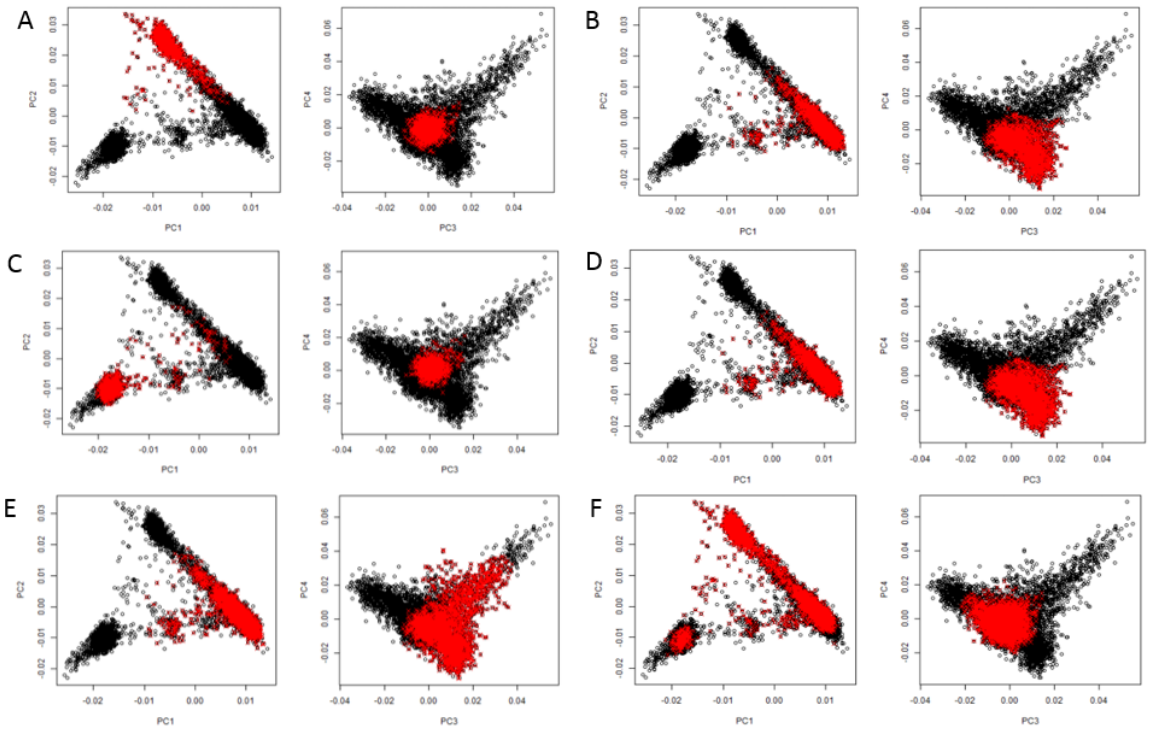

Figure 2.16: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 9. Subjects are colored red if they belong to (A): cluster 1206, (B): cluster 1620, (C): cluster 1765, (D): cluster 1869, (E): cluster 2480, (F): cluster 2971. Subjects not belonging to respective clusters are colored black in every plot.

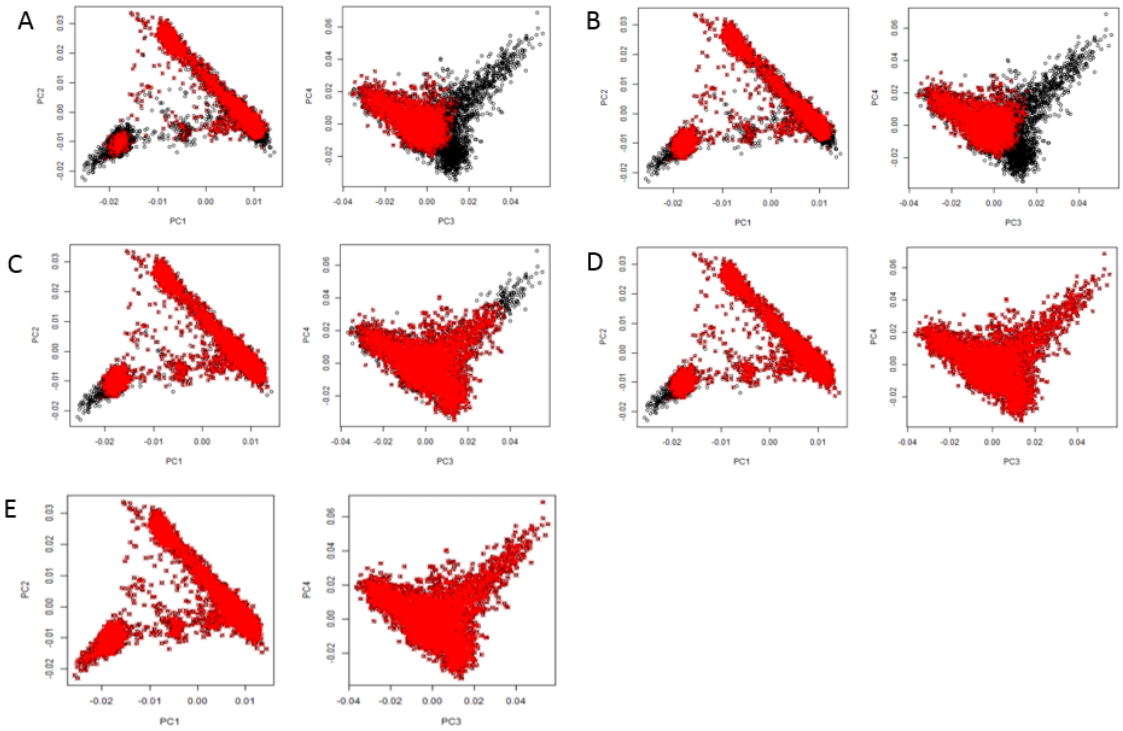

Figure 2.17: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL - set 10. Subjects are colored red if they belong to (A): cluster 3776, (B): cluster 4921, (C): cluster 7401, (D): cluster 8355, (E): cluster 8961. Subjects not belonging to respective clusters are colored black in every plot.

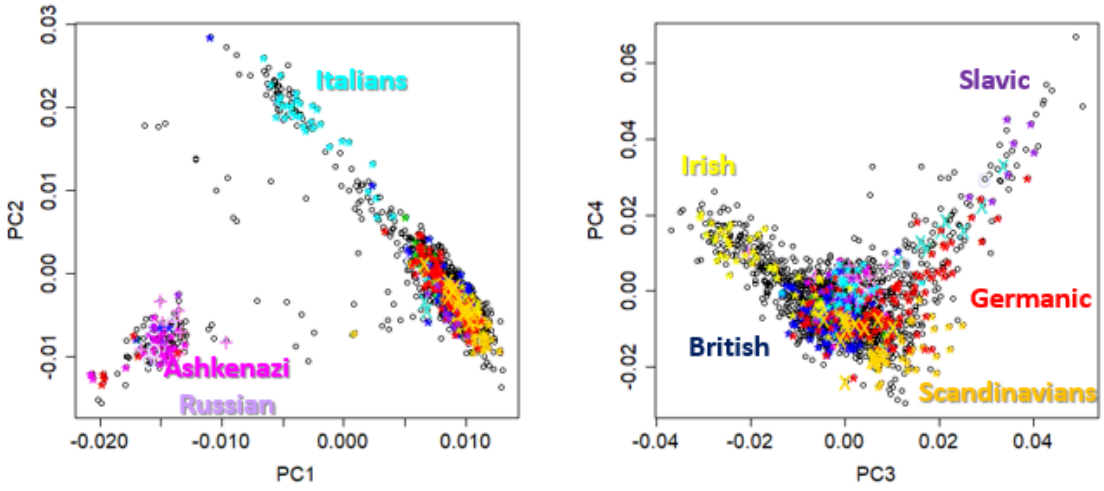

Figure 2.18: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of subjects in the NECS. Subjects are labeled by ethnicity using the information about mother tongue and places of birth of subjects and their parents (Sebastiani et al., 2012).

EL in sub-populations of Europeans. It also appears to have an ethnic-specific effect on surviving past age 90 in the HRS dataset. To account for the varying sample sizes of clusters, we computed the power to detect significant associations in clusters using the G*Power software (Faul et al., 2009).

2.3.2.1 rs3764814 and extreme longevity

rs3764814 is a coding SNP in the gene *USP42* which is located on chromosome 7. We recently found this SNP to be very strongly associated with EL in Europeans ignoring population specific effects (Sebastiani et al., 2017b). The global MAF of rs3764814 is 0.28, but it becomes much rarer in Europeans: 0.07. The MAF of rs3764814 in our dataset is 0.09 and it increases 1.5 times in centenarians as compared to controls: 0.12 in cases and 0.08 in controls. Table 2.11 summarizes the results of PopCluster analysis for rs3764814 on EL. Figure 2.19 presents a hierar-

Table 2.11: Complete list of clusters for rs3764814 and EL.

Cluster	OR	95% CI	P-value	MAF	Power, %
828	2.24	[1.49, 3.36]	9.87E-05	0.086	100
316 (583)	2.89	[1.61, 5.17]	0.0004	0.085	100
721	1.91	[1.31, 2.79]	0.0007	0.093	100
805 (2971)	2.22	[1.37, 3.60]	0.001	0.088	100
611	2.03	[1.23, 3.35]	0.006	0.075	100
2971 (805)	1.3	[1.07, 1.57]	0.009	0.089	100
1145	1.47	[1.08, 1.99]	0.01	0.105	100
126	3.75	[1.28, 11.00]	0.02	0.075	100
606	1.61	[1.07, 2.42]	0.02	0.094	100
249	1.3	[0.63, 2.71]	0.48	0.064	50
583 (316)	0.85	[0.49, 1.49]	0.57	0.071	47

Cluster: label for the cluster which reflect cluster size, e.g. cluster labeled *583* consists of 583 subjects (If in the final dendrogram structure, a cluster has a sibling, it is reported here in parentheses.); OR: odds ratio for EL in carriers of the allele; 95% CI: 95% confidence interval for the OR; P-value: p-value of the association; MAF: minor allele frequency in the cluster; Power, %: power of detecting a given OR with a given number of subjects.

chical tree of this dataset with the clusters returned for this SNP as highlighted in yellow. PopCluster identified two clusters (clusters *249* and *583* in Table 2.11 and black dots in Figure 2.20) in which the association of rs3764814 did not reach statistical significance. Since clusters 249 and 583 are not sibling clusters, we can only conclude that there is no significant association of rs3764814 and EL in these two groups. Note that this is different than saying the effects are the same. In Figure 2.20, the subjects depicted as red dots belong to clusters for which the association between SNP rs3764814 and EL is significant or borderline significant. Using partially known information on subjects' ancestry, such as birth places and native languages of grandparents (Solovieff et al., 2010), we identified the subjects without an association as being enriched of Danish descent (Sebastiani et al., 2017c).

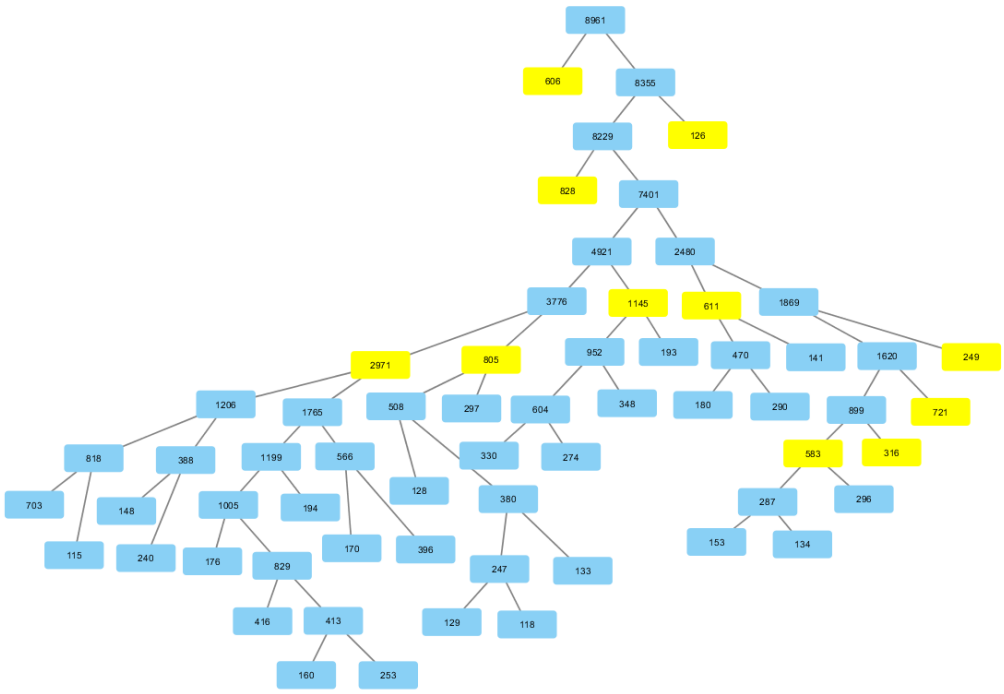

Figure 2.19: Full hierarchical tree structure returned for the analysis of EL dataset before pruning of redundant clusters step. Highlighted in yellow are the clusters that were identified by PopCluster as having ethnic-specific effects of SNP rs3764814 on EL. Numbers inside the nodes represent the number of subjects in each cluster. Visualization was done using Cytoscape (Shannon et al., 2003).

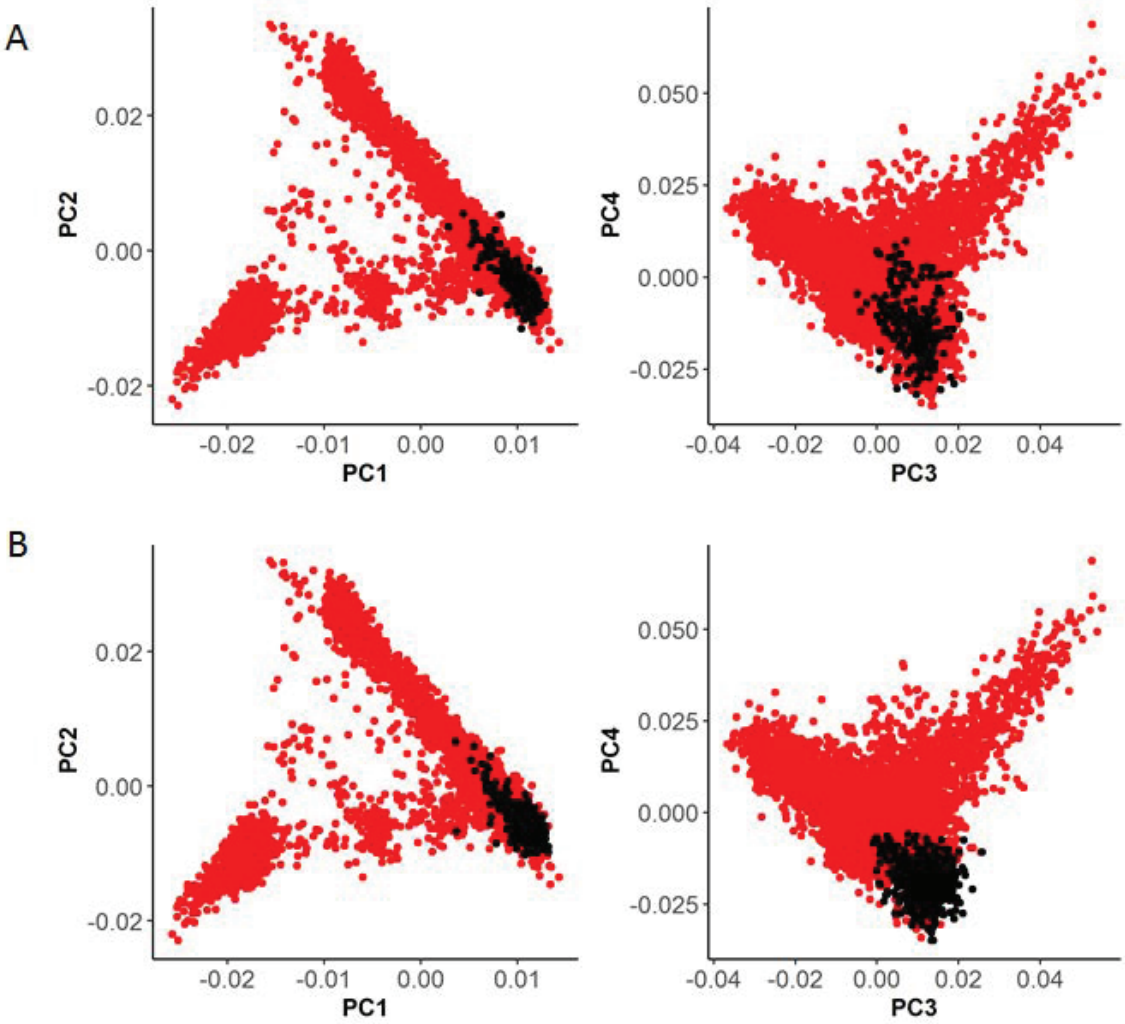

Figure 2.20: Ethnic groups in which the effect of SNP rs3764814 on EL did not reach statistical significance. The scatter plots display the principal components PC1-PC4 calculated using genome-wide genotype data of all subjects in the study of EL. Subjects colored in black belong to (Panel A): cluster *249*, (Panel B): cluster *583* as defined in Table 2.11.

Table 2.12: Complete list of clusters for rs72834698 returned as an output from PopCluster run on HRS dataset with phenotype of surviving past age 90.

Cluster	OR	95% CI	P-value	MAF	Power, %
8128 (236)	1.26	[1.06, 1.50]	0.008	0.098	100
236 (8128)	0.34	[0.10, 1.16]	0.09	0.083	100
811	0.64	[0.31, 1.28]	0.21	0.075	100
160	1.01	[0.50, 2.01]	0.99	0.181	5

Table 2.13: Complete list of clusters detected by PopCluster for 11 SNPs and HRS.

https://open.bu.edu/handle/2144/29809

Following the link you will find *HRS-results.csv* file with all the results for the analysis of PopCluster on 11 SNPs and surviving past age 90. Column *Cluster* contains labels for the clusters. Only 3 out of 11 SNPs have significant associations with a phenotype of surviving past the age of 90: rs72834698, rs5882, and rs405509. Legend for the table in the *HRS-results.csv* file: OR: odds ratio for surviving past 90 in carriers of the allele; P-value: p-value of the association.

2.3.2.2 rs72834698 and survival past age 90

We used PopCluster to analyze the association between SNP rs72834698 and surviving past the age of 90 in the HRS dataset. The analysis identified one large cluster of 8128 subjects in which this SNP had a significant association with survival past age 90 (cluster *8128* in Table 2.12 and red dots in Figure 2.22-A). Note that this association is not significant after correction for multiple testing. Based on self-reported ethnicity labels provided with HRS dataset, the group of subjects that is not in this cluster (black dots in Figure 2.22-A) is enriched of "Hispanic, Mexican" subjects. Figure 2.21 presents a hierarchical tree with clusters returned for rs72834698 in yellow. For more results on this analysis, see Table 2.13 and Figure 2.22.

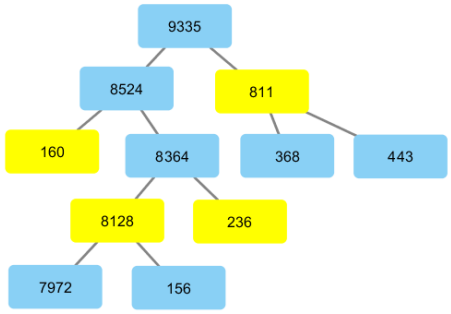

Figure 2.21: Full hierarchical tree structure returned for the analysis of survival past age 90 HRS dataset before pruning of redundant clusters step. Highlighted in yellow are the clusters that were identified by PopCluster as having ethnic-specific effects of SNP rs3764814 on surviving past age 90. Numbers inside the nodes represent the number of subjects in each cluster.

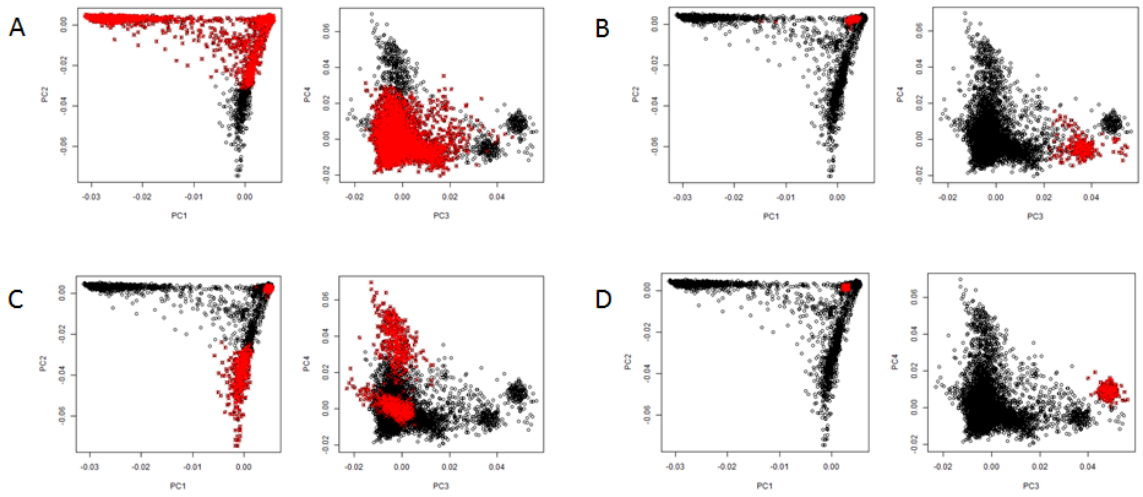

Figure 2.22: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the HRS. Subjects are colored red if they belong to (A): cluster 8128, (B): cluster 236, (C): cluster 811, (D): cluster 160. Subjects not belonging to respective clusters are colored black in every plot.

2.4 DISCUSSION

Currently most of the genetics studies are based on data generated in subjects of specific European ancestry, and sometimes the results of the genetic association studies do not generalize to other populations (Martin et al., 2017). The issue of underrepresentation of non-European populations in genetic studies is slowly being addressed (Popejoy & Fullerton, 2016); and it is important to adapt current techniques to account for the different allele frequencies and genetic effects in those populations. There are methods that have been proposed to account for the heterogeneity of variants and phenotype associations in different populations. For example, the generalized linear mixed model association test (GMMAT) accounts for population stratification and varying binary phenotype frequencies in different populations (Chen et al., 2016). The GMMAT corrects p-values and effect estimates in the genetic association studies in the presence of non-constant mean-variance re-

lationship for a binary phenotype; however, it does not identify the varying effect sizes in the populations. Another approach, XP-BLUP, predicts individual genetic risk scores for heterogeneous subjects by incorporating multi- and trans- ethnic information in the analysis (Coram et al., 2017). The novelty of PopCluster is to provide a heuristic search to discover heterogeneous effects when the sub-populations are unknown.

There are many consequences of not being able to identify the varying genetic effects in the studies that consist of only or a majority of European samples. This problem is particularly important in genetic association studies that aim to discover new drug targets. Currently there are several high-selling medications that do not help or even hurt the majority of people who take them (Schork, 2015). Another area that would benefit the delineation of population specific genetic effects is genetic risk prediction. When a genetic marker for a trait is identified using predominantly European populations, using this marker for prediction of disease risk in non-Europeans may result in a higher false positive diagnostic rate (The PLOS Medicine Editors et al., 2016; Manrai et al., 2016).

Various factors, such as genetics, diet, lifestyle and endemic infectious diseases, contribute to varying allele frequencies and genetic effects in different populations (Petrovski & Goldstein, 2016; Kelly et al., 2017; Rosenberg et al., 2002). In addition, different genetic markers can be associated with the same disease phenotype in different populations (Schork, 1997). PopCluster performs the association studies in populations with varying genetic effects on a phenotype to account for the diverse ancestry and environmental backgrounds.

PopCluster can also be used as a step before performing meta-analysis when working with multi and trans-ethnic studies. The algorithm facilitates identifi-

cation of populations with heterogeneous genetic effects. Subsequently, separate GWASes can be performed on the detected sub-populations, and the results can be combined using tools such as transethnic meta-analysis (Morris, 2011).

In the evaluation we tested datasets with a small number of related individuals ($\sim$ 14%) and the algorithm worked well in those cases. However, when the number of related individuals is large, proper corrections for relatedness are important. In our implementation of the algorithm, we use the R geeglm function from the geepack package (Hojsgaard et al., 2006) to fit the regression model. If the dataset includes related individuals, PopCluster can use a generalized estimating equation (GEE) to adjust for within-family correlation (Wang et al., 2013). In our examples we only used a binary phenotype. However, in the implementation of PopCluster, there is an option to choose the probability distribution of the outcome in the regression model so the algorithm can be used to analyze continuous phenotypes.

PopCluster has several limitations that I outline below. One of the limitations of our algorithm is that even though it finds ethnicity-specific associations that otherwise would have been missed, breaking the dataset into smaller clusters makes the association testing less powerful. Additionally, if the initial dataset has a small number of samples that belong to genetically very different group compared to the rest of the samples, PopCluster might not be able to identify the presence of ethnicity-dependent effects as it would not process clusters below the root node of the dendrogram (Figure 2.1). In such situation, I recommend to remove these distinct samples from the dataset, and re-run PopCluster on the updated set of samples. In situations when genetic variants do not have heterogeneous associations with a phenotype in different populations, PopCluster might lead to overfit-

ting and identify differential associations between clusters. Thus, it is important to have a replication for all the findings. Another constraint is that PopCluster accepts the data with quality control performed beforehand. For example, systematic differences in genotyping of the data could bias the principal component analysis. In our examples, we performed quality control on genome-wide genotype data so that highly polymorphic regions and SNPs in high LD are removed, and that the strand direction is consistent for all the studies, etc. However, some additional sources of bias may always be possible and it might be useful to verify that the clusters represent ethnical differences if appropriate label data for some of the subjects are known. In our examples, we verified that the clusters represent European ethnicities using subjects and their parents places of birth, or mother tongue. This step is not necessary, but it is an addition to validating the results.

Overall, I am hopeful that the use of the PopCluster's methodology will contribute to more precise estimate of genetic associations in the presence of population heterogeneity and ultimately better use of genetic findings in precision medicine.

CHAPTER 3

Varying effect of *APOE* alleles on extreme longevity in European ethnicities

3.1 INTRODUCTION

Apolipoprotein E (APOE) is a class of proteins involved in lipid metabolism with functions determined by alleles of the gene *APOE*. The gene has 3 alleles e2, e3, and e4 defined by combinations of genotypes of the SNPs rs7412 and rs429358 (Schachter et al., 1994; Weisgraber et al., 1981). APOE is a well-studied gene with multiple effects on aging and longevity. The e4 allele is a well-established risk factor for late onset of Alzheimers disease (Corder et al., 1993; Yip et al., 2005; Liu et al., 2013). We and others have demonstrated that having an APOE e4 allele has a deleterious effect on longevity that decreases the odds to reach extreme human lifespan (Schachter et al., 1994; Sebastiani et al., 2019). The e3 allele is the "neutral allele" in many ethnicities, while e2 is the allele that promotes longevity and healthy aging (Schachter et al., 1994; Sebastiani et al., 2019; Schupf et al., 2013; Wu & Zhao, 2016).

The frequency of the *APOE* alleles varies among human populations (Corbo & Scacchi, 1999). For example, the most common e3 allele frequency varies from 54% in African Pygmies to 90% in Southern Italians and Sardinians. The frequency of the e4 allele varies from 5% in Sardinians to 41% in African Pygmies. It has been reported that the frequency of the e4 allele increases with latitude due to the natural selection to protect against low-cholesterol levels (Eisenberg et al., 2010). Additionally, the e4 allele is associated with better resistance to adverse non-industrialized environments, specifically to parasites and infections in children (van Exel et al., 2017; Trumble et al., 2017).

Studies have suggested that *APOE* e4 has different effects on the risk of Alzheimer's disease in Europeans, African-Americans and Hispanics (Liu et al., 2013; Hendrie et al., 2014; Campos et al., 2013), while the role of ethnicity on the effect of *APOE* e2 on longevity and neuroprotection is unknown. To investigate the ethnic-specific effect of *APOE* e2 and e4 on extreme longevity, I used PopCluster algorithm (described in Chapter 2) to search for ethnically different clusters of Europeans in which the effect of *APOE* e2 and e4 on extreme longevity changes.

3.2 MATERIALS AND METHODS

3.2.1 Study Populations

We used genome-wide genotype data from a consortium of four studies of EL and healthy aging: the SICS, the LGP, the LLFS, and the NECS (Table 2.1). The SICS is a study of longevity that focused enrollment of long lived individuals in the South of Italy (Malovini et al., 2011). The LGP is a study of longevity that enrolled long lived individuals who were of Ashkenazy Jewish descent, survived to at least age 95 years old, and were dementia free at the time of enrollment (Barzilai et al., 2003). Some siblings, offspring and spouses of offspring were also enrolled and additional unrelated population controls were selected based on lack of familial longevity. The LLFS is a family-based study of healthy aging and longevity that recruited approximately 550 families and 5,000 family members selected for familial longevity (Newman et al., 2011; Ash et al., 2009). Participants were enrolled at three American field centers (Boston, Pittsburgh and New York), and a European field center in Denmark. The NECS is a study of centenarians, some of the long-lived siblings, offspring, offspring spouses and additional unrelated controls selected because their parents died before reaching the median age survival of their

birth year cohort (Sebastiani & Perls, 2012). The study recruits centenarians worldwide. All subjects consented and the studies were approved by local Institutional Review Boards (IRBs).

3.2.2 Genotype Data

Genome-wide genotype data for all studies were generated using Illumina SNP arrays (Sebastiani et al., 2012) and imputed to the 1000 genomes haplotypes phase I using IMPUTE2 and standard protocol (Howie et al., 2012). Imputation was preceded by pre-phasing with SHAPEIT (Delaneau et al., 2012). More details on the datasets can be found in in (Sebastiani et al., 2017b). *APOE* alleles were inferred from SNPs rs7412 and rs429358 that were either genotyped using real time PCR or imputed using IMPUTE2 (Sebastiani et al., 2019). Cases were defined as individuals who lived past the 1 percentile survival age from the 1900 birth year cohort based on US Social Security Administration cohort life tables (Bell & Miller, 2005), i.e. age 96 and greater for males, and 100 years and greater for females. Controls were defined either as individuals who died before reaching the threshold age, or as random subjects from the general population. The combined datasets contain several European ethnicities and information about place of birth and mother/father tongue.

3.2.3 Statistical Analysis

I used PopCluster to find ethnic-specific clusters of subjects with varying effect of *APOE* e2 and e4 on EL. PopCluster discovers subsets of individuals characterized by significantly different effects of a genetic variant by, first, clustering subjects based on principal components of genome-wide genotype data and, then, recur-

Table 3.1: *APOE* genotype distribution in the studies of extreme longevity.

genotypes	e2e2	e2e3	e3e3	e3e4	e4e4
No. subjects	56	1362	5901	1497	126

sively comparing the effect of a variant on phenotype between genetically closest clusters (Gurinovich et al., 2019).We used logistic regression adjusted by sex and four principal components calculated from the genome-wide genotype data of subjects to estimate cluster-specific associations between *APOE* alleles and EL. To adjust p-values for multiple testing we use p < 0.05/(number of clusters returned by the algorithm). Our evaluation showed that this correction maintains a family-wide error rate < 5%.

To evaluate the effect of *APOE* e2 and e4 on EL independently of each other, we conducted two analyses. In one analysis, we removed all carriers of e4, and used an additive genetic model with e3e3 coded as 0 (5901 subjects), e2e3 as 1 (1362 subjects) and e2e2 as 2 (56 subjects). Similarly, to evaluate the effect of e4 on EL we removed all carriers of e2, and used an additive genetic model with e3e3 coded as 0 (5901 subjects), e3e4 as 1 (1497 subjects) and e4e4 as 2 (126 subjects). Total genotype counts are presented in Table 3.1.

3.3 RESULTS

I analyzed the ethnic-specific association of EL with *APOE* alleles in 2143 cases of EL, and 6825 controls summarized in Table 2.1. PopCluster identified 13 ethnic-specific clusters in which the effect of *APOE* e2 on EL varied (Table 3.2), but only 2 with a significant, positive effect on EL after correction for multiple comparisons ($p < 0.05/13 = 0.0038$) (first two rows in Table 3.2). Similarly, PopCluster

identified 12 ethnic-specific clusters in which the effect of *APOE* e4 on EL varied (Table 3.3), and in 5 clusters *APOE* e4 was significantly and negatively associated with EL after correcting for multiple comparisons ($p < 0.05/12 = 0.004$) (first five rows in Table 3.3). Partially known information on subjects' ancestry, such as birth places and native languages of grandparents, is available for about 63% of subjects (27). We labelled each cluster with ethnicity based on the following rules followed in order: (1) If more than 50% of subjects in a cluster are enriched of a certain ethnicity, the cluster was labeled by that ethnicity. (2) If there is no known ethnicity with more than 50% subjects represented in a cluster, but the scatter plots of genome-wide principal components (Figures 3.1-3.3) of a cluster are localized, the best guess was made based on the scatter plots. (3) If neither of two previous conditions hold, a cluster was labeled as mixed.

Figures 3.1-A and -B display scatter plots of the first 4 genome-wide principal components for the 2 clusters in which PopCluster discovered significant effects of *APOE* e2. In both clusters, carriers of e2 have increased odds for EL compared to carriers of e3e3: $OR = 2.24$, 95% CI: 1.35, 3.73 in the cluster enriched of Danish ancestry, and $OR = 2.12$, 95% CI: 1.44, 3.13 the cluster enriched of Ashkenazi Jewish ancestry. The difference of effects between these two clusters however did not reach statistical significance ($p = 0.87$). In all other clusters, the genetic association did not reach statistical significance although several clusters had sample size > 180 that is the minimum size required to have 80% power to detect an odds ratio of 2, assuming a level of significance of 0.004.

Figures 3.2-D, 3.3-C, and 3.1-E, -A, -B display the scatter plots of the first 4 genome-wide principal components for the 5 clusters in which the association between *APOE* e4 and EL is statistically significant, after correction for multiple test-

Table 3.2: Associations between *APOE* e2 and EL in ethnic-specific clusters.

Label	No. subjects	Enriched ethnicity	OR	95% CI	p
Ashkenazi Jews_1 (*)	977	Ashkenazi Jews (0.64)	2.12	1.44, 3.13	0.0002
Danish_1 (*)	437	Danish(0.93)	2.24	1.35, 3.73	0.002
British_1(**)	452	British (0.10)	1.78	1.08, 2.93	0.02
Polish (**)	150	Polish (0.10)	2.5	0.95, 6.57	0.06
Mixed_1 (**)	974	Danish (0.47)	1.38	0.92, 2.06	0.127
Mixed_2 (**)	828	Germans (0.06)	1.43	0.91, 2.25	0.122
South Italians (*)	1309	South Italians (0.77)	1.27	0.92, 1.76	0.14
Irish (**)	1141	Irish (0.04)	1.36	0.88, 2.10	0.16
Ashkenazi Jews_2 (*)	235	Ashkenazi Jews (0.70)	1.52	0.74, 3.14	0.25
British_2 (**)	964	British (0.10)	0.83	0.58, 1.17	0.29
Danish_2 (*)	198	Danish (0.89)	0.71	0.29, 1.73	0.45
Danish_3 (**)	766	Danish (0.26)	1.13	0.75, 1.69	0.56
Russians (**)	757	Russians (0.03)	1.13	0.71, 1.78	0.61

Label: This was inferred by either the enriched ethnicity of subjects (>50% in the cluster) (*) or based on the PCA plots for this cluster (**) (underscore and number mean that there are more than 1 distinct cluster that are labeled the same). **No. subjects**: total number of subjects in the cluster; **Enriched ethnicity**: enriched ethnicity with proportion of subjects with enriched ethnicity from the subjects with known information on their ancestry; **OR**: odds ratio for EL comparing carriers of one copy of *APOE* e2 to e3e3 carriers; **p**: p-value (association is significant if $p < 0.05/13 = 0.004$).

Table 3.3: Associations between *APOE* e4 and EL in ethnic-specific clusters.

Label	No. subjects	Enriched ethnicity	OR	95% CI	p
British_3 (**)	1416	British (0.10)	0.3	0.21, 0.44	4.42E-10
Danish_4 (**)	559	Danish (0.33)	0.44	0.26, 0.72	0.001
Mixed_1 (**)	974	Danish (0.47)	0.49	0.32, 0.78	0.002
Ashkenazi Jews_1 (*)	977	Ashkenazi Jews (0.64)	0.48	0.30, 0.77	0.003
Danish_5 (*)	635	Danish (0.92)	0.47	0.28, 0.78	0.004
Russians (**)	757	Russians (0.03)	0.51	0.31, 0.83	0.006
Irish (**)	1141	Irish (0.04)	0.56	0.35, 0.91	0.02
Mixed_2 (**)	828	Germans (0.06)	0.57	0.33, 0.99	0.046
Ashkenazi Jews_2 (*)	235	Ashkenazi Jews (0.70)	0.43	0.18, 1.01	0.05
South Italians (*)	1309	South Italians (0.77)	0.82	0.56, 1.20	0.31
Polish (**)	150	Polish (0.10)	0.52	0.10, 2.61	0.42
British_4 (**)	207	British (0.06)	1.32	0.54, 3.22	0.54

OR: odds ratio for EL comparing carriers of one copy of *APOE* e4 to e3e3 carriers; p: p-value (association is significant if $p < 0.05/12 = 0.004$).

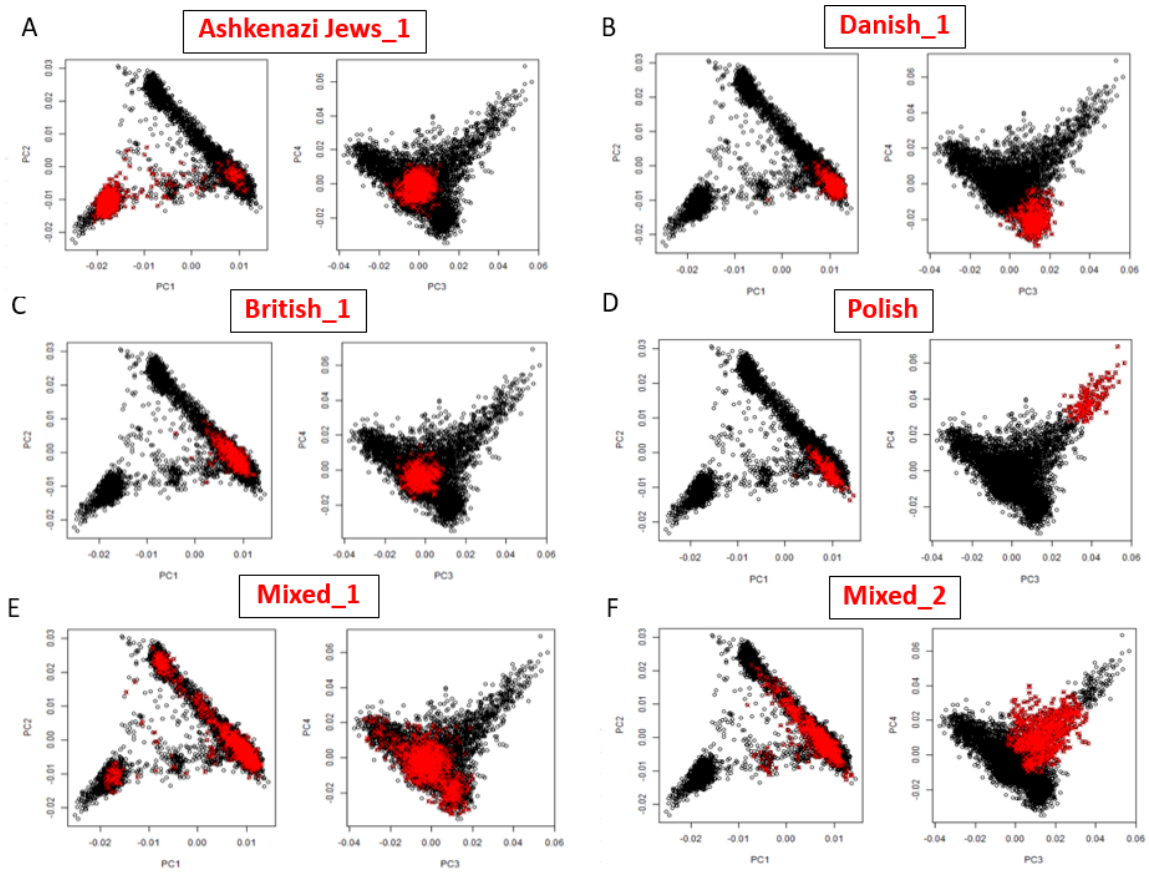

Figure 3.1: Scatter plots of principal components PC1-PC2 and PC3-PC4 from genome-wide genotype data of all subjects in the study of EL. Subjects are colored red if they belong to (A): cluster with 977 subjects enriched of Ashkenazi Jews; (B): cluster with 437 subjects enriched of Danish subjects; (C) cluster with 452 subjects enriched of British subjects; (D) cluster with 150 subjects enriched of Polish subjects; (E) cluster with 974 subjects of mixed ancestry; (F) cluster with 828 subjects of mixed ancestry.

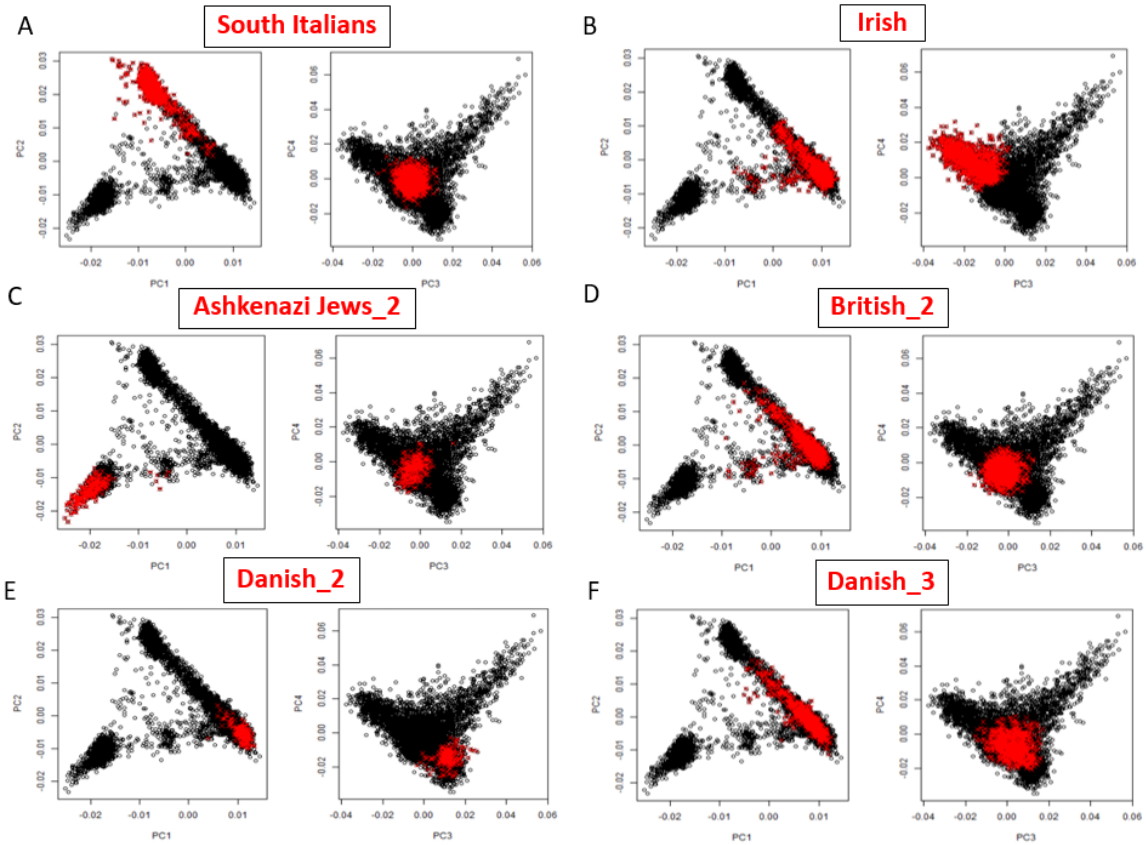

Figure 3.2: Scatter plots of principal components PC1-PC2 and PC3-PC4 from genome-wide genotype data of all subjects in the study of EL. Subjects are colored red if they belong to (A): cluster with 1309 subjects enriched of Italians; (B): cluster with 1141 subjects enriched of Irish subjects; (C) cluster with 235 subjects enriched of Ashkenazi Jews; (D) cluster with 964 subjects enriched of British subjects; (E) cluster with 198 subjects enriched of Danish subjects; (F) cluster with 766 subjects enriched of Danish subjects.

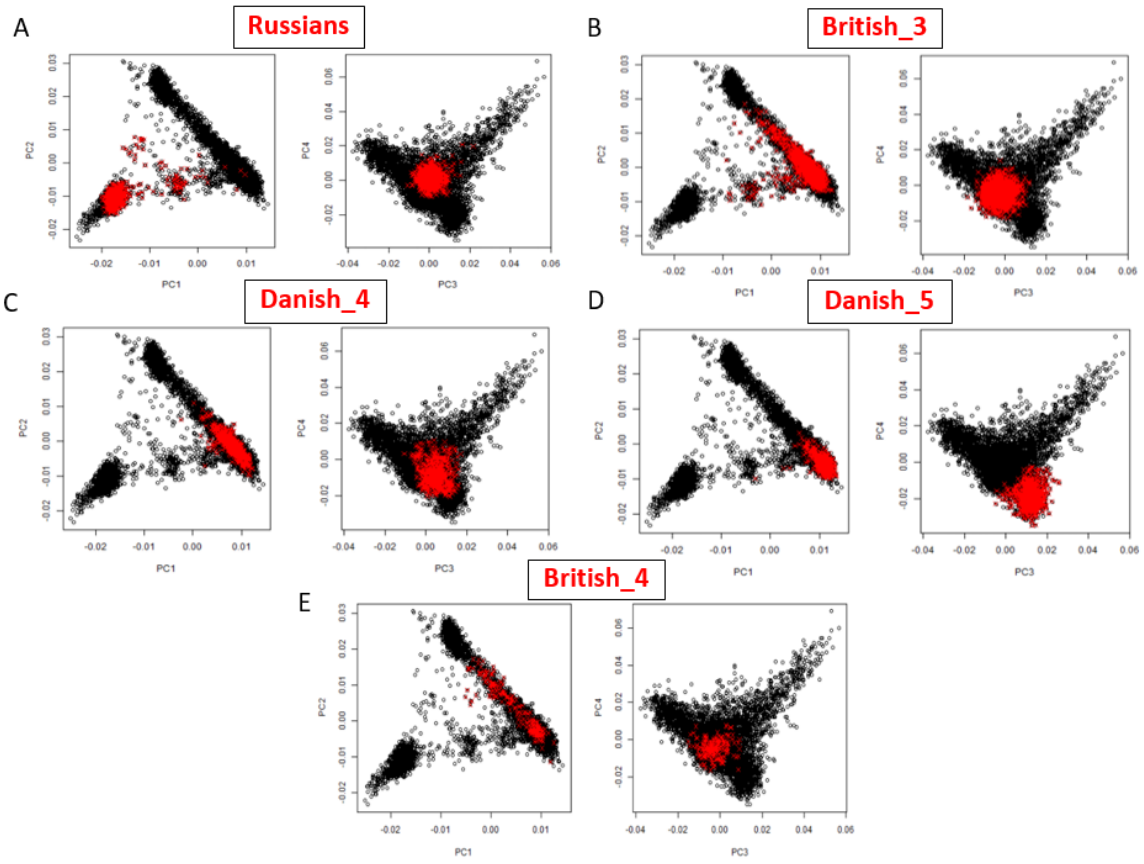

Figure 3.3: Scatter plots of principal components PC1-PC2 and PC3-PC4 from genome-wide genotype data of all subjects in the study of EL. Subjects are colored red if they belong to (A): cluster with 757 subjects enriched of Russians; (B): cluster with 1416 subjects enriched of British subjects; (C) cluster with 559 subjects enriched of Danish subjects; (D) cluster with 635 subjects enriched of Danish subjects; (E) cluster with 207 subjects enriched of British subjects.

ing. In all 5 clusters, the effect of *APOE* e4 is deleterious on longevity with worst effect in subjects of British ancestry ($OR = 0.3$, 95% CI: 0.21, 0.44) and slightly less severe effect in subjects with North Eastern Europeans ethnicities. None of the pairwise differences of genetic effects between the 5 clusters was statistically significant. However, when the effect of the cluster enriched of British ancestry was compared to the effect of the other 4 clusters combined, the difference was borderline significant ($p = 0.06$). In the other clusters, the genetic association did not reach statistical significance and it is noticeable that in the cluster enriched of Southern Italians with $N = 1309$, the effect of e4 on EL was substantially smaller than in other ethnic groups ($OR = 0.82$, $p = 0.31$) (Table 3.3). We present next the results in two specific European ethnic groups.

3.3.1 Effect of *APOE* in Italians

The cluster enriched of Southern Italians with $N = 1309$ includes 77% of subjects of South Italian ancestry (Tables 3.2, 3.3 and Figure 3.2-A) with 805 subjects from the SICS (Table 2.1) who live in South Italy, and 504 subjects who live in U.S.A. since they were enrolled from different studies (Figure 3.4). In this cluster, neither the effect of *APOE* e2 nor e4 were statistically significant when the data were analyzed without adjustment to the country of residence (carriers of e2: OR for EL = 1.27, 95% CI: 0.92, 1.76; carriers of e4: OR for EL = 0.82, 95% CI: 0.56, 1.20). To investigate the interaction between *APOE* alleles and country of residence, we analyzed the data in this cluster using a logistic regression model that included the *APOE* effect, sex, country of residence indicator variable coded as 0 for subjects living in U.S.A., and 1 for subjects living in Italy, and the interaction term between the indicator variable and the genetic effect. We did not adjust for PCs because

Table 3.4: Gene-environment model parameters (logistic regression) testing association between *APOE* e4 and EL in cluster 1309 enriched of subjects of South Italian descent.

	Estimate	**SE**	**p**
(Intercept)	$\beta_0 = -0.68$	0.28	0.02
APOE	$\beta_1 = -1.25$	0.48	0.01
Sex	$\beta_2 = -0.30$	0.16	0.06
ENV	$\beta_3 = -0.40$	0.17	0.02
APOE x ENV	$\beta_4 = 1.44$	0.52	0.01

β_0, ..., β_4 are model parameters; variable *ENV* is coded 0 if subjects live in the U.S.A., and 1 if they live in Italy. The model is adjusted by *Sex* (coded as 1 for males and 0 for females).

the dataset is genetically homogeneous and we conducted separate analyses for the effect of *APOE* e2 and e4. We did not detect a statistical significant interaction between residence and the effect of *APOE* e2. However, the model with the *APOE* e4 allele had a significant interaction between residence and the e4 effect (Table 3.4). EL was 71% less likely in Italians with one copy of e4 vs e3e3 who live in U.S.A. ($OR = 0.29$, 95% CI: 0.11, 0.73). However, there was no significant association between e4 and EL in Italians who live in Italy ($OR = 1.21$, 95% CI: 0.79, 1.86), although a sample size of 805 provides more than 75% power to detect an $OR > 1.2$. In order to remove confounding due to population structure, we also re-analyzed the data after removing 157 subjects who mostly live in South Italy, but whose ethnicity is more consistent with Northern Italians based on the PC1-PC2 plot (Figure 3.4). After removing these subjects, the statistical estimates were similar with the estimates from the analyses done on the whole cluster 1309. Specifically, EL was 81% less likely in Italians with one copy of e4 vs e3e3 who live in U.S.A. ($OR = 0.19$, 95% CI: 0.04, 0.80), and there was no significant association between e4 and EL in Italians who live in Italy ($OR = 1.21$, 95% CI: 0.79, 1.85)

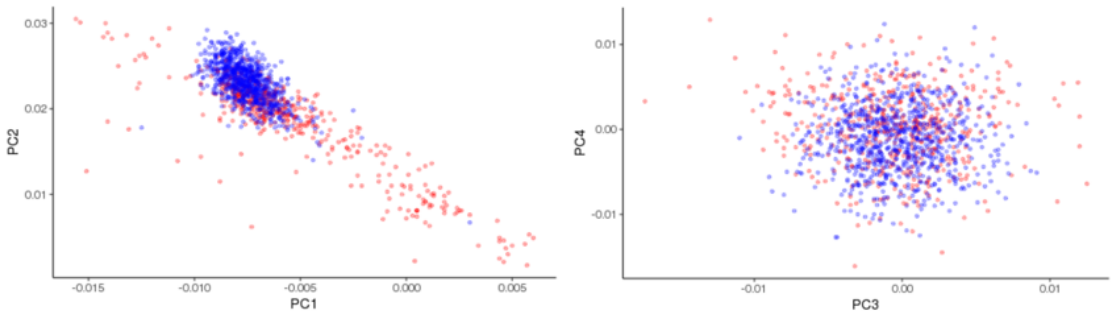

Figure 3.4: Scatter plots of principal components PC1-PC2 and PC3-PC4 from genome-wide genotype data of all subjects in the EL study. Subjects are colored in red if they are in the cluster of 1309 subjects of Southern Italian descent and live in South Italy. Subjects in blue denote individuals in this cluster who live in the U.S.A. On the scatter plot in the left panel, subjects with $PC1 > -0.005$ are more consistent with Northern Italian ethnicity, rather than Southern Italian ethnicity.

3.3.2 Effect of *APOE* in Danish

There were five distinct clusters enriched of subjects of Danish ancestry found to have ethnic-specific effects of *APOE* alleles on EL (Tables 3.2 and 3.3), i.e. clusters with 437 (Danish_1), 198 (Danish_2) and 766 (Danish_3) subjects for *APOE* e2, and clusters with 559 (Danish_4) and 635 (Danish_5) subjects for *APOE* e4. The LLFS enrolled participants in both the U.S.A. and Denmark, and therefore we could examine how country of residence modifies the effects of *APOE* on EL. Table 3.5 summarizes the distribution of Danish subjects in the 5 clusters by residence (Denmark or U.S.A.). Interestingly, while e2 was significantly positively associated with EL in the Danish_1 cluster that includes 72% of Danish living in Denmark ($OR = 2.24$, 95% CI: 1.35, 3.73), this association in the Danish_3 cluster that includes only 7% of Danish living in Denmark was not significant ($OR = 1.13$, 95% CI: 0.75, 1.69), and the ORs were significantly different ($p = 0.04$). The OR for EL in carriers of e4 in the Danish_4 cluster that includes 90% of Danish living in the U.S.A. ($OR = 0.44$,

Table 3.5: Distribution of countries where subjects live for clusters enriched of Danish ethnicity.

Cluster, No. subjects	Live in U.S.A.	Live in Denmark
Danish_1, 437	28%	72%
Danish_2, 198	40%	60%
Danish_3, 766	93%	7%
Danish_4, 559	90%	10%
Danish_5, 635	29%	71%

Detailed information for each cluster can be found in Tables 3.2 and 3.3. We refer to the clusters in which large majority (> 65%) of subjects live in U.S.A. or Denmark as clusters with Danish living in the U.S.A. (clusters Danish_3 and Danish_4) versus Danish living in Denmark (clusters Danish_1 and Danish_5) respectively.

95% CI: 0.26, 0.72) was slightly smaller than the OR for EL in carriers of e4 in the Danish_5 cluster that includes only 29% of Danish living in the U.S.A. ($OR = 0.47$, 95% CI: 0.28, 0.78), and the ORs were not significantly different ($p = 0.86$).

3.4 DISCUSSION

APOE e2 and e4 alleles are known to have an effect on EL (Schachter et al., 1994; Sebastiani et al., 2019; Schupf et al., 2013) but the analyses in this chapter suggest that the magnitude of these associations is ethnic-specific among Europeans. I used a novel algorithm to search for clusters of individuals characterized by specific genetic ancestry and varying genetic effects. Our analysis discovered one group of North-Eastern European ancestry that demonstrated a strong protective effect of *APOE* e2 on EL, and 2 groups of North European ancestry with different, deleterious effects of *APOE* e4 on EL. While with larger sample sizes the genetic association between *APOE* e2 and EL could become statistically significant in more European ethnicities, our analysis suggests that the protective effect of *APOE* e2 on EL in most European ethnicities is smaller than the effect in Ashkenazi Jewish

/ Northern European and Danish subjects living in Denmark.

Conomos et al. in (Conomos et al., 2015) have shown that the PCA of genetic data might capture family relatedness instead of population structure when applied to the datasets with relatedness. Even though our combined dataset contains 14% related individuals (Table 2.1), the evaluation of PopCluster on this very dataset had demonstrated that the algorithm worked well (Gurinovich et al., 2019).

I also provided evidence that the genetic effect of *APOE* alleles changes based on country of residence in addition to genetic ancestry, suggesting the presence of environmental risk factors that modify the genetic effects of *APOE* after controlling for genetic ancestry. For example, our analysis showed that the deleterious effect of *APOE* e4 in subjects with Southern Italian ancestry differs between those living in the South of Italy or the U.S.A. These results suggest that factors related to living in the South of Italy may mitigate the deleterious effect of *APOE* e4. The results are consistent with previous findings that the Mediterranean diet reduces the risk of Alzheimers disease (Sindi et al., 2015), and *APOE* e4 carriers versus non-carriers might have an exaggerated or different response to nutrition and other factors in relation to Alzheimers and cognitive function (Kivipelto et al., 2008; Hanson et al., 2015; Bos et al., 2019). Similarly, my analyses showed that the protective effect of the e2 allele in subjects with Danish ancestry is stronger in those individuals who live in Denmark and becomes much weaker in individuals of Danish ancestry who live in the U.S.A. The overall diet composition (energy/protein/fat/carbohydrate amounts) in Denmark and the U.S.A. is comparable; however, the distribution of consumption of saturated and unsaturated fatty acids varies between Denmark and the U.S.A. diets (Auestad et al., 2015; Harika et al., 2013). Another difference between the two countries' diets is Denmark's higher consumption of dairy prod-

ucts and fish compared to the U.S.A. (Auestad et al., 2015). These differences are suggestive of complex gene-environment interaction of *APOE* and nutrition on EL that could lead to the development of natural interventions for healthy aging.

The APOE protein is essential for healthy cholesterol metabolism and central nervous system cholesterol transport. Total APOE levels in plasma in very old individuals were found to be associated with lower total cholesterol and LDL cholesterol levels, which in turn were associated with the *APOE* e2 allele (Muenchhoff et al., 2017). The *APOE* e4 allele has been associated with abnormal lipid metabolism in cerebrospinal fluid, and reduced capacity to deliver neuronal cholesterol (Liu et al., 2013). Detrimental effects of *APOE* e4 may be alleviated through diet interventions (Ordovas, 2002), specifically Mediterranean diet (increased omega-3 fatty acids) (Bos et al., 2019; Grimm et al., 2017). Additionally, APOE e4 carriers may be more sensitive to cholesterol and saturated fatty acids (Carvalho-Wells et al., 2012). *APOE* e2 carriers with metabolic syndrome might benefit from diet interventions as well (Fallaize et al., 2017). Overall, these results are consistent with the hypothesis of an interaction between *APOE* and nutrition that differs by European ethnicity.

CHAPTER 4

GWAS of rare variants and extreme longevity

4.1 INTRODUCTION

The strong heritability of EL supports the hypothesis that this is a genetically-regulated trait. Several GWASes have identified common genetic variants that are associated with EL (Sebastiani et al., 2017b; Broer & van Duijn, 2015; Stallard et al., 2018). However, there are only a few significant hits that have been replicated in multiple studies, e.g. SNPs of *APOE* gene (Sebastiani et al., 2017b; Pilling et al., 2017). We and others hypothesize that due to the rarity of the EL phenotype, less common or rare variants could be important in deciphering the heritability of EL (Broer & van Duijn, 2015). In this chapter I will describe a new study that I conducted to discover uncommon SNPs that are associated with EL.

I conduct a GWAS of EL in a case-control dataset of 4216 individuals, including 1317 individuals who survived past the 1% age of survival of birth and sex specific cohorts and 2899 controls defined as people who did not, with median age at death or last contact of 104 years. The dataset consists of subjects from the NECS and Illumina controls. All of the subjects genotype data were imputed with the Haplotype Reference Consortium (HRC) panel of 65,000 haplotypes using Michigan Imputation server which resulted in about 9 million high-imputation quality SNPs including about 4.5 million rare and uncommon SNPs (MAF < 0.05). The associations are tested using a mixed effect logistic regression model with genotype-based kinship covariance of the random effects to adjust for cryptic relations using the package GENESIS. The analysis discovers 61 genome-wide significant SNPs (p < 5E-08) including fifteen new loci in chromosomes 4, 6, 7, 8, 9, 10, 14 and 15 in

addition to the *APOE* locus. Additionally, I use serum protein data available for a subset of subjects and find significant protein quantitative trait loci (pQTLs) which suggest new biological mechanism involved in extreme human longevity.

4.2 MATERIALS AND METHODS

4.2.1 Study populations

4.2.1.1 New England Centenarian Study

The New England Centenarian Study (NECS) is a study of centenarians, their long-lived siblings, offspring, and additional unrelated controls selected because their parents died before reaching the median age survival of their birth year cohort (Sebastiani & Perls, 2012). The study began by recruiting centenarians in the Boston metropolitan area in 1994 and expanded in the late 1990s to include North America and English speaking countries. The age of participants is carefully validated (Young et al., 2011), and participants are followed-up annually. Genome-wide genotype data of 2105 samples were previously generated using Illumina SNP arrays (Sebastiani et al., 2012). Recently additional 370 DNA samples with 284 centenarians were genotyped using Illumina Global Screening Array. The combined 2475 NECS subjects were imputed to the HRC panel (version r1.1 2016) of 64,940 haplotypes with 39,635,008 sites using the Michigan Imputation Server (Das et al., 2016). Eagle2 (Loh et al., 2016) and European population were selected for phasing and quality control respectively. All subjects provided informed consent approved by the Boston University Medical Campus IRB.

4.2.1.2 Illumina controls

To increase statistical power, we added controls selected from the Illumina repository. This repository includes approximately 6000 samples of various races and ethnicities used as controls of a variety of GWASs. The data can be accessed using the protocol available from here: http://www.illumina.com/documents/icontroldb/document_purpose.pdf. We used this set of controls as referent population in the study of longevity since we expect that only a small portion of them would become centenarians. By pooling controls from different studies we also expect no bias (e.g. controls with no cardiovascular disease, or controls with no cancer). Genome-wide genotype data were generated with a variety of Illumina SNP arrays and data were cleaned carefully in (Sebastiani et al., 2012). We selected 3613 subjects to genetically match the European ethnicity of subjects from the NECS using genome-wide PCs. Genotype data were imputed to the HRC panel using the Michigan Imputation Server as in the NECS.

4.2.2 GWAS dataset

4.2.2.1 Selection of subjects

NECS and Illumina imputed datasets were aggregated for pair-wise genome-wide identity-by-descent (IBD) estimation and PCA. For these analyses, we first excluded the SNPs in high-linkage regions, such as Major Histocompatibility Complex region on chromosome 6 and region of inversion polymorphism on chromosome 8. Additionally, ambiguous SNPs (AT, TA, GC, CG) and variants that are larger than a single variation were removed. Next, we pruned the genotype data to only keep independent SNPs. All these and following analyses (unless mentioned otherwise) were done using PLINK (Purcell et al., 2007; Chang et al., 2015)

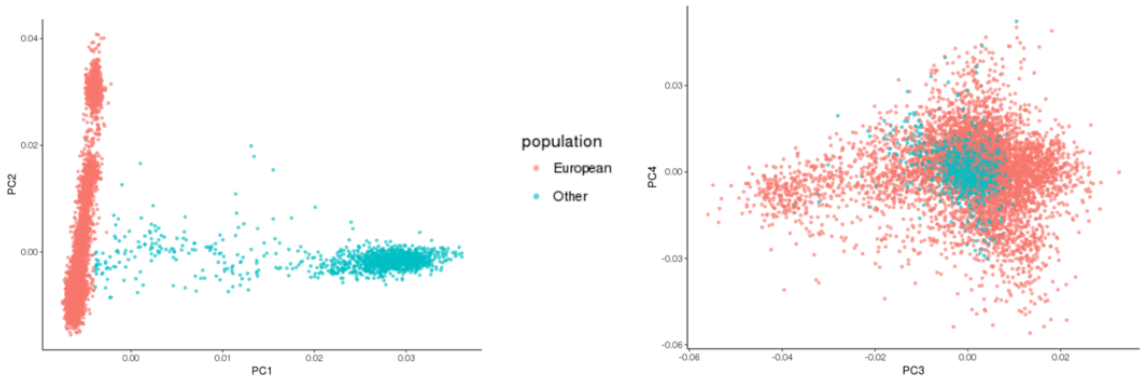

Figure 4.1: Scatter plots of principal components PC1-PC2 and PC3-PC4 from genome-wide genotype data of 6,088 genotyped and imputed subjects from the NECS and Illumina control repository. Subjects are colored in red if we identified them as Europeans through removal of a tail on PC1-PC2 plot and re-calculation of PCs.

and scripts implemented in the R programming language (R Core Team, 2018). IBD analysis was conducted to detect samples contamination, swaps, duplications, and to validate known family relation. In addition, IBD analysis was used in the PCA estimation. PCA was performed using the EIGENSOFT software (Price et al., 2006). Genome-wide principal components calculated using genotype data of all 6,088 subjects are presented in Figure 4.1. The ethnicity labels come from the experimental cut-off and re-calculation of principal components to distinguish European subjects from others. Since the majority of subjects in our dataset are of European descent, we continued the analyses using only European subjects. Re-calculated PCs for the European subjects are presented in Figure 4.2. The final dataset consisted of 4,216 subjects of European descent with 1,317 cases with median age = 104 and standard deviation = 3.7. EL was defined as survival past the 1 percentile age based on a birth cohort: 96 for 1900 birth cohort, 97 - for 1910, 98 - for 1920 for males; 100 - for females for all birth cohorts.

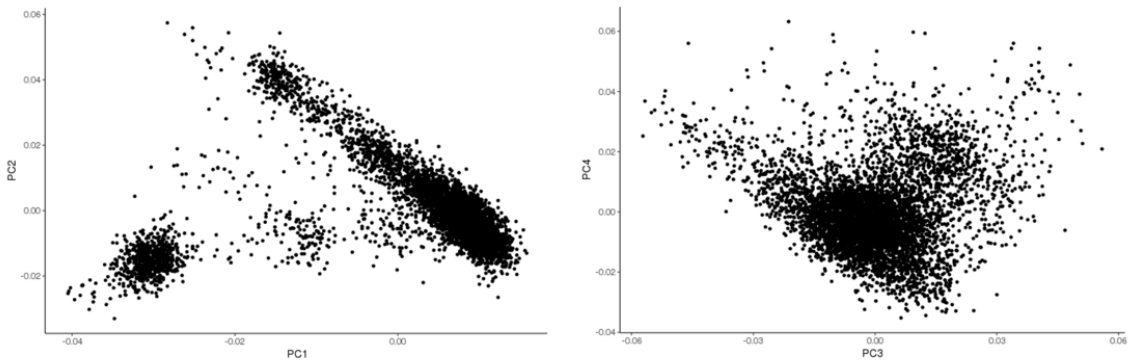

Figure 4.2: Scatter plots of re-calculated principal components PC1-PC2 and PC3-PC4 from genome-wide genotype data of European subjects in NECS and Illumina control repository (depicted as red dots in Figure 4.1).

4.2.2.2 Selection of SNPs

To select a set of SNPs to analyze, we first removed duplicate, monomorphic and ambiguous SNPs (AT, TA, GC, CG). Next, we identified SNPs that are of high-imputation quality ($Rsq > 0.7$) and are present in both NECS and Illumina genotype datasets, which resulted in 9,039,731 SNPs. Out of these SNPs, 4,593,958 have MAF < 0.05, and 2,915,050 have MAF < 0.01.

4.2.3 Protein data

A custom-designed aptamer profiling platform was used at SomaLogic Inc. (Boulder, US) to measure protein levels, as previously described (Davies et al., 2012; Emilsson et al., 2018; Hathout et al., 2015). Serum samples were selected from 227 participants (79 centenarians, 83 centenarians' offspring, and 65 controls) who were alive at least one year after the blood draw and were free of major aging-related diseases at least one year from the time of the blood draw. The 227 serum samples from the NECS biorepository were assayed with 5,034 SOMAmers. The

Table 4.1: Subjects with SOMAscan protein data.

	Centenarians	Offspring	Controls
Numbers	77	82	62
Mean age at serum yrs (sd)	105.7 (3.7)	71.0 (9.1)	70.6 (5.2)
Mean age at last contact	107.5	80.2	78.7
% alive (as of Dec. 2017)	31%	84%	84%
% Female	66%	66%	55%

Mean age at serum yrs (sd): mean age at serum draw in years with standard deviation; mean age at last contact: mean age at death for deceased subjects and age at last follow up for those who are still alive (based on 2017 follow up).

samples were randomized into analytic batches of 84 samples or less and the plates were assayed as a set, to avoid biases from technical procedures and sample processing. The SOMAscan results passed a quality control assessment for median intra- and inter-assay variability similar to variability previously reported in the SOMAscan assays (Candia et al., 2017). Genome-wide genotype data are available for the 221 subjects (77 centenarians, 82 offspring, and 62 controls) with the protein data (Table 4.1).

4.2.4 Statistical analysis

4.2.4.1 GENESIS

First, we used the GENESIS R package version 2.12.2 (Conomos et al., 2019) to perform the GWAS of EL. GENESIS contains functions for analyzing genetic data from samples with population structure and/or relatedness. The VCF files obtained from the imputation step were processed and converted to the genomic data structure (GDS) files which are accepted by GENESIS. GENESIS was used to calculate kinship estimates using the KING algorithm (Manichaikul et al., 2010) and PCs based on a pruned independent set of SNPs. Kinship estimates were re-

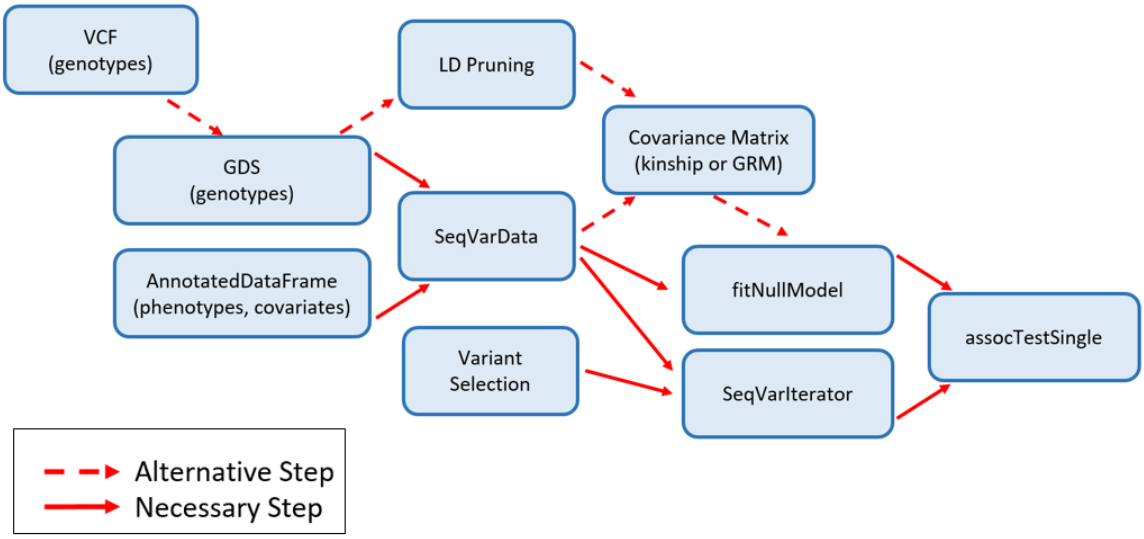

Figure 4.3: GENESIS GWAS flow chart. Modified by Zeyuan Song from (Conomos et al., 2019).

computed after adjusting for ancestry (PC1 and PC2). The general workflow of GWAS as done with GENESIS is depicted in Figure 4.3. We fit a logistic mixed effect model with genotype-based kinship to test SNP-phenotype associations:

$$\log\left(\frac{p(EL)}{1-p(EL)}\right) = \beta_0+\beta_1 SNP+\beta_2 PC_1+\beta_3 PC_2+\beta_4 PC_3+\beta_5 PC_4+\beta_6 Sex+\beta_7 Kinship, \tag{4.1}$$

where $p(EL)$ is the probability of a subject having the phenotype of EL expressed as 0 for its absence and 1 for presence; $\beta_0, ..., \beta_7$ are model parameters; and the variable SNP is the number of coded alleles in the genotypes, i.e. additive genetic model. The model is adjusted by PC_1, ..., PC_4, calculated using the independent set of SNPs, and additional covariates, such as Sex and kinship estimate ($Kinship$).

4.2.4.2 Bayesian logistic regression

Because GENESIS uses large sample approximation to compute p-values, we implemented a step-2 Bayesian procedure to validate the significant associations. Bayesian logistic regression model was adjusted for sex and four genome-wide PCs, and random effects to model the within family correlation. The model was implemented using the rjags R package (Plummer, 2018) which is an interface to the JAGS MCMC library.

4.2.4.3 pQTL analysis

The expression data of 4,785 proteins were log-transformed, and for each protein, values that differed from the protein's mean by more than three standard deviations were removed. The following linear regression model was fit for each protein-SNP combination:

$$Protein = \beta_0 + \beta_1 SNP + \beta_2 Gender + \beta_3 Age.serum, \tag{4.2}$$

where $Protein$ is the log-transformed value of the amount of protein; $\beta_0, ..., \beta_3$ are model parameters; and the variable SNP is the number of coded alleles in the genotypes, i.e. additive genetic model. The model is adjusted by $Gender$ and $Age.serum$ (age at blood draw). We selected significant protein-SNP pairs using a false discovery rate (FDR) < 0.05 as level of significance to correct for multiple testing.

Table 4.2: Genome-wide significant loci as returned by GENESIS.

SNP	Chr	Ref/Alt	CA (CAF)	Score	Score.pval	Genes
rs429358	19	T/C	C (0.11)	-11	3E-28	*APOE*
rs796804885	7	T/G	G (0.63)	8	4E-15	*EVX1*, *HIBADH*
rs10457056	6	C/T	T (0.01)	8	4E-14	*EPM2A*
rs9671417	14	A/G	G (0.01)	7	1E-11	*FSCB*
rs6977506	7	A/G	G (0.1)	7	6E-11	*DOCK4*
rs7182629	15	A/C	C (0.01)	6	8E-10	*ZNF609*
rs10973748	9	G/T	T (0.04)	6	3E-09	*SHB*, *ALDH1B1*
rs1987475	7	G/A	A (0.67)	-6	3E-09	*TRY2P*, *MTRNR2L6*

Chr: chromosome; Ref/Alt: reference and alternative alleles; CA (CAF): coded allele with coded allele frequency; Score: the chi-squared score test statistic; Score.pval: the p-value based on the score test statistic; Genes: closest gene/genes (annotation was done using ANNOVAR (Hakonarson et al., 2010)).

4.3 RESULTS

4.3.1 GENESIS

The GWAS implemented with GENESIS identified 426 SNPs with p-value < 1E-05, and 35 SNPs with genome-wide level of significance (p-value < 1E-08). The QQ-plot and Manhattan plot of results are presented in Figures 4.4 and 4.5 respectively. There are eight visible genome-wide significant loci on the Manhattan plot, and the top significant SNPs from these loci are presented in Table 4.2.

4.3.2 Bayesian logistic regression

To validate the results, we fit a Bayesian logistic regression model as described in Section 4.2.4.2 to the top 426 SNPs identified by GENESIS. The results are comparable with GENESIS, except for three SNPs with very small coded allele counts. GENESIS found those SNPs to be significantly associated with EL; however, the

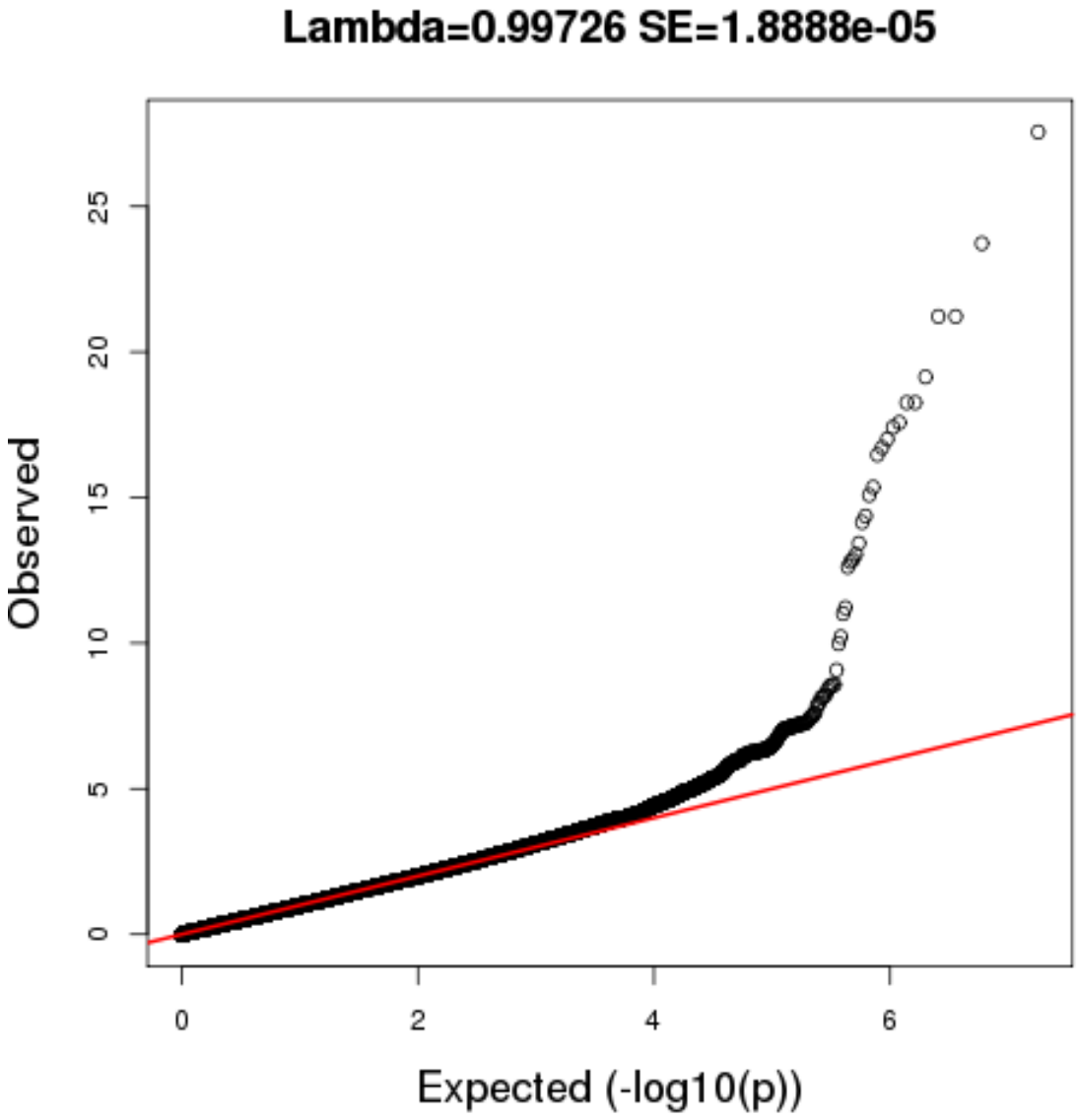

Figure 4.4: QQ plot of the GWAS of EL as implemented by GENESIS.

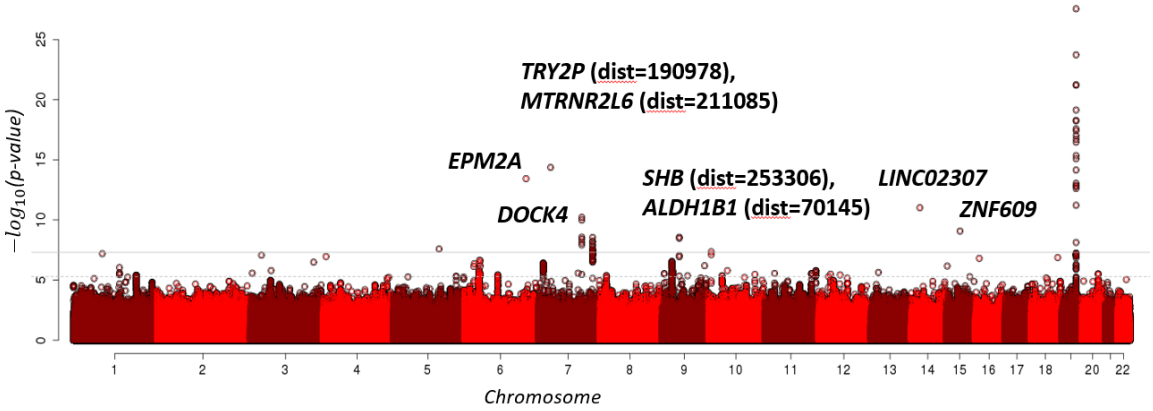

Figure 4.5: Manhattan plot of the GWAS of EL as implemented by GENESIS.

Bayesian analysis did not identify them as very significant as it could not be inferred from such small allele counts (Table 4.3). There are 61 SNPs that were found to have a genome-wide level significant association with EL by Bayesian logistic regression, with sixteen distinct loci. The top significant SNPs from those loci are presented in Table 4.4. All of the SNPs (except for rs429358 SNP of *APOE* gene) are novel genome-wide level significant targets of EL. Next, I will refer to alleles that have a protective effect on EL (OR > 1) and become more common in centenarians (CAF in cases > CAF in controls) as longevity alleles. Out of 16 loci, there are five rare (CAF in controls < 0.05) longevity alleles: two SNPs on chromosome 6: rs10457056 (*EPM2A*) and rs12524467 (*LY6G6D*), one on chromosome 9: rs10973748 (*SHB*, *ALDH1B1*), one on chromosome 14: rs9671417 (*FSCB*), and one on chromosome 15: rs7182629 (*ZNF609*). Additionally, there are four uncommon (0.05 < CAF in controls < 0.1) longevity alleles: rs13129138 (*SLC2A9*, *WDR1*) on chromosome 4, rs1042151 (*HLA-DPB1*) on chromosome 6, rs6977506 (*DOCK4*) on chromosome 7, and rs72771826 (*LINC00707*, *SFMBT2*) on chromosome 10.

Out of sixteen genome-wide level significant loci (Table 4.4), five SNPs are sig-

Table 4.3: SNPs for which GWAS results generated by GENESIS are not reliable.

SNP	Chr	Ref/Alt	CA (cases/controls)	p-value (GENESIS))	p-value (Bayesian)
rs60969364	9	T/C	C (0.002/0)	3E-06	0.06
rs186886142	9	G/A	A (0.002/0)	3E-06	0.06
rs527967377	3	G/A	A (0.003/0)	3E-07	0.003

CA (cases/controls): coded allele with its frequency in cases and controls.

Table 4.4: Genome-wide significant loci as identified by Bayesian analysis.

SNP	Ref/Alt	CA (cases/controls)	OR	P-value	Genes
rs796804885 (7)	T/G	G (0.69/0.6)	1.53	0	*EVX1*, *HIBADH*
rs429358 (19)	T/C	C (0.05/0.14)	0.32	0	*APOE*
rs10457056 (6)	C/T	T (0.02/0.004)	7.59	8E-15	*EPM2A*
rs6977506 (7)	A/G	G (0.14/0.08)	1.72	7E-12	*DOCK4*
rs10973748 (9)	G/T	T (0.06/0.03)	2.21	7E-12	*SHB*, *ALDH1B1*
rs7182629 (15)	A/C	C (0.02/0.003)	6.26	1E-10	*ZNF609*
rs2855963 (7)	G/A	A (0.63/0.7)	0.73	4E-10	*TRY2P*, *MTRNR2L6*
rs1042151 (6)	A/G	G (0.1/0.06)	1.66	1E-09	*HLA-DPB1*
rs9671417 (14)	A/G	G (0.02/0.001)	13.89	2E-09	*FSCB*
rs72771826 (10)	G/A	A (0.15/0.1)	1.55	2E-09	*LINC00707*, *SFMBT2*
rs6460902 (7)	G/A	A (0.48/0.41)	1.3	2E-09	*TMEM106B*
rs12524467 (6)	G/A	A (0.04/0.02)	2.27	2E-09	*LY6G6D*
rs13129138 (4)	C/A	A (0.13/0.09)	1.58	3E-09	*SLC2A9*, *WDR1*
rs7838131 (8)	G/A	A (0.62/0.57)	1.3	4E-09	*GATA4*
rs2008074 (9)	A/C	C (0.18/0.14)	1.44	5E-09	*LINC01503*, *LINC00963*
rs1537374 (9)	A/G	G (0.48/0.56)	0.77	7E-09	*CDKN2B-AS1*

SNP: SNP id with its chromosome in parentheses; Ref/Alt: reference and alternative alleles; OR: odds ratio for EL in carriers of the allele; P-value: p-value of the association; Genes: closest gene/genes (annotation was done using ANNOVAR (Hakonarson et al., 2010)).

nificant eQTLs (obtained from the GTEx Portal on 03/04/2019). SNP rs2855963 (chr 7) is a significant cis-eQTL for *TRBV7-4* gene in whole blood. SNP rs1042151 (chr 6) is a significant cis-eQTL for *HLA-DPA1*, *HLA-DPB2*, *HLA-DRB5*, *NOTCH4*, *RPL32P1* genes in multiple tissues. SNP rs6460902 (chr 7) is a significant cis-eQTL for *TMEM106B* gene in multiple tissues. SNP rs12524467 (chr 6) is a significant cis-eQTL for *C4A*, *LTA*, *LY6G5B*, *LY6G5C*, *SKIV2L*, *TNXA* genes, and trans-eQTL for `Y_RNA` gene (chr 20) in multiple tissues. SNP rs7838131 (chr 8) is a significant cis-eQTL for *AF131215*, *AF131216*, *BLK*, *C8orf49*, *CTSB*, *DEFB134*, *ENPP7P12*, *FAM167A*, *FAM66A*, *FAM90A25P*, *FDFT1*, *NEIL2*, *RP11-148O21*, *RP11-351I21*, *RP11-481A20*, and *TDH* genes in multiple tissues.

4.3.3 Replication

This new GWAS replicated SNPs on or near *APOE* gene which is a widely studied and replicated gene in multiple other studies of longevity and ageing, including ours (Sebastiani et al., 2017b; Pilling et al., 2017). I discuss this gene in great detail in Chapter 3 of this dissertation. Additionally, this new GWAS of EL replicated all, except for one, significant SNPs from our previous GWAS of EL (Sebastiani et al., 2017b). However, some of the genome-wide level significant SNPs from the previous GWAS did not reach genome-wide level of signficance in the new GWAS. The SNP rs7185374 on chromosome 16 that did not replicate was not in the tested dataset, and was probably dropped due to law imputation quality score or during other quality control steps.

We found SNP rs1537374 (*CDKN2B-AS1*) on chromosome 9 to have negative association with EL ($OR = 0.77$) (Table 4.4), which replicates the results of one of the previous analyses from our group (Sebastiani et al., 2012). This SNP was also

implicated in the variation of lifespan from a recent large study of 1 million parent lifespans (Timmers et al., 2019). Additionally, it has previously been found to be significantly associated with various cardiometabolic traits, such as ankle brachial index (Murabito et al., 2012), abdominal aortic aneurysm (Jones et al., 2017), and myocardial infarction (the CARDIoGRAMplusC4D Consortium, 2015).

We found SNP rs915179 (*LMNA*) on chromosome 1 (Ref/Alt = A/G with CAF (G) = 0.42) to have positive ($OR = 1.28$) significant (but not genome-wide level significant) association with EL (p-value = 1E-06), which replicates the results of two other analyses from our and other groups (Sebastiani et al., 2012; Conneely et al., 2012). This SNP was also found in the LongevityMap database (Budovsky et al., 2013) as being significantly associated with longevity.

4.3.4 pQTL analysis

I combined normalized 4785 protein and top 426 SNPs data for the 221 subjects described in Table 4.1. First, I removed SNPs with zero variance and SNPs with MAF < 0.025 for the subjects with non-missing protein data. Next, I fit linear models described by Equation 4.2 for each of the remaining protein-SNP pairs. After protein-SNP pairs were sorted by smallest to largest p-values, SNPs within 10,000 bps of the most significant SNP-protein association, were removed for each SNP-protein pair. After these QC steps, 366,260 protein-SNP pairs were left for further analysis and interpretation. This number ($N = 366,260$) was used as a number of tests for multiple comparison correction.

I find 42 protein-SNP pairs with a significant association at 5% FDR. 29 of these associations are with SNPs from the *APOE* region. Previously in a separate analysis (Sebastiani et al., SUBMITTED), we correlated *APOE* genotypes with the pro-

tein data using the same dataset as for this analysis. We discovered a signature of 16 proteins that are associated with *APOE* genotypes (Table 4.5), which is almost an identical set of proteins discovered here through individual pQTLs of SNPs on chromosome 19. This protein signature is also consistent with the results from (Emilsson et al., 2018). Here I show the correlation results between APOE genotypes and the proteins. The 16 proteins that passed the significance threshold include 9 overexpressed in carriers of the e2 allele (Figure 4.6-A), and 7 overexpressed in carriers of the e4 allele (Figure 4.6-B). Besides APOE and APOB, the other proteins have not previously been reported as directly associated with the *APOE* genotypes. The pattern of APOB expression by *APOE* genotype is consistent with results published in (Muenchhoff et al., 2017) and (Soares et al., 2012), and the rare e2e2 genotype was associated with the lowest APOB level. The level of the APOE probe included in this list is lowest in carriers of e2 and increases in carriers of e3 and e4. Interestingly, the effect of the e2 allele on most proteins was additive in the log-scale, as shown by the almost linear change of log-expression for ordered genotypes in Figures 4.6-A and -B. The genetic effect was recessive on APOB, and dominant on BIRC2. The 16 proteins have a variety of functions including regulation of cell proliferation, cell surface receptor, protein binding, and immune system. We annotated the proteins' functions using the human protein atlas (https://www.proteinatlas.org/) (Uhlén et al., 2015), Entrez (Maglott et al., 2005), Ensembl (Frankish et al., 2017), and DAVID (Sherman et al., 2008; Huang et al., 2009) (annotations retrieved on 09/2018). Below is a summary of each protein's functions:

- **BIRC2** (11q22.2 - Baculoviral IAP Repeat Containing 2): This protein is a member of a family of proteins that inhibits apoptosis by binding to tumor

necrosis factor receptor-associated factors TRAF1 and TRAF2, probably by interfering with activation of ICE-like proteases. This encoded protein inhibits apoptosis induced by serum deprivation and menadione, a potent inducer of free radicals.

- **CEP57** (11q21 - Centrosomal Protein 57): This is a cytoplasmic protein called Translokin. This protein localizes to the centrosome and has a function in microtubular stabilization. The N-terminal half of this protein is required for its centrosome localization and for its multimerization, and the C-terminal half is required for nucleating, bundling and anchoring microtubules to the centrosomes. This protein specifically interacts with fibroblast growth factor 2 (FGF2), sorting nexin 6, Ran-binding protein M and the kinesins KIF3A and KIF3B, and thus mediates the nuclear translocation and mitogenic activity of the FGF2. It also interacts with cyclin D1 and controls nucleocytoplasmic distribution of the cyclin D1 in quiescent cells. This protein is crucial for maintaining correct chromosomal number during cell division.

- **S100A13** (1q21.3 - S100 Calcium Binding Protein A13): This protein is a member of the S100 family of proteins containing 2 EF-hand calcium-binding motifs. S100 proteins are localized in the cytoplasm and/or nucleus of a wide range of cells, and involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation. This protein is widely expressed in various types of tissues with a high expression level in thyroid gland. In smooth muscle cells, this protein co-expresses with other family members in the nucleus and in stress fibers, suggesting diverse functions in signal transduction.

- **LRRN1** (3p26.2 - Leucine Rich Repeat Neuronal 1): No summary was available for this protein.

- **VPS29** (12q24.11 - Vacuolar Protein Sorting-Associated Protein 29): This protein is a component of a large multimeric complex, termed the retromer complex, which is involved in retrograde transport of proteins from endosomes to the trans-Golgi network. This VPS protein may be involved in the formation of the inner shell of the retromer coat for retrograde vesicles leaving the prevacuolar compartment.

- **PSME1** (14q12 - Proteasome Activator Subunit 1): The 26S proteasome is a multicatalytic proteinase complex with a highly ordered structure composed of 2 complexes, a 20S core and a 19S regulator. The 20S core is composed of 4 rings of 28 non-identical subunits; 2 rings are composed of 7 alpha subunits and 2 rings are composed of 7 beta subunits. The 19S regulator is composed of a base, which contains 6 ATPase subunits and 2 non-ATPase subunits, and a lid, which contains up to 10 non-ATPase subunits. Proteasomes are distributed throughout eukaryotic cells at a high concentration and cleave peptides in an ATP/ubiquitin-dependent process in a non-lysosomal pathway. An essential function of a modified proteasome, the immunoproteasome, is the processing of class I MHC peptides. The immunoproteasome contains an alternate regulator, referred to as the 11S regulator or PA28, that replaces the 19S regulator. Three subunits (alpha, beta and gamma) of the 11S regulator have been identified. This gene encodes the alpha subunit of the 11S regulator, one of the two 11S subunits that is induced by gamma-interferon. Three alpha and three beta subunits combine to form a heterohexameric ring.

- **APOE** (19q13.32 - Apolipoprotein E): This protein is a major apoprotein of the chylomicron. It binds to a specific liver and peripheral cell receptor, and is essential for the normal catabolism of triglyceride-rich lipoprotein constituents.

- **TBCA** (5q14.1 - Tubulin Folding Cofactor A): This protein is one of four proteins (cofactors A, D, E, and C) involved in the pathway leading to correctly folded beta-tubulin from folding intermediates. Cofactors A and D are believed to play a role in capturing and stabilizing beta-tubulin intermediates in a quasi-native confirmation. Cofactor E binds to the cofactor D/beta-tubulin complex; interaction with cofactor C then causes the release of beta-tubulin polypeptides that are committed to the native state.

- **UBA2** (19q13.11 - Ubiquitin Like Modifier Activating Enzyme 2): Post-translational modification of proteins by the addition of the small protein SUMO (Small Ubiquitin-like Modifier), or sumoylation, regulates protein structure and intracellular localization. SAE1 (SUMO1 Activating Enzyme Subunit 1) and UBA2 form a heterodimer that functions as a SUMO-activating enzyme for the sumoylation of proteins.

- **C5orf38** (5p15.33 - Chromosome 5 Open Reading Frame 38): No summary was available for this protein.

- **CTF1** (16p11.2 - Cardiotrophin 1): This protein is a secreted cytokine that induces cardiac myocyte hypertrophy in vitro. It has been shown to bind and activate the ILST/gp130 receoptor.

- **KMT2C** (7q36.1 - Lysine Methyltransferase 2C): This protein is a nuclear protein with an AT hook DNA-binding domain, a DHHC-type zinc finger, six PHD-type zinc fingers, a SET domain, a post-SET domain and a RING-type

zinc finger. This protein is a member of the ASC-2/NCOA6 complex (ASCOM), which possesses histone methylation activity and is involved in transcriptional coactivation.

- **KIN** (10p14 - Kin17 DNA and RNA Binding Protein): This protein is a nuclear protein that forms intranuclear foci during proliferation and is redistributed in the nucleoplasm during the cell cycle. Short-wave ultraviolet light provokes the relocalization of the protein, suggesting its participation in the cellular response to DNA damage.

- **APOB** (2p24.1 - Apolipoprotein B): This protein is the main apolipoprotein of chylomicrons and low density lipoproteins. It occurs in plasma as two main isoforms, apoB-48 and apoB-100: the former is synthesized exclusively in the gut and the latter in the liver. The intestinal and the hepatic forms of apoB are encoded by a single gene from a single, very long mRNA. The two isoforms share a common N-terminal sequence. The shorter apoB-48 protein is produced after RNA editing of the apoB-100 transcript at residue 2180 (CAA->UAA), resulting in the creation of a stop codon, and early translation termination.

- **CRYZL1** (21q22.11 - Crystallin Zeta Like 1): This protein has sequence similarity to zeta crystallin, also known as quinone oxidoreductase. This zeta crystallin-like protein also contains an NAD(P)H binding site.

- **CKAP2** (13q14.3 - Cytoskeleton Associated Protein 2): This is a cytoskeleton-associated protein that stabalizes microtubules and plays a role in the regulation of cell division. The protein is itself regulated through phosphorylation at multiple serine and threonine residues.

Table 4.5: Signature of 16 biomarkers associated with *APOE* genotypes.

Uniprot	geneID	e2e2	e2e3	e2e4	e3e4	p
Q13490	BIRC2	5.87	3.23	3.67	0.90	1.55E-61
Q86XR8	CEP57	1.62	1.23	1.22	1.01	1.96E-28
Q99584	S100A13	0.51	0.78	0.47	0.67	2.69E-23
Q6UXK5	LRRN1	0.89	0.96	1.17	1.40	1.17E-19
Q9UBQ0	VPS29	1.55	1.18	1.26	0.98	2.84E-19
Q06323	PSME1	1.51	1.27	1.22	0.99	5.23E-19
P02649	APOE	0.86	0.77	1.15	1.16	9.19E-11
O75347	TBCA	1.08	1.02	0.88	0.83	1.60E-10
Q9UBT2	UBA2	1.77	1.13	1.31	0.98	5.44E-10
Q86SI9	C5orf38	0.73	0.84	0.88	1.29	7.68E-10
Q16619	CTF1	0.94	0.95	1.20	1.11	9.63E-09
Q8NEZ4	KMT2C	1.33	1.11	1.06	0.99	2.15E-08
O60870	KIN	1.23	1.08	1.22	1.02	3.10E-07
P04114	APOB	0.50	0.86	0.97	1.07	3.36E-06
O95825	CRYZL1	0.89	0.88	0.70	1.11	7.07E-06
Q8WWK9	CKAP2	1.33	1.12	1.08	0.96	8.12E-06

Columns e2e2, e2e3, e2e4, e3e4 report fold change of protein level relative to e3e3. P is p-value from F-test after adjusting for sex, age at blood draw, and length of sample storage.

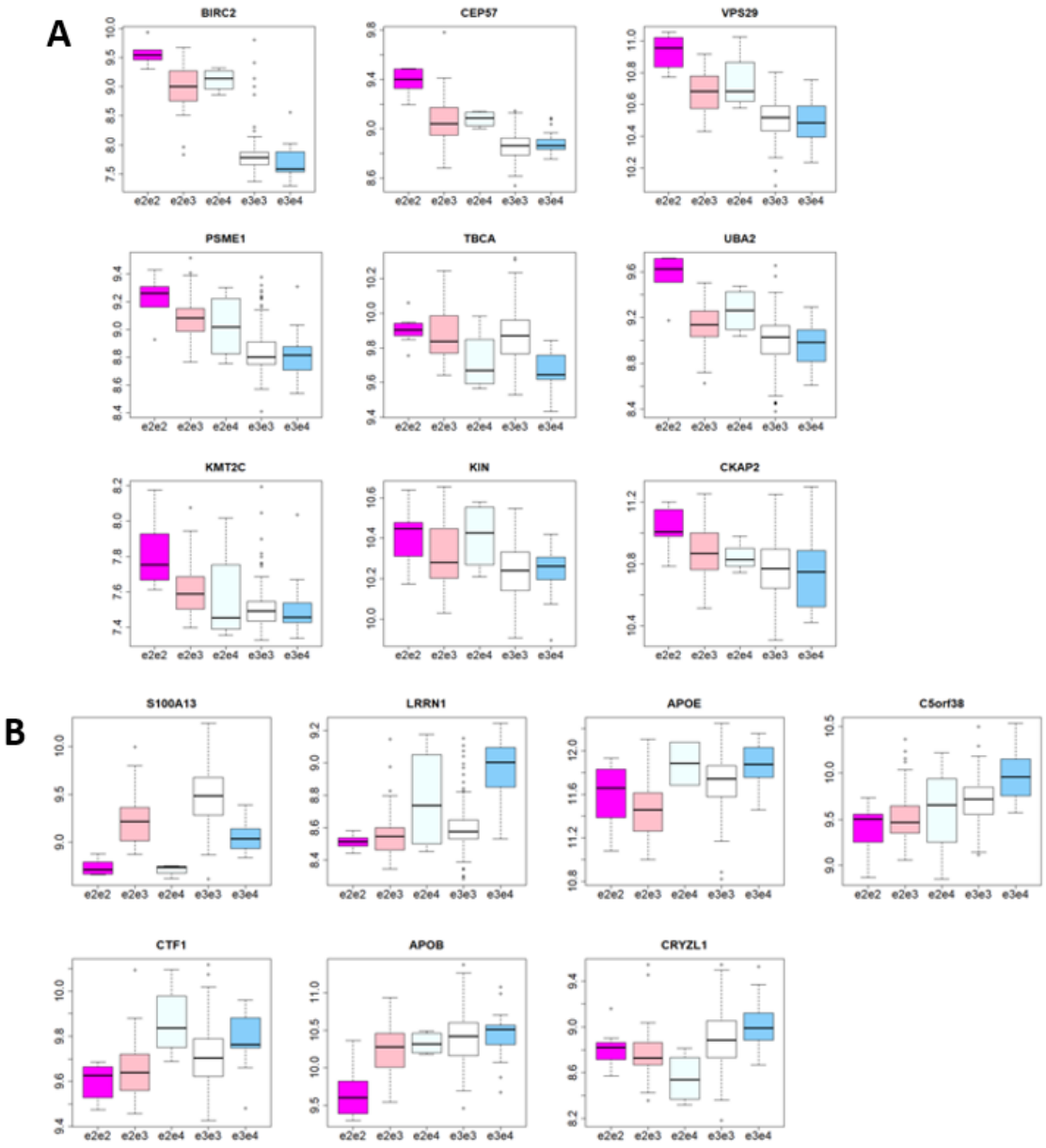

Figure 4.6: Distribution of protein intensity in log scale by *APOE* genotypes. (A) Boxplot of 9 proteins that increase expression in carriers of the e2. (B) Boxplot of 7 proteins that increase expression in carriers of the e4.

The rest of the significant pQTLs at 5% FDR are the SNPs not on chromosome 19 (Table 4.6), with 9 distinct protein-SNP pairs (Table 4.7 and Figure 4.7):

- **SNP rs11757313 is trans-pQTL for C7orf73 protein**: SNP rs11757313 is located on chromosome 6 and is a rare SNP (CAF (C) in controls = 0.005) that becomes more common in centenarians (CAF in cases = 0.018) with a strong protective effect on EL (OR = 3.78, p.value = 7.68E-08). SNP rs11757313 is a trans-pQTL of C7orf73 (7q33) which is a short transmembrane mitochondrial protein 1. Having the C allele of rs11757313 is associated with increased levels of protein C7orf73 (Figure 4.7-A). According to the Human Protein Atlas (Uhlen et al., 2017), C7orf73 is a favorable prognostic marker for ovarian cancer, meaning higher levels of C7orf73 are associated with higher survival from the ovarian cancer.

- **SNP rs114202986 is cis-pQTL for MICA protein**: SNP rs114202986 is located on chromosome 6 and is a rare SNP (CAF (C) in controls = 0.04) that becomes more common in centenarians (CAF in cases = 0.08) with a protective effect on EL (OR = 1.67, p.value = 1.19E-06). SNP rs114202986 is a cis-pQTL of MICA (6p21.33). Having the C allele of rs114202986 is associated with decreased levels of protein MICA (Figure 4.7-B). This SNP is also significant eQTL of MICA, HCG27, HLA-B, HLA-C, POU5F1, PSORS1C3, `Y_RNA` genes in multiple tissues (obtained from the GTEx Portal on 02/27/2019). Having the C allele of rs114202986 is associated with increased levels of gene MICA. Lower levels of MICA protein are associated with higher survival from the cervical cancer (Uhlen et al., 2017). MICA is the highly polymorphic major histocompatability complex class I chain-related protein A. This protein product is expressed on the cell surface, although unlike canonical class I

molecules it does not seem to associate with beta-2-microglobulin. It is a ligand for the NKG2-D type II integral membrane protein receptor. The protein functions as a stress-induced antigen that is broadly recognized by intestinal epithelial gamma delta T cells (provided by RefSeq, Jan 2014).

- **SNP rs10457056 is cis-pQTL for GLP1R protein and trans-pQTL for SEMA4B protein**: SNP rs10457056 is located on chromosome 6 and is a rare SNP (CAF (T) in controls = 0.004) that becomes more common in centenarians (CAF in cases = 0.024) with strong protective effect on EL (OR = 7.59, p.value = 8.22E-15). SNP rs10457056 is cis-pQTL of GLP1R (6p21.2). Having the T allele of rs10457056 is associated with decreased levels of the protein GLP1R (Figure 4.7-D). The GLP1R is a 7-transmembrane protein that functions as a receptor for glucagon-like peptide 1 (GLP-1) hormone, which stimulates glucose-induced insulin secretion. This receptor, which functions at the cell surface, becomes internalized in response to GLP-1 and GLP-1 analogs, and it plays an important role in the signaling cascades leading to insulin secretion. It also displays neuroprotective effects in animal models. The protein is an important drug target for the treatment of type 2 diabetes and stroke (provided by RefSeq, Apr 2016). Additionally, SNP rs10457056 is trans-pQTL of SEMA4B (15q26.1). Having the T allele of rs10457056 is associated with increased levels of protein SEMA4B (Figure 4.7-C); however, only 1 subject in the dataset has TT genotype with relatively low level of SEMA4B. Lower levels of SEMA4B protein are associated with higher survival from renal and lung cancers (Uhlen et al., 2017). Among SEMA4B's related pathways are apoptotic pathways in synovial fibroblasts and GPCR pathway. It inhibits axonal extension by providing local signals to specify territories inaccessible

for growing axons.

- **SNP rs73210911 is trans-pQTL for HCAR2 protein**: SNP rs73210911 is located on chromosome 7 and is an uncommon SNP (CAF (A) in controls = 0.09) that becomes more common in centenarians (CAF in cases = 0.13) with a protective effect on EL (OR = 1.61, p.value = 9.67E-10). SNP rs73210911 is a trans-pQTL of HCAR2 (12q24.31). Having the A allele of rs73210911 is associated with increased levels of the protein HCAR2 (Figure 4.7-E). The HCAR2 protein acts as a high affinity receptor for both nicotinic acid (also known as niacin) and (D)-beta-hydroxybutyrate and mediates increased adiponectin secretion and decreased lipolysis through G(i)-protein-mediated inhibition of adenylyl cyclase. This pharmacological effect requires nicotinic acid doses that are much higher than those provided by a normal diet. HCAR2 Mediates nicotinic acid-induced apoptosis in mature neutrophils. Receptor activation by nicotinic acid results in reduced cAMP levels which may affect activity of cAMP-dependent protein kinase A and phosphorylation of target proteins, leading to neutrophil apoptosis. HCAR2 is an FDA approved drug target for Acipimox (trade name Olbetam in Europe). Acipimox is a niacin derivative used as a lipid-lowering agent. It reduces triglyceride levels and increases HDL cholesterol.

- **SNP rs2734161 is trans-pQTL for GLTP protein**: SNP rs2734161 is located on chromosome 7 and is a common SNP (CAF (C) in controls = 0.3) that becomes more common in centenarians (CAF in cases = 0.37) with a protective effect on EL (OR = 1.37, p.value = 2.39E-07). SNP rs2734161 is a trans-pQTL of GLTP (12q24.11). Having the C allele of rs2734161 is associated with increased levels of protein GLTP (Figure 4.7-F). Higher levels of GLTP protein

are associated with higher survival from thyroid and cervical cancers, and lower survival from the liver cancer (Uhlen et al., 2017). The GLTP protein is similar to bovine and porcine proteins which accelerate transfer of certain glycosphingolipids and glyceroglycolipids between membranes. It is thought to be a cytoplasmic protein (provided by RefSeq, Jul 2008). SNP rs2734161 is also significant eQTL of *TRBV7-4* gene in whole blood with the T allele being associated with increased levels of the gene levels (obtained from the GTEx Portal on 02/27/2019).

- **SNP rs35776335 is trans-pQTL for ACVRL1 protein**: SNP rs35776335 is located on chromosome 8 and is an uncommon SNP (CAF (T) in controls = 0.18) that becomes less common in centenarians (CAF in cases = 0.15) with a negative effect on EL (OR = 0.71, p.value = 4.30E-07). The SNP rs35776335 is trans-pQTL of ACVRL1 (12q13.13). Having the T allele of rs35776335 is associated with increased levels of protein ACVRL1 (Figure 4.7-G). ACVRL1, sometimes termed ALK1, shares similar domain structures with other closely related ALK or activin receptor-like kinase proteins that form a subfamily of receptor serine/threonine kinases. Mutations in this gene are associated with hemorrhagic telangiectasia type 2, also known as Rendu-Osler-Weber syndrome 2 (provided by RefSeq, Jul 2008). SNP rs35776335 is also significant eQTL for *AF131215*, *CTSB*, *FAM66A*, *FAM85A*, *FDFT1*, *NEIL2*, *RP11-148O21*, and *RP11-351I21* in multiple tissues (obtained from the GTEx Portal on 02/27/2019).

- **SNP rs10973751 is cis-pQTL for TOR4A protein**: SNP rs10973751 is located on chromosome 9 and is a rare SNP (CAF (C) in controls = 0.01) that becomes more common in centenarians (CAF in cases = 0.03) with a protective ef-

fect on EL (OR = 2.37, p.value = 2.07E-06). SNP rs10973751 is a cis-pQTL of TOR4A (9q34.3). Having the C allele of rs10973751 is associated with increased levels of protein TOR4A (Figure 4.7-H). Higher levels of TOR4A protein are associated with lower survival from liver and colorectal cancers (Uhlen et al., 2017). TOR4A is related to the pathway of response to elevated platelet cytosolic Ca2+.

- **SNP rs72485059 is trans-pQTL for ARPP21 protein**: SNP rs72485059 is located on chromosome 10 and is an uncommon SNP (CAF (T) in controls = 0.14) that becomes more common in centenarians (CAF in cases = 0.17) with a protective effect on EL (OR = 1.42, p.value = 1.09E-07). SNP rs72485059 is a trans-pQTL of ARPP21 (3p22.3). Having the T allele of rs72485059 is associated with increased levels of protein ARPP21 (Figure 4.7-I). ARPP21 is a cAMP-regulated phosphoprotein which is enriched in the caudate nucleus and cerebellar cortex. A similar protein in mouse may be involved in regulating the effects of dopamine in the basal ganglia (provided by RefSeq, Jun 2012).

4.4 DISCUSSION

We performed a GWAS of EL using a large set of high-quality imputed variants. The analysis discovered 61 genome-wide significant SNPs (p< 5E-08) including fifteen new loci in chromosomes 4, 6, 7, 8, 9, 10, 14 and 15 in addition to the *APOE* locus. From these novel significant results, there were nine rare (CAF in controls < 0.05) and uncommon (0.05 < CAF in controls < 0.1) variants with protective effect on EL (OR > 1): rs13129138 (chr 4), rs10457056 (chr 6), rs12524467 (chr 6), rs1042151

Table 4.6: Description of SNPs not on chromosome 19 which were identified as significant pQTLs.

SNP	CA (cases/controls)	OR	P-value	Genes
rs11757313 (6)	C (0.018/0.005)	3.78	7.68E-08	*NRSN1*
rs114202986 (6)	C (0.08/0.04)	1.67	1.19E-06	*MICA*
rs10457056 (6)	T (0.024/0.004)	7.59	8.22E-15	*EPM2A*
rs73210911 (7)	A (0.13/0.09)	1.61	9.67E-10	*DOCK4*
rs2734161 (7)	C (0.37/0.3)	1.37	2.39E-07	*TRY2P*, *MTRNR2L6*
rs35776335 (8)	T (0.15/0.18)	0.71	4.30E-07	*CTSB*, *DEFB136*
rs10973751 (9)	C (0.03/0.01)	2.37	.07E-06	*SHB*, *ALDH1B1*
rs72485059 (10)	T (0.17/0.14)	1.42	1.09E-07	N/A

SNP: SNP id with its chromosome in parentheses; OR: odds ratio for EL in carriers of the allele (from the Bayesian logistic regression); P-value: p-value of the association between SNP and EL; Genes: closest gene/genes (annotation was done using ANNOVAR (Hakonarson et al., 2010)).

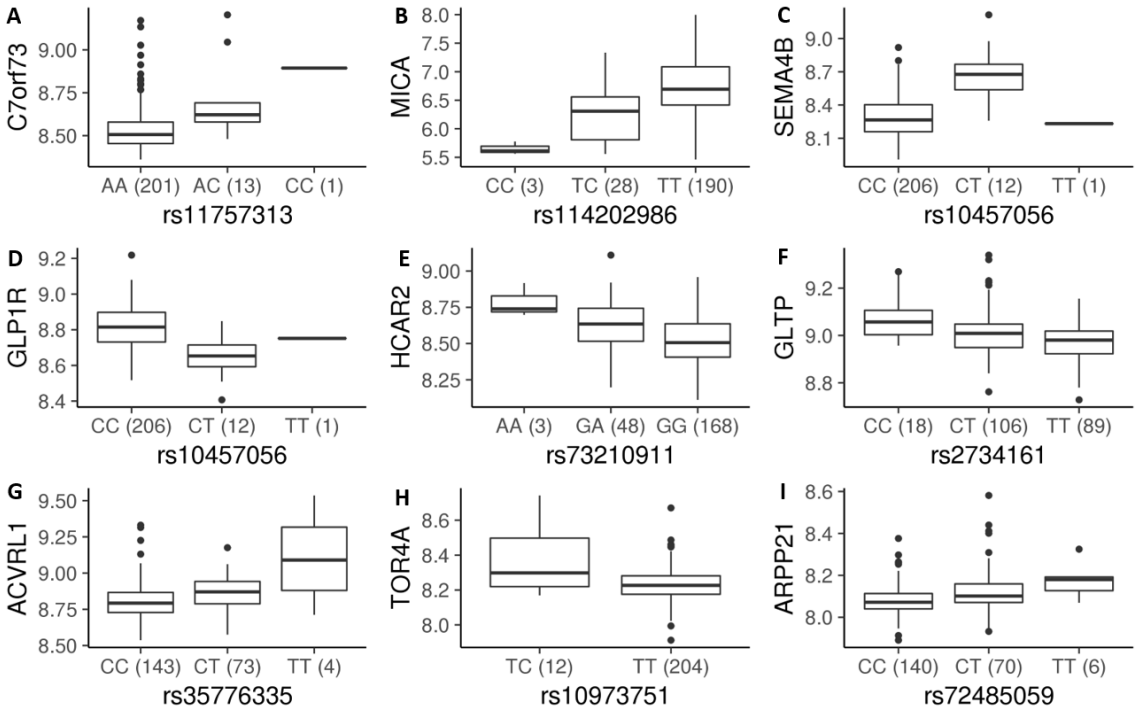

Figure 4.7: Boxplots of distribution of protein intensity in log scale by significant pQTLs that are not on chromosome 19.

Table 4.7: Protein-SNP pQTLs for SNPs not on chromosome 19.

pQTL	Protein (Chr)	SNP (Chr)	CA (CAF)	Estimate (SE)	P
trans-	C7orf73 (7)	rs11757313 (6)	C (0.03)	0.17 (0.03)	3E-06
cis-	MICA (6)	rs114202986 (6)	C (0.08)	-0.43 (0.09)	8E-07
cis-	GLP1R (6)	rs10457056 (6)	T (0.03)	-0.11 (0.02)	3E-06
trans-	SEMA4B (15)	rs10457056 (6)	T (0.03)	0.23 (0.05)	7E-07
trans-	HCAR2 (12)	rs73210911 (7)	A (0.12)	0.12 (0.02)	29E-07
trans-	GLTP (12)	rs2734161 (7)	C (0.33)	-0.05 (0.01)	1E-06
trans-	ACVRL1 (12)	rs35776335 (8)	T (0.18)	0.08 (0.02)	5E-06
cis-	TOR4A (9)	rs10973751 (9)	C (0.03)	0.14 (0.03)	5E-06
trans-	ARPP21 (3)	rs72485059 (10)	T (0.19)	0.05 (0.01)	5E-06

pQTL: whether the SNP is associated with levels of a protein on the same chromosome (-cis), or different one (-trans); Protein (Chr): protein with its chromosome; SNP (Chr) : SNP with its chromosome; CA (CAF): coded allele with coded allele frequency; Estimate (SE): estimate represents log of fold change of protein level depending on SNPs genotype (standard error of the estimate); P: p-value of the association

(chr 6), rs6977506 (chr 7), rs10973748 (chr 9), rs72771826 (chr 10), rs9671417 (chr 14), rs7182629 (chr 15).

Additionally, we discovered a signature in serum of 16 proteins that are associated with different *APOE* genotypes. We also identified nine new significant pQTLs in serum for SNPs not near the APOE locus (including four rare and three uncommon pQTLs) that suggest new biological mechanisms involved in extreme human longevity.

Our GWAS replicated most of the hits from our previous GWAS of EL (Sebastiani et al., 2017b). In addition, SNP rs1537374 (chr 9) from our results has also been recently identified to have an association with a lifespan (Timmers et al., 2019) and various cardiometabolic traits (Murabito et al., 2012; Jones et al., 2017; the CARDIoGRAMplusC4D Consortium, 2015). Furthermore, SNP rs915179 (chr 1) from our results has also been previously identified as associated with longevity (Conneely et al., 2012; Budovsky et al., 2013).

CHAPTER 5

Conclusions

Over the last decade, several studies have provided evidence that many centenarians delay or escape aging-related diseases, such as cardiovascular and Alzheimers diseases, and that more than 90% of people living to 100 are functionally independent at the mean age of 93 years and thus markedly delay disability (Hitt et al., 1999; Terry et al., 2008). Many who live to 105 years and older, thus truly approaching the limit of human lifespan, also compress the age of onset of these diseases and disability towards the end of their very long lives (Andersen et al., 2012). We and other have hypothesized that centenarians posses protective genetic and molecular profiles that can be leveraged to promote healthy aging and develop novel therapeutics for aging-related disease. Over the last decade, there have been multiple GWASs of EL. However, they have produced limited findings, despite the strong heritability of this trait. There are two possible reasons for this limitation. First, genetic variants might have a varying effect on EL in different populations, and GWAS applied to a dataset as a whole may not pinpoint such differences. Second, most of the published genetic association studies of extreme human longevity have searched for common variants, but the findings published so far point to rare variants, or rare recessive genotypes that are associated with living to extreme old age, such as the *APOE* e2 allele. In this dissertation, I tried to address these issues through three projects. Below I will summarize each of the projects and discuss their potential expansion and future directions.

First, I developed PopCluster algorithm that identifies if a genetic variant of interest has statistically different effect on a phenotype in different ethnic clusters. PopCluster could be extended in a few different ways. For example, I applied hi-

erarchical clustering to identify different populations because of its deterministic nature; however, other clustering approaches could be used in a similar manner on a set of principal components inferred from the genome-wide genetics data (Solovieff et al., 2010). The current implementation of PopCluster is not designed to analyze genome-wide genotype data and can be used to re-analyze the associations between the SNPs that reach a certain level of significance in a standard genome-wide association study. For future work, the implementation of PopCluster could be optimized to handle larger genome-wide genotype datasets.

For the second project, I investigated ethnic-specific effects of *APOE* alleles on EL, along with the country of residence of subjects. The analyses suggest possible interaction of the *APOE* gene with the environment. Future investigations into diet and *APOE* genotype interaction might point at viable nutrition interventions to reduce the deleterious effect of *APOE* e4 allele. Additionally, accounting for ethnic-specific differences in the drug development process would contribute to higher drug efficacy for more populations (Schork, 2015).

For the last project, I conducted GWAS of rare and uncommon variants of EL. I found five rare and four uncommon SNPs that are significantly associated with EL. Additionally, we discovered sixteen proteins associated with *APOE* genotypes and nine pQTLs of non-*APOE* SNPs. Preliminary pQTL results suggest they can help to understand the biological mechanism that links genotype to phenotype and to identify targets that can be manipulated. In the future, we plan to verify our top hits using whole-genome sequencing data from our lab and collaborators'. Specifically, we plan to impute genotype data from other studies (LLFS, SICS and LGP), and to conduct a mega-analysis focused on rare variants. There have been limited effort in generating whole-genome sequences of centenarian genomes, and

an important step to discover genetic variants of extreme human longevity will be to conduct whole-genome sequence studies of centenarians and family members.

BIBLIOGRAPHY

Andersen, S. L., Sebastiani, P., Dworkis, D. A., Feldman, L., & Perls, T. T. (2012). Health span approximates life span among many supercentenarians: Compression of morbidity at the approximate limit of life span. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *67A*(4), 395–405.

Ash, A. S., Christensen, K., Province, M., Sebastiani, P., Perls, T. T., Rossi, W., & Hadley, E. C. (2009). A family longevity selection score: Ranking sibships by their longevity, size, and availability for study. *American Journal of Epidemiology*, *170*(12), 1555–1562.

Atzmon, G., Schechter, C., Greiner, W., Davidson, D., Rennert, G., & Barzilai, N. (2004). Clinical phenotype of families with longevity. *Journal of the American Geriatrics Society*, *52*(2), 274–277.

Auestad, N., Hurley, J., Fulgoni, V., & Schweitzer, C. (2015). Contribution of food groups to energy and nutrient intakes in five developed countries. *Nutrients*, *7*(6), 4593–618.

Barzilai, N., Atzmon, G., Schechter, C., Schaefer, E., Cupples, A., Lipton, R., Cheng, S., & Shuldiner, A. (2003). Unique lipoprotein phenotype and genotype associated with exceptional longevity. *JAMA*, *290*(15), 2030–40.

Bell, F., & Miller, M. (2005). Life tables for the United States Social Security area 1900-2100. *Actuarial Study*, *16*.

Bos, M. M., Noordam, R., Blauw, G. J., Slagboom, P. E., Rensen, P. C. N., & van Heemst, D. (2019). The apoe e4 isoform: Can the risk of diseases be reduced by environmental factors? *The Journals of Gerontology: Series A*, *74*(1), 99–107.

Broer, L., & van Duijn, C. (2015). GWAS and meta-analysis in aging/longevity. In P. G. Atzmon (Ed.) *Longevity Genes*, chap. 847. New York, NY: Advances in Experimental Medicine and Biology. Springer.

Budovsky, A., Craig, T., Wang, J., Tacutu, R., Csordas, A., Lourenco, J., Fraifeld, V. E., & de Magalhaes, J. P. (2013). LongevityMap: a database of human genetic variants associated with longevity. *Trends in Genetics*, *29*(10), 559 – 560.

Campos, M., Edland, S., & Peavy, G. (2013). An exploratory study of APOE-e4 genotype and risk of Alzheimer's disease in Mexican Hispanics. *Journal of the American Geriatrics Society*, *61*(6), 1038–1040.

Candia, J., Cheung, F., Kotliarov, Y., Fantoni, G., Sellers, B., Griesman, T., Huang, J., Stuccio, S., Zingone, A., Ryan, B. M., Tsang, J. S., & Biancotto, A. (2017). Assessment of variability in the SOMAscan assay. *Scientific Reports*, *7*(14248).

Carlson, C. S., Matise, T. C., North, K. E., Haiman, C. A., Fesinmeyer, M. D., Buyske, S., Schumacher, F. R., Peters, U., Franceschini, N., Ritchie, M. D., et al. (2013). Generalization and dilution of association results from European GWAS in populations of non-European ancestry: The PAGE study. *PLoS Biology*, *11*(9).

Carvalho-Wells, A. L., Jackson, K. G., Lockyer, S., Lovegrove, J. A., & Minihane, A. M. (2012). APOE genotype influences triglyceride and C-reactive protein responses to altered dietary fat intake in UK adults. *The American Journal of Clinical Nutrition*, *96*(6), 1447–1453.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7.

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., Szpiro, A. A., Chen, W., Brehm, J. M., Celedón, J. C., Redline, S., Papanicolaou, G. J., Thornton, T. A., Laurie, C. C., Rice, K., & Lin, X. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, *98*(4), 653–666.

Cingolani, P., Platts, A., Coon, M., Nguyen, T., Wang, L., Land, S., Lu, X., & Ruden, D. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92.

Conneely, K. N., Capell, B. C., Erdos, M. R., Sebastiani, P., Solovieff, N., Swift, A. J., Baldwin, C. T., Budagov, T., Barzilai, N., Atzmon, G., Puca, A. A., Perls, T. T., Geesaman, B. J., Boehnke, M., & Collins, F. S. (2012). Human longevity and common variations in the LMNA gene: a meta-analysis. *Aging Cell*, *11*(3), 475–481.

Conomos, M. P., Gogarten, S. M., Brown, L., Chen, H., Rice, K., Sofer, T., Thornton, T., & Yu, C. (2019). *GENESIS: GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness*. R package version 2.12.2, https://github.com/UW-GAC/GENESIS.

Conomos, M. P., Miller, M. B., & Thornton, T. A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, *39*(4), 276–293.

Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L., & Tang, H. (2017). Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. *The American Journal of Human Genetics*, *101*(2), 218 – 226.

Corbo, R. M., & Scacchi, R. (1999). Apolipoprotein E (APOE) allele distribution in the world. Is APOE*4 a 'thrifty' allele? *Annals of Human Genetics*, *9*, 301–310.

Corder, E., Saunders, A., Strittmatter, W., Schmechel, D., Gaskell, P., Small, G., Roses, A., Haines, J., & Pericak-Vance, M. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, *261*(5123), 921–923.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G. R., & Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*, 1284–1287.

Davies, D. R., Gelinas, A. D., Zhang, C., Rohloff, J. C., Carter, J. D., O'Connell, D., Waugh, S. M., Wolk, S. K., Mayfield, W. S., Burgin, A. B., Edwards, T. E., Stewart, L. J., Gold, L., Janjic, N., & Jarvis, T. C. (2012). Unique motifs and hydrophobic interactions shape the binding of modified DNA ligands to protein targets. *Proceedings of the National Academy of Sciences*, *109*(49), 19971–19976.

Delaneau, O., Marchini, J., & Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, *9*, 179–181.

Eisenberg, D. T., Kuzawa, C. W., & Hayes, M. G. (2010). Worldwide allele frequencies of the human apolipoprotein E gene: Climate, local adaptations, and evolutionary history. *American Journal of Physical Anthropology*, *143*(1), 100–111.

Emilsson, V., Ilkov, M., Lamb, J. R., Finkel, N., Gudmundsson, E. F., Pitts, R., Hoover, H., Gudmundsdottir, V., Horman, S. R., Aspelund, T., Shu, L., Trifonov, V., Sigurdsson, S., Manolescu, A., Zhu, J., Olafsson, Ö., Jakobsdottir, J., Lesley, S. A., To, J., Zhang, J., Harris, T. B., Launer, L. J., Zhang, B., Eiriksdottir, G., Yang, X., Orth, A. P., Jennings, L. L., & Gudnason, V. (2018). Co-regulatory networks of human serum proteins link genetics to disease. *Science*, *361*(6404), 769–773.

Epstein, M. P., Allen, A. S., & Satten, G. A. (2007). A simple and improved correction for population stratification in case-control studies. *The American Journal of Human Genetics*, *80*(5), 921 – 930.

Evert, J., Lawler, E. V., Bogan, H., & Perls, T. T. (2003). Morbidity profiles of centenarians: survivors, delayers, and escapers. *The journals of gerontology. Series A, Biological sciences and medical sciences*, *58 3*, 232–7.

Fallaize, R., Carvalho-Wells, A. L., Tierney, A. C., Marin, C., Kiec-Wilk, B., Dembinska-Kiec, A., Drevon, C. A., DeFoort, C., Lopez-Miranda, J., Riserus, U., Saris, W. H., Blaak, E. E., Roche, H. M., & Lovegrove, J. A. (2017). APOE genotype influences insulin resistance, apolipoprotein CII and CIII according to plasma fatty acid profile in the Metabolic Syndrome. *Scientific Reports*, *7*(6274).

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Frankish, A., Vullo, A., Zadissa, A., Yates, A., Thormann, A., Parker, A., Gall, A., Moore, B., Walts, B., Aken, B. L., Cummins, C., GirÃşn, C. G., Ong, C. K., Sheppard, D., Staines, D. M., Murphy, D. N., Zerbino, D. R., Ogeh, D., Perry, E., Haskell, E., Martin, F. J., Cunningham, F., Riat, H. S., Schuilenburg, H., Sparrow, H., Lavidas, I., Loveland, J. E., To, J. K., Mudge, J., Bhai, J., Taylor, K., Billis, K., Gil, L., Haggerty, L., Gordon, L., Amode, M., Ruffier, M., Patricio, M., Laird, M. R., Muffato, M., Nuhn, M., Kostadima, M., Langridge, N., Izuogu, O. G., Achuthan, P., Hunt, S. E., Janacek, S. H., Trevanion, S. J., Hourlier, T., Juettemann, T., Maurel, T., Newman, V., Akanni, W., McLaren, W., Liu, Z., Barrell, D., & Flicek, P. (2017). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761.

Fries, J. F. (1980). Aging, natural death, and the compression of morbidity. *New England Journal of Medicine*, *303*(3), 130–135.

Giuliani, C., Sazzini, M., Pirazzini, C., Bacalini, M. G., Marasco, E., Ruscone, G. A. G., Fang, F., Sarno, S., Gentilini, D., Di Blasio, A. M., Crocco, P., Passarino, G., Mari, D., Monti, D., Nacmias, B., Sorbi, S., Salvarani, C., Catanoso, M., Pettener, D., Luiselli, D., Ukraintseva, S., Yashin, A., Franceschi, C., & Garagnani, P. (2018). Impact of demography and population dynamics on the genetic architecture of human longevity. *Aging*, *10*(8), 1947–1963.

Grimm, M. O. W., Michaelson, D. M., & Hartmann, T. (2017). Omega-3 fatty acids, lipids, and apoE lipidation in Alzheimers disease: a rationale for multi-nutrient dementia prevention. *Journal of Lipid Research*, *58*(11), 2083–2101.

Gurinovich, A., Bae, H., Farrell, J. J., Andersen, S. L., Monti, S., Puca, A., Atzmon, G., Barzilai, N., Perls, T. T., & Sebastiani, P. (2019). PopCluster: an algorithm to identify genetic variants with ethnicity-dependent effects. *Bioinformatics*, *btz017*.

Hakonarson, H., Li, M., & Wang, K. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164.

Hanson, A., Bayer, J., Baker, L., Cholerton, B., VanFossen, B., Trittschuh, E., Rissman, R., Donohue, M., Moghadam, S., Plymate, S., & Craft, S. (2015). Differential effects of meal challenges on cognition, metabolism, and biomarkers for Apolipoprotein E e4 carriers and adults with mild cognitive impairment. *Journal of Alzheimer's Disease*, *48*(1), 205–218.

Harika, R., Eilander, A., Alssema, M., Osendarp, S., & Zock, P. (2013). Intake of fatty acids in general populations worldwide does not meet dietary recommendations to prevent coronary heart disease: A systematic review of data from 40 countries. *Annals of Nutrition & Metabolism*, *63*, 229–238.

Hathout, Y., Brody, E., Clemens, P. R., Cripe, L., DeLisle, R. K., Furlong, P., Gordish-Dressman, H., Hache, L., Henricson, E., Hoffman, E. P., Kobayashi, Y. M., Lorts, A., Mah, J. K., McDonald, C., Mehler, B., Nelson, S., Nikrad, M., Singer, B., Steele, F., Sterling, D., Sweeney, H. L., Williams, S., & Gold, L. (2015). Large-scale serum protein biomarker discovery in Duchenne muscular dystrophy. *Proceedings of the National Academy of Sciences*, *112*(23), 7153–7158.

Hendrie, H. C., Murrell, J., Baiyewu, O., Lane, K. A., Purnell, C., Ogunniyi, A., Unverzagt, F. W., Hall, K., Callahan, C. M., Saykin, A. J., et al. (2014). APOE e4 and the risk for Alzheimer disease and cognitive decline in African Americans and Yoruba. *International Psychogeriatrics*, *26*(6), 977–985.

Hitt, R., Young-Xu, Y., Silver, M., & Perls, T. (1999). Centenarians: the older you get, the healthier you have been. *The Lancet*, *354*(9179), 652.

Hojsgaard, S., Halekoh, U., & Yan, J. (2006). The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, *15/2*, 1–11.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959.

Huang, D. W. a. . W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, *4*(1), 44–57.

Ismail, K., Nussbaum, L., Sebastiani, P., Andersen, S., Perls, T., Barzilai, N., & Milman, S. (2016). Compression of morbidity is observed across cohorts with exceptional longevity. *Journal of the American Geriatrics Society*, *64*(8), 1583–1591.

Jones, G. T., Tromp, G., Kuivaniemi, H., Gretarsdottir, S., Baas, A. F., Giusti, B., Strauss, E., et al. (2017). Meta-analysis of genome-wide association studies for abdominal aortic aneurysm identifies four new disease-specific risk loci. *Circulation Research*, *120*(2), 341–353.

Kelly, D. E., Hansen, M. E., & Tishkoff, S. A. (2017). Global variation in gene expression and the value of diverse sampling. *Current Opinion in Systems Biology*, *1*, 102–108.

Kimmel, G., Jordan, M. I., Halperin, E., Shamir, R., & Karp, R. M. (2007). A randomization test for controlling population stratification in whole-genome association studies. *The American Journal of Human Genetics*, *81*(5), 895 – 905.

Kivipelto, M., Rovio, S., Ngandu, T., KÃěreholt, I., Eskelinen, M., Winblad, B., Hachinski, V., Cedazo-Minguez, A., Soininen, H., Tuomilehto, J., & Nissinen, A. (2008). Apolipoprotein E e4 magnifies lifestyle risks for dementia: a population-based study. *Journal of Cellular and Molecular Medicine*, *12*(6b), 2762–2771.

Liu, C.-C., Kanekiyo, T., Xu, H., & Bu, G. (2013). Apolipoprotein E and Alzheimer disease: risk, mechanisms, and therapy. *Nature Reviews Neurology*, *9*(2), 106–118.

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & Price, A. L. (2016). Reference-based phasing using the haplotype reference consortium panel. *Nature Genetics*, *48*, 1443–1448.

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, *33*(suppl_1), D54–D58.

Malovini, A., Illario, M., Iaccarino, G., Villa, F., Ferrario, A., Roncarati, R., Anselmi, C. V., Novelli, V., Cipolletta, E., Leggiero, E., et al. (2011). Association study on long-living individuals from Southern Italy identifies rs10491334 in the CAMKIV gene that regulates survival proteins. *Rejuvenation Research*, *14*, 283–291.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873.

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J., & Kohane, I. S. (2016). Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, *375*(7), 655–665.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., & Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *American Journal of Human Genetics*, *100*(4), 635–649.

Mathew, S. S., Barwell, J., Khan, N., Lynch, E., Parker, M., & Qureshi, N. (2017). Inclusion of diverse populations in genomic research and health services: Genomix workshop report. *Journal of Community Genetics*, *8*(4), 267–273.

Morris, A. P. (2011). Transethnic meta-analysis of genomewide association studies. *Genetic Epidemiology*, *35*(8), 809–822.

Muenchhoff, J., Song, F., Poljak, A., Crawford, J. D., Mather, K. A., Kochan, N. A., Yang, Z., Trollor, J. N., Reppermund, S., Maston, K., Theobald, A., Kirchner-Adelhardt, S., Kwok, J. B., Richmond, R. L., McEvoy, M., Attia, J., Schofield, P. W., Brodaty, H., & Sachdev, P. S. (2017). Plasma apolipoproteins and physical and cognitive health in very old individuals. *Neurobiology of Aging*, *55*, 49 – 60.

Murabito, J. M., White, C. C., Kavousi, M., Sun, Y. V., Feitosa, M. F., Nambi, V., Lamina, C., Schillert, A., Coassin, S., Bis, J. C., Broer, L., Crawford, D. C., Franceschini, N., et al. (2012). Association between chromosome 9p21 variants and the ankle-brachial index identified by a meta-analysis of 21 genome-wide association studies. *Circulation: Cardiovascular Genetics*, *5*(1), 100–112.

Need, A. C., & Goldstein, D. B. (2009). Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics*, *25*(11), 489–494.

Newman, A. B., Glynn, N. W., Taylor, C. A., Sebastiani, P., Perls, T. T., Mayeux, R., Christensen, K., Zmuda, J. M., Barral, S., Lee, J. H., et al. (2011). Health and function of participants in the Long Life Family Study: A comparison with other cohorts. *Aging*, *3*(1), 63–76.

Ordovas, J. M. (2002). Gene-diet interaction and plasma lipid responses to dietary intervention. *Biochemical Society Transactions*, *30*(2), 68–73.

Petrovski, S., & Goldstein, D. B. (2016). Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biology*, *17*(1), 157.

Pilling, L. C., Kuo, C.-L., Sicinski, K., Tamosauskaite, J., Kuchel, G. A., Harries, L. W., Herd, P., Wallace, R., Ferrucci, L., & Melzer, D. (2017). Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging*, *9*(12), 2504–2520.

Plummer, M. (2018). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-8, https://CRAN.R-project.org/package=rjags.

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature Comment*, *538*(7624), 161–164.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetic*, *38*, 904–909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science*, *298*(5602), 2381–2385.

Schachter, F., Faure-Delanef, L., Guenot, F., Rouger, H., Froguel, P., Lesueur-Ginot, L., & Cohen, D. (1994). Genetic associations with human longevity at the APOE and ACE loci. *Nature Genetics*, *6*, 29–32.

Schork, N. J. (1997). Genetics of complex disease: Approaches, problems, and solutions. *American Journal of Respiratory and Critical Care Medicine*, *156*(4).

Schork, N. J. (2015). Personalized medicine: Time for one-person trials. *Nature Comment*, *520*, 609–611.

Schupf, N., Barral, S., Perls, T., Newman, A., Christensen, K., Thyagarajan, B., Province, M., Rossi, W. K., & Mayeux, R. (2013). Apolipoprotein e and familial longevity. *Neurobiology of Aging*, *34*(4), 1287 – 1291.

Sebastiani, P., Andersen, S., McIntosh, A., Nussbaum, L., Stevenson, M., Pierce, L., Xia, S., Salance, K., & Perls, T. (2016a). Familial risk for exceptional longevity. *North American Actuarial journal*, *20*(1), 57–64.

Sebastiani, P., Bae, H., Gurinovich, A., Soerensen, M., Puca, A., & Perls, T. T. (2017a). Limitations and risks of meta-analyses of longevity studies. *Mechanisms of Ageing and Development*, *165*(Part B), 139 – 146.

Sebastiani, P., Gurinovich, A., Bae, H., Andersen, S., Malovini, A., Atzmon, G., Villa, F., Kraja, A. T., Ben-Avraham, D., Barzilai, N., Puca, A., & Perls, T. (2017b). Four genome-wide association studies identify new extreme longevity variants. *The Journals of Gerontology: Series A*, *glx027*.

Sebastiani, P., Gurinovich, A., Bae, H., Andersen, S. L., & Perls, T. T. (2017c). Assortative mating by ethnicity in longevous families. *Frontiers in Genetics*, *8*, 186.

Sebastiani, P., Gurinovich, A., Nygaard, M., Sasaki, T., Sweigart, B., Bae, H., Andersen, S. L., Villa, F., Atzmon, G., Christensen, K., Arai, Y., Barzilai, N., Puca, A., Christiansen, L., Hirose, N., & Perls, T. T. (2019). APOE alleles and extreme human longevity. *The Journals of Gerontology: Series A*, *74*(1), 44–51.

Sebastiani, P., Monti, S., Morris, M., Gurinovich, A., et al. (SUBMITTED). A serum protein signature of APOE genotypes in centenarians. *Aging Cell*.

Sebastiani, P., Nussbaum, L., Andersen, S. L., Black, M. J., & Perls, T. T. (2016b). Increasing sibling relative risk of survival to older and older ages and the importance of precise definitions of "aging", "life span", and "longevity". *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *71*(3), 340–346.

Sebastiani, P., & Perls, T. T. (2012). The genetics of extreme longevity: Lessons from the New England Centenarian Study. *Frontiers in Genetics*, *3*(277).

Sebastiani, P., Solovieff, N., DeWan, A. T., Walsh, K. M., Puca, A., Hartley, S. W., Melista, E., Andersen, S., Dworkis, D. A., Wilk, J. B., Myers, R. H., Steinberg, M. H., Montano, M., Baldwin, C. T., Hoh, J., & Perls, T. T. (2012). Genetic signatures of exceptional longevity in humans. *PLOS ONE*, *7*(1), 1–22.

Sebastiani, P., Sun, F., Andersen, S., Lee, J., Wojczynski, M., Sanders, J., Yashin, A., Newman, A., & Perls, T. (2013). Families enriched for exceptional longevity also have increased health-span: Findings from the long life family study. *Frontiers in Public Health*, *1*, 38.

Shannon, P., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498–504.

Sherman, B. T., Huang, D. W., & Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13.

Sindi, S., Mangialasche, F., & Kivipelto, M. (2015). Advances in the prevention of Alzheimer's Disease. *F1000Prime Reports*, *7*(50).

Soares, H. D., Potter, W. Z., Pickering, E., Kuhn, M., Immermann, F. W., Shera, D. M., Ferm, M., Dean, R. A., Simon, A. J., Swenson, F., Siuciak, J. A., Kaplow, J., Thambisetty, M., Zagouras, P., Koroshetz, W. J., Wan, H. I., Trojanowski, J. Q., Shaw, L. M., & Biomarkers Consortium Alzheimer's Disease Plasma Proteomics Project, f. t. (2012). Plasma biomarkers associated with the Apolipoprotein E genotype and Alzheimer disease . *Archives of Neurology*, *69*(10), 1310–1317.

Solovieff, N., Hartley, S. W., Baldwin, C. T., Perls, T. T., Steinberg, M. H., & Sebastiani, P. (2010). Clustering by genetic ancestry using genome-wide SNP data. *BMC Genetics*, *11*(1), 108.

Sonnega, A., Faul, J. D., Ofstedal, M. B., Langa, K. M., Phillips, J. W., & Weir, D. R. (2014). Cohort profile: the Health and Retirement Study (HRS). *International Journal of Epidemiology*, *43*(2), 576–585.

Stallard, E., Bagley, O., Kulminski, A. M., Wu, D., Fang, F., Akushevich, I., Arbeev, K. G., Ukraintseva, S. V., Yashin, A. I., Arbeeva, L. S., Wojczynski, M. K., An, P., Christensen, K., Newman, A. B., Boudreau, R. M., Province, M. A., Thielke, S., Perls, T. T., & Elo, I. (2018). Genetics of human longevity from incomplete data: new findings from the Long Life Family Study. *The Journals of Gerontology: Series A*, *73*(11), 1472–1481.

Terry, D. F., Sebastiani, P., Andersen, S. L., & Perls, T. T. (2008). Disentangling the roles of disability and morbidity in survival to exceptional old age. *Archives of Internal Medicine*, *168*(3), 277–283.

the CARDIoGRAMplusC4D Consortium (2015). A comprehensive 1000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, *47*, 1121–1130.

The PLOS Medicine Editors, Rid, A., Johansson, M. A., Leung, G., Valantine, H., Burchard, E. G., Oh, S. S., & Zimmerman, C. (2016). Towards equity in health: researchers take stock. *PLoS Medicine*, *13*(11).

Thomas T. Perls, M. H. S. (1999). *Living to 100: Lessons in Living to your Maximum Potential at any Age*. Basic Books.

Timmers, P. R., Mounier, N., Lall, K., Fischer, K., Ning, Z., Feng, X., Bretherick, A. D., Clark, D. W., eQTLGen Consortium, Shen, X., Esko, T., Kutalik, Z., Wilson, J. F., & Joshi, P. K. (2019). Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife*, *8*, e39856.

Torkamani, A., Pham, P., Libiger, O., Bansal, V., Zhang, G., Zeeland, A. A. S.-V., Tewhey, R., Topol, E. J., & Schork, N. J. (2012). Clinical implications of human population differences in genome-wide rates of functional genotypes. *Frontiers in Genetics*, *3*, 211.

Trumble, B. C., Stieglitz, J., Blackwell, A. D., Allayee, H., Beheim, B., Finch, C. E., Gurven, M., & Kaplan, H. (2017). Apolipoprotein E4 is associated with improved cognitive function in Amazonian forager-horticulturalists with a high parasite burden. *The FASEB Journal*, *31*(4), 1508–1515.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K.,

Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., & Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220).

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., Sanli, K., von Feilitzen, K., Oksvold, P., Lundberg, E., Hober, S., Nilsson, P., Mattsson, J., Schwenk, J. M., Brunnström, H., Glimelius, B., Sjöblom, T., Edqvist, P.-H., Djureinovic, D., Micke, P., Lindskog, C., Mardinoglu, A., & Ponten, F. (2017). A pathology atlas of the human cancer transcriptome. *Science*, *357*(6352).

van Exel, E., Koopman, J. J. E., Bodegom, D. v., Meij, J. J., Knijff, P. d., Ziem, J. B., Finch, C. E., & Westendorp, R. G. J. (2017). Effect of APOE e4 allele on survival and fertility in an adverse environment. *PLOS ONE*, *12*(7), 1–13.

Wang, K. (2009). Testing for genetic association in the presence of population stratification in genome-wide association studies. *Genetic Epidemiology*, *33*(7), 637–645.

Wang, X., Lee, S., Zhu, X., Redline, S., & Lin, X. (2013). GEE-based SNP set association test for continuous and discrete traits in family-based association studies. *Genetic Epidemiology*, *37*(8), 778–786.

Weisgraber, K. H., Rall, S. C., & Mahley, R. W. (1981). Human E apoprotein heterogeneity. Cysteine-arginine interchanges in the amino acid sequence of the apo-E isoforms. *Journal of Biological Chemistry*, *256*(17), 9077–9083.

Wu, L., & Zhao, L. (2016). ApoE2 and Alzheimer's disease: time to take a closer look. *Neural Regeneration Research*, *11*(3), 412–413.

Yip, A. G., McKee, A. C., Green, R. C., Wells, J., Young, H., Cupples, L. A., & Farrer, L. A. (2005). APOE, vascular pathology, and the AD brain. *Neurology*, *65*(2), 259–265.

Young, R. D., Desjardins, B., McLaughlin, K., Poulain, M., & Perls, T. T. (2011). Typologies of extreme longevity myths. *Current Gerontology and Geriatrics Research*, *2010*, 423087.

CURRICULUM VITAE